VEL@DravidianLangTech: Sentiment Analysis of Tamil and Tulu

Kishore Kumar Ponnusamy[1], **Charmathi Rajkumar**[2],
Prasanna Kumar Kumaresan[3], **Elizabeth Sherly**[4]
Ruba Priyadharshini[5]
[1] Guru Nanak College, Chennai, Tamil Nadu, India
[2] The American College, Madurai, Tamil Nadu, India
[3] Insight SFI Research Centre for Data Analytics, University of Galway, Ireland
[4]Digital University Kerala, India
[5]The Gandhigram Rural Institute Deemed University, Tamil Nadu, India
{kishorep16002, charmathirajkumar}@gmail.com
prasanna.kumaresan@insight-centre.org
sherly@iiitmk.ac.in, rubapriyadharshini.a@gmail.com

Abstract

We participated in the Sentiment Analysis in Tamil and Tulu - DravidianLangTech 2023-RANLP 2023 task in the team name of VEL. This research focuses on addressing the challenge of detecting sentiment analysis in social media code-mixed comments written in Tamil and Tulu languages. Code-mixed text in social media often deviates from strict grammar rules and incorporates non-native scripts, making sentiment identification a complex task. To tackle this issue, we employ pre-processing techniques to remove unnecessary content and develop a model specifically designed for sentiment analysis detection. Additionally, we explore the effectiveness of traditional machine-learning models combined with feature extraction techniques. Our best model logistic regression configurations achieve impressive macro F1 scores of 0.43 on the Tamil test set and 0.51 on the Tulu test set, indicating promising results in accurately detecting instances of sentiment in code-mixed comments.

1 Introduction

The constant increase in Web and social media usage by the global audience has resulted in a substantial increase in the quantity of textual data expressing opinions (Shanmugavadivel et al., 2022). This publicly available textual data contains valuable insights that can be utilized in numerous disciplines, including marketing, finance, politics, and security (Shaik et al., 2023). These insights provide an excellent opportunity for individuals and businesses to discover the opinions of a user base, allowing them to make informed decisions regarding the enhancement of their brands and services (Chakravarthi et al., 2022).

However, as more opinionated data becomes available on the Internet, it becomes exceedingly difficult to read and trace emotions. This issue can be addressed using techniques of sentiment analysis (Divya et al., 2023). Sentiment analysis is regarded as the computational study of sentiments, opinions, emotions, and evaluations in order to comprehend better a person's reactions and attitudes towards multiple entities (Vasantharajan et al., 2023; Mercha and Benbrahim, 2023).

Due to its prospective applications, the computational treatment of opinion, sentiment, and subjectivity has recently garnered a great deal of attention. Sentiment analysis identifies the perspective(s) underlying a text segment; an example application is classifying a movie review as "thumbs up" or "thumbs down." The situation is more complicated for low-resource languages like Tamil and Tulu due to the poor performance of multilingual or language-specific models and the lack of adequate benchmark datasets (Thenmozhi et al., 2023; Subramanian et al., 2023; Navaneethakrishnan et al., 2023; Chakravarthi, 2023).

Tools and models for sentiment analysis have undergone substantial development over the course of several years, particularly for European languages. On the other hand, comparable tools for Dravidian languages are difficult to come by (Chinnaudayar Navaneethakrishnan et al., 2023; Kumaresan et al., 2022). This is due to the fact that cutting-edge pre-processing tools for Indian languages, such as POS taggers, shallow parsers, and other similar programs, are not easily accessible (Thavareesan and Mahesan, 2019, 2020b; Mahata et al., 2021). It has been observed that people in bilingual and multilingual communities often combine a number of languages when they are speaking or

Third Workshop on Speech and Language Technologies for Dravidian Languages,
pages 211–216, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-085-4_030

writing informally (B and A, 2021; Karim et al., 2022). Younger generations are increasingly using Dravidian code-mixed languages in a variety of contexts, including but not limited to advertising, entertainment, and social media (Priyadharshini et al., 2022; Subramanian et al., 2022). Examples of these languages include Malayalam and Tamil. Roman script is the most prevalent form used to write the language (Mandalam and Sharma, 2021; Gupta et al., 2021; Chakravarthi, 2022a,b).

We participated in sentiment analysis shared tasks for Tamil and Tulu (Hedge et al., 2023), which focused on detecting sentiment in social media comments. The task was organized by DravidianLangTech 2023. Using the provided dataset, we developed three machine-learning models that utilized TF-IDF feature extraction. Among our models, the logistic regression model yielded the best results, achieving a macro F1 score of 0.43 for Tamil and a macro F1 score of 0.51 for Tulu. These scores indicate the effectiveness of our approach in accurately identifying instances of sentiment in social media comments for the Tamil and Tulu languages. Our participation in this shared task has provided valuable insights into detecting and understanding discriminatory behavior on online platforms.

2 Related Work

Initially, sentiment analysis relies heavily on artificial feature selection and a sentiment dictionary for classification (Thavareesan and Mahesan, 2020a; Jiang et al., 2023). Three levels of sentiment analysis were investigated: document, sentence, and aspect. To categorize the prevailing sentiment of the entire document and sentence, the tasks at the document and sentence levels are virtually identical. On the other hand, the objective at the facet level is to identify opinions expressed about entities and aspects of entities. Earlier research proposed a number of sentiment analysis methods. There are three categories of these approaches: (i) lexicon-based, (ii) machine learning, and (iii) hybrid. The vast majority of research in this area concentrates on a lexicon-based approach that compiles sentiment words. All sentiment analysis tasks require machine learning approaches, the majority of which are either supervised or semi-supervised. A supervised approach is a common solution for sentiment-related tasks. On the basis of labeled datasets, it is frequently employed to ascertain the polarity of sentiments. The advantage of this method is its capacity to adapt and build efficiently trained models for a specific domain. The primary disadvantage of this method, however, is its inefficiency when applied to diverse data across domains. This is due to domain dependence and data scarcity in specific domains. Since 2011, deep learning has been utilized in the field of sentiment analysis research.

3 Task details

Sentiment analysis is the task of identifying subjective opinions or emotional responses about a given topic. It has been an active area of research in the past two decades in academia and industry. There is an increasing demand for sentiment detection on social media texts which are largely code-mixed for Dravidian languages. Code-mixing is a prevalent phenomenon in a multilingual community, and code-mixed texts are sometimes written in non-native scripts. Systems trained on monolingual data fail on code-mixed data due to the complexity of code-switching at different linguistic levels in the text. The shared Task - A presents a new gold standard corpus for sentiment detection of code-mixed text in Dravidian languages (Tamil-English, and Tulu-English)[1].

The goal of this task is to identify the sentiment polarity of the code-mixed dataset of comments/posts in Tamil-English and Tulu-English collected from social media. The dataset description was tabulated in the Table 1. The comment/post may contain more than one sentence, but the average sentence length of the corpora is 1. Each comment/post is annotated with sentiment polarity at the comment/post level. This dataset also has class imbalance problems depicting real-world scenarios. Our proposal aims to encourage research that will reveal how sentiment is expressed in code-mixed techniques on social media.

4 Methodology

Our methodology for the sentiment analysis shared task involved implementing three traditional machine-learning models and performing several data processing steps. We started by importing essential packages such as Pandas, NumPy, and NLTK[2] with sklearn library[3]. These packages

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

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

[3] https://scikit-learn.org/stable/

Table 1: Dataset description for Tamil and Tulu

Tamil			
Classes	**Train**	**Dev**	**Test**
Positive	20,070	2,257	73
Negative	4,271	480	338
unknown_state	5,628	611	137
Mixed_feelings	4,020	438	101
Tulu			
Positive	3,118	369	344
Negative	646	90	60
unknown_state	1,719	202	197
Mixed_feelings	974	120	107

played a crucial role in tasks such as data loading, cleaning, tokenization, vectorization, and modeling.

To begin, we read the train, development, and test files using Pandas, which contained the text data used for training and evaluating the sentiment analysis models. We then proceeded to clean the files by removing unwanted links, '@' symbols, and other characters that could potentially impact the accuracy of our models. Tokenization was performed using the NLTK package, breaking down the text data into individual words or terms. Lowercasing the contents of the files was another important step to ensure consistency, as it helped improve the models' ability to detect sentiment analysis accurately. Once the data was preprocessed, we used the Term Frequency - Inverse Document Frequency (TFIDF) vectorizer method to transform the text data into a numeric representation that could be utilized by our machine-learning models. The models employed for training were logistic regression, multinomial naive Bayes, and linear SVC. We trained these models using the preprocessed train data. To evaluate their performance, we utilized the development data, analyzing metrics such as the confusion matrix[4] and classification report[5]. The macro F1 scores obtained from the development data allowed us to determine the best-performing model.

In the final stage, we assessed the selected best model's performance using the test data. This step provided insights into the model's generalizability and effectiveness in handling new instances of sentiment analysis. Our comprehensive methodology ensured a robust approach to sentiment analysis in social media comments, leveraging traditional machine-learning techniques and effective data processing steps.

[4] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html

[5] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html

5 Results and Discussion

In our evaluation of the machine learning models for sentiment analysis, we employed several performance metrics, including Accuracy (ACC), Macro Precision (M_P), Macro Recall (M_R), Macro F1 (M_F1), Weighted Precision (W_P), Weighted Recall (W_R), and Weighted F1 (W_F1) scores. Our experimentation involved three different models: Logistic Regression (LR), Multinomial Naive Bayes (MNB), and Linear Support Vector Classification (L-SVC), all utilizing TF-IDF feature extraction.

For the Tamil language, the macro F1 scores obtained for the models were as follows: logistic regression achieved a score of 0.43, Multinomial Naive Bayes obtained 0.20, and Linear Support Vector Classification achieved 0.41. Among these models, logistic regression performed particularly well in accurately identifying sentiment in Tamil social media comments when using the TF-IDF vectorizer. The overall classification report is tabulated in Table 2.

Moving on to the Tulu language, we obtained the following macro F1 scores: logistic regression achieved a score of 0.51, Multinomial Naive Bayes obtained 0.25, and Linear Support Vector Classification achieved 0.49. Once again, logistic regression demonstrated superior performance in accurately detecting sentiment in Tulu social media comments when utilizing TF-IDF feature extraction. The overall classification report is tabulated in Table 3.

These results indicate that the logistic regression model outperformed the other models across both Tamil and Tulu languages in terms of macro F1 scores. The TF-IDF vectorizer played a crucial role in capturing important features and enabling accurate sentiment analysis. Overall, our experiments show the effectiveness of logistic regression coupled with TF-IDF feature extraction for sentiment analysis in social media comments for both Tamil and Tulu languages. These findings highlight the potential of machine learning techniques in identifying sentiment and understanding discriminatory

Table 2: Results of the machine learning models with TFIDF in classification report for Tamil

Classifiers	ACC	M_P	M_R	M_F1	W_P	W_R	W_F1
LR	0.64	0.50	0.41	**0.43**	0.59	0.64	0.60
MNB	0.61	0.40	0.26	0.20	0.53	0.61	0.46
L-SVC	0.61	0.46	0.39	0.41	0.56	0.61	0.57

Table 3: Results of the machine learning models with TFIDF in classification report for Tulu

Classifiers	ACC	M_P	M_R	M_F1	W_P	W_R	W_F1
LR	0.66	0.60	0.50	**0.51**	0.64	0.66	0.62
MNB	0.55	0.34	0.30	0.25	0.46	0.55	0.42
L-SVC	0.61	0.53	0.47	0.49	0.58	0.61	0.59

behavior on online platforms in multilingual settings.

6 Conclusion

This study focused on the detection of sentiment analysis in the Tamil and Tulu languages using machine learning models. By employing traditional models with TF-IDF feature extraction, we achieved impressive results, with a macro F1 score of 0.43 on the Tamil test set and 0.51 on the Tulu test set. These findings underscore the effectiveness of our machine learning approach in accurately classifying sentiment analysis instances. We conclude that the TF-IDF with logistic regression (LR) model is a suitable choice for this task, offering potential applications in addressing social issues and promoting inclusivity in online spaces. Future research can further improve the model's performance through fine-tuning techniques and larger datasets. This study contributes to advancing sentiment analysis and highlights the importance of accurate sentiment classification in understanding and addressing social dynamics in multilingual online platforms.

References

Bharathi B and Agnusimmaculate Silvia A. 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.

Bharathi Raja Chakravarthi. 2022a. Hope speech detection in YouTube comments. *Social Network Analysis and Mining*, 12(1):75.

Bharathi Raja Chakravarthi. 2022b. Multilingual hope speech detection in english and dravidian languages. *International Journal of Data Science and Analytics*, 14(4):389–406.

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. 2022. 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.

Subalalitha Chinnaudayar Navaneethakrishnan, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Malliga Subramanian, Prasanna Kumar Kumaresan, Bharathi, Lavanya Sambath Kumar, and Rahul Ponnusamy. 2023. Findings of shared task on sentiment analysis and homophobia detection of youtube comments in code-mixed dravidian languages. In *Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '22, page 18–21, New York, NY, USA. Association for Computing Machinery.

S. Divya, N. Sripriya, Daphne Evangelin, and G. Saai Sindhoora. 2023. Opinion classification on code-mixed Tamil language. In *Speech and Language Technologies for Low-Resource Languages*, pages 155–168, Cham. Springer International Publishing.

Akshat Gupta, Sai Krishna Rallabandi, and Alan W Black. 2021. Task-specific pre-training and cross lingual transfer for sentiment analysis in Dravidian code-switched languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 73–79, Kyiv. Association for Computational Linguistics.

Asha Hedge, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rahul Ponnusamy, Subalalitha Chinnaudayar Navaneethakrishnan, Lavanya S.K, Durairaj Thenmozhi, Martha Karunakar, Shreya Sriram, 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.

Lei Jiang, Yuan Li, Jing Liao, Ziwei Zou, and Caoqing Jiang. 2023. Research on non-dependent aspect-level sentiment analysis. *Knowledge-Based Systems*, 266:110419.

Md Rezaul Karim, Sumon Kanti Dey, Tanhim Islam, Md Shajalal, and Bharathi Raja Chakravarthi. 2022. Multimodal hate speech detection from bengali memes and texts. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 293–308. Springer.

Prasanna Kumar Kumaresan, Rahul Ponnusamy, Elizabeth Sherly, Sangeetha Sivanesan, and Bharathi Raja Chakravarthi. 2022. Transformer based hope speech comment classification in code-mixed text. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 120–137. Springer.

Sainik Mahata, Dipankar Das, and Sivaji Bandyopadhyay. 2021. Sentiment classification of code-mixed tweets using bi-directional RNN and language tags. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 28–35, Kyiv. Association for Computational Linguistics.

Asrita Venkata Mandalam and Yashvardhan Sharma. 2021. Sentiment analysis of Dravidian code mixed data. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 46–54, Kyiv. Association for Computational Linguistics.

El Mahdi Mercha and Houda Benbrahim. 2023. Machine learning and deep learning for sentiment analysis across languages: A survey. *Neurocomputing*, 531:195–216.

Subalalitha Chinnaudayar Navaneethakrishnan, Sathiyaraj Thangasamy, Nithya R, Info-farmer, and Neechalkaran. 2023. Exploring the opportunities and challenges in contributing to Tamil Wikimedia. In *Speech and Language Technologies for Low-Resource Languages*, pages 253–262, Cham. Springer International Publishing.

Ruba Priyadharshini, Bharathi raja Chakravarthi, Sajeetha Thavareesan, Dhivya Chinnappa, Durairaj Thenmozhi, and Rahul Ponnusamy. 2022. Overview of the dravidiancodemix 2021 shared task on sentiment detection in tamil, malayalam, and kannada. In *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '21, page 4–6, New York, NY, USA. Association for Computing Machinery.

Thanveer Shaik, Xiaohui Tao, Christopher Dann, Haoran Xie, Yan Li, and Linda Galligan. 2023. Sentiment analysis and opinion mining on educational data: A survey. *Natural Language Processing Journal*, 2:100003.

Kogilavani Shanmugavadivel, Malliga Subramanian, Prasanna Kumar Kumaresan, Bharathi Raja Chakravarthi, B Bharathi, Subalalitha Chinnaudayar Navaneethakrishnan, Lavanya Sambath Kumar, Thomas Mandl, Rahul Ponnusamy, Vasanth Palanikumar, et al. 2022. Overview of the shared task on sentiment analysis and homophobia detection of youtube comments in code-mixed dravidian languages.

Malliga Subramanian, Ramya Chinnasamy, Prasanna Kumar Kumaresan, Vasanth Palanikumar, Madhoora Mohan, and Kogilavani Shanmugavadivel. 2023. Development of multi-lingual models for detecting hope speech texts from social media comments. In *Speech and Language Technologies for Low-Resource Languages*, pages 209–219, Cham. Springer International Publishing.

Malliga Subramanian, Rahul Ponnusamy, Sean Benhur, Kogilavani Shanmugavadivel, Adhithiya Ganesan, Deepti Ravi, Gowtham Krishnan Shanmugasundaram, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2022. Offensive language detection in tamil youtube comments by adapters and cross-domain knowledge transfer. *Computer Speech & Language*, 76:101404.

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.

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.

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.

Durairaj Thenmozhi, C. Jerin Mahibha, S. Kayalvizhi, M. Rakesh, Y. Vivek, and V. Poojesshwaran. 2023. Paraphrase detection in Indian languages using deep learning. In *Speech and Language Technologies for Low-Resource Languages*, pages 138–154, Cham. Springer International Publishing.

Charangan Vasantharajan, Ruba Priyadharshini, Prasanna Kumar Kumarasen, Rahul Ponnusamy, Sathiyaraj Thangasamy, Sean Benhur, Thenmozhi Durairaj, Kanchana Sivanraju, Anbukkarasi Sampath, and Bharathi Raja Chakravarthi. 2023. Tamilemo: Fine-grained emotion detection dataset for Tamil.

In *Speech and Language Technologies for Low-Resource Languages*, pages 35–50, Cham. Springer International Publishing.