Selam@DravidianLangTech:Sentiment Analysis of Code-Mixed Dravidian Texts using SVM Classification

Selam Abitte Kanta
Instituto Politécnico Nacional
Centro de Investigación en Computación
selaminadady300@gmail.com

Grigori Sidorov
Instituto Politécnico Nacional
Centro de Investigación en Computación
sidorov@cic.ipn.mx

Abstract

Sentiment analysis in code-mixed text written in Dravidian languages. Specifically, Tamil-English and Tulu-English. This paper describes the system paper of the RANLP-2023 shared task. The goal of this shared task is to develop systems that accurately classify the sentiment polarity of code-mixed comments and posts. be provided with development, training, and test data sets containing code-mixed text in Tamil-English and Tulu-English. The task involves message-level polarity classification, to classify YouTube comments into positive, negative, neutral, or mixed emotions. This Code-Mix was compiled by RANLP-2023 organizers from posts on social media. We use classification techniques SVM and achieve an F1 score of 0.147 for Tamil-English and 0.518 for Tulu-English.

1 Introduction

Social media platforms have become significant sources of user-generated content, providing a wealth of information about people's opinions, emotions, and attitudes. Sentiment analysis, a subfield of natural language processing, aims to automatically classify the sentiment or emotional tone expressed in textual data. Currently, in the area of NLP, different researchers are developing different NLP applications in code-mixed datasets. Some of the applications are code-mixed sentiments analysis

which involves identifying subjective opinions or emotional responses, has gained significant attention in both academia and industry over the past two decades. One emerging challenge in sentiment analysis is the detection of sentiment in social media texts, particularly in Dravidian languages, where code-mixing is prevalent (Shahiki Tash et al., 2022). Social media platforms have become more integrated into this digital era and have impacted various people's perceptions of networking and socializing (Tonja et al., 2022) machine translation detection, sentiment analysis and language identification (Gemeda Yigezu et al., 2022).

Code-mixing refers to the phenomenon where multiple languages or language varieties are used within a single conversation or text. This linguistic practice is prevalent in multilingual societies, particularly in regions where diverse languages coexist. In the context of Dravidian languages, such as Tamil and English, code-mixing has gained prominence on social media platforms, where users freely express their thoughts using a mixture of both languages. Code-mixing or code-switching is the alternation between two or more languages at the level of the document, paragraph, comments, sentence, phrase, word, or morpheme. It is a distinctive aspect of conversation or dialogue in bilingual and multilingual societies (Barman et al., 2014)

As the classification model for sentiment analysis, we propose using the Support Vector Machine (SVM) algorithm. SVM has proven effective in various natural language processing tasks, including sentiment analysis (Mullen and Collier, 2004).

By utilizing SVM, we aim to leverage its ability to capture complex patterns in text data and its flexibility in handling high-dimensional feature spaces. In social media, low-resourced languages such as Tamil and Malayalam have been increasingly used along with English (Najiha and Romadhony, 2023)

The contribution of this research lies in advancing the understanding of sentiment expression in code-mixed scenarios on social media, within the context of Dravidian languages of Tamil-English and Tulu-English language. Accurate identification of sentiment polarity at the message level provides valuable insights into user emotions and attitudes in code-mixed interactions. also known as opinion mining,(Nandwani and Verma, 2021) is a natural language processing technique that aims to determine the sentiment expressed in a piece of text (Liu et al., 2010).

By applying sentiment analysis to code-mixed

Third Workshop on Speech and Language Technologies for Dravidian Languages,
pages 176–179, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-085-4_024

interactions, researchers and analysts can gain a deeper understanding of how users feel and their attitudes toward specific topics or situations. accurate identification of sentiment polarity in code-mixed interactions can provide valuable insights into user emotions and attitudes(Saura et al., 2023). Advancements in code-mixed sentiment analysis can contribute to a better understanding of user sentiment in multilingual communities, social media, customer support, and other domains where code-mixing is prevalent(Agüero-Torales et al., 2021).

The findings of this research will facilitate the development of more robust sentiment analysis techniques for analyzing multilingual social media data, enabling improved understanding and interpretation of user sentiments across diverse linguistic contexts(Hegde et al.).

2 Related Work

Along with language-specific preprocessing techniques, the implemented model used sub-word level representations to incorporate features at the morpheme level, the smallest meaningful unit of any language.

It was evaluated by weighted average F1-score, the subword level approach achieved the 5th highest score 47 in the Tamil task, and the 12th rank in the Malayalam task (Shahiki Tash et al., 2022). People use code-mixing because it is much easier and more effective to express their feelings, grab the attention of others, and are not fluent in one of the languages used (Wongso et al., 2022). As a multilingual country, people commonly mix or switch from their regional or native language to Indonesian (Najiha and Romadhony, 2023). It is frequently heard in daily conversations in the neighborhood or on social media. To improve abusive language detection in English social media communications, (Felbo et al., 2017) used the 'deepmoji' technique, which was first announced in 2017(Chakravarthi et al., 2023).

This strategy is primarily based on pretraining a neural network model for offensive language classification using emojis as poorly supervised training labels. A lexical syntactic feature architecture was proposed to strike a balance between identifying offensive content and potentially offensive users in social media (Luo et al., 2015) the challenge of cross-lingual classification due to linguistic differences between languages. mentions that the SVM and KNN algorithms were effective for this task, showcasing the importance of selecting appropriate algorithms for different languages(Ahani et al.)

This data on polarity can help in understanding public opinion. Furthermore, including sentiment analysis can improve the performance of tasks such as recommendation system (Andrew and Gao, 2007) to train some machine learning classifiers with various syntax-based n-gram features.

3 Task A Description and Dataset

The shared task on sentiment analysis in Tamil and Tulu focuses on message-level polarity classification. The dataset provided for this task consists of code-mixed text in Dravidian languages, namely Tamil-English and Tulu-English. Tamil-English code-switched, sentiment-annotated corpus containing 15,744 comment posts from YouTube.(Chakravarthi et al., 2020). and the code-mixed Tulu annotated corpus of 7,171 YouTube comments is created.(Hegde et al., 2022)

The comments and posts in the dataset may contain more than one sentence, but the average sentence length across the corpora is one. Each comment and post is annotated with sentiment polarity at the message level, indicating whether it expresses a positive, negative, neutral, or mixed emotional sentiment.

These datasets reflect the real-world scenarios of social media texts, exhibiting class imbalance issues commonly encountered in sentiment analysis tasks. The datasets provided in this task will facilitate the exploration of innovative approaches and techniques for sentiment analysis in multilingual and multicultural contexts.

4 Methods

For sentiment analysis tasks in code-mixed comments and posts in Tamil-English and Tulu-English, we propose employing the Support Vector Machines (SVM) model as the classification technique. SVM has proven effective in various natural language processing tasks, including sentiment analysis, and has demonstrated robustness in handling high-dimensional feature spaces.

4.1 Feature Engineering

To represent the textual data as numerical features suitable for SVM classification, we will explore various feature extraction techniques. This may include traditional approaches such as bag-of-words (BoW), term frequency-inverse document

frequency (TF-IDF), or more advanced methods like word embeddings (e.g., Word2Vec or GloVe) or contextual embeddings (e.g., BERT or Roberta). By representing the text as feature vectors, we can capture the important information relevant to sentiment analysis.

4.2 Model Construction

we trained the SVM model using the labeled training dataset. The SVM algorithm aims to find an optimal hyperplane that separates the different sentiment classes in the feature space. By adjusting the hyperparameters of the SVM model, such as the kernel function and regularization parameters, we can fine-tune the model's performance.

Such as accuracy, precision, recall, and F1 score. Cross-validation techniques like k-fold cross-validation may be employed to ensure the robustness of the results. Additionally, we will analyze the model's performance on the development dataset to identify potential areas for improvement. Parameters that were used in SVM and TF-IDF were as follows. For the SVM classifier, we used C=0.1, kernel='poly', degree=3, and gamma='scale'. For the TF-IDF vectorizer, we used analyzer='char_wb', ngram_range=(2,6), min_df=0, and norm='l1'.

5 Results

In the shared task, a Support Vector Machine (SVM) model was employed for message-level polarity classification of code-mixed text in Tamil-English and Tulu-English. The evaluation metric used was the F1 score, which provides a measure of the model's performance across all sentiment classes. The results obtained for the SVM model on the provided datasets were as follows in Table 1.

Run	language	F1-score
Run1	Tamil	0.147
Run1	Tulu	0.518

Table 1: F1-Score

6 Conclusion

This study shows how different languages may be identified in code-mix data using a classifier that uses two algorithms, SVM and TF-IDF. The first technique produces better results, with the best weighted average F1-score of 0.147 and 0.518.

Acknowledgments

The work was done with partial support from the Mexican Government through grant A1-S-47854 of CONACYT, Mexico, grants 20232138, 20232080, and 20231567 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico, and acknowledge the support of Microsoft through the Microsoft Latin America Masters Award.

References

Marvin M Agüero-Torales, José I Abreu Salas, and Antonio G López-Herrera. 2021. Deep learning and multilingual sentiment analysis on social media data: An overview. *Applied Soft Computing*, 107:107373.

Zahra Ahani, Grigori Sidorov, Olga Kolesnikova, and Alexander Gelbukh. Hope speech detection from text using tf-idf features and machine learning algorithms.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 13–23, Doha, Qatar. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Manoj Balaji Jagadeeshan, Vasanth Palanikumar, and Ruba Priyadharshini. 2023. Offensive language identification in dravidian languages using mpnet and cnn. *International Journal of Information Management Data Insights*, 3(1):100151.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John P McCrae. 2020. Corpus creation for sentiment analysis in code-mixed tamil-english text. *arXiv preprint arXiv:2006.00206*.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524*.

Mesay Gemeda Yigezu, Atnafu Lambebo Tonja, Olga Kolesnikova, Moein Shahiki Tash, Grigori Sidorov, and Alexander Gelbukh. 2022. Word level language identification in code-mixed Kannada-English texts using deep learning approach. In *Proceedings of the*

19th International Conference on Natural Language Processing (ICON): Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts, pages 29–33, IIIT Delhi, New Delhi, India. Association for Computational Linguistics.

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.

Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rahul Ponnusamy, SUBALALITHA CN, Lavanya S K, Thenmozhi D, Martha Karunakar, Shreya Shreeram, and Sarah" Aymen. Findings of the shared task on sentiment analysis in tamil and tulu code-mixed text.

Bing Liu et al. 2010. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2(2010):627–666.

A-Li Luo, Yong-Heng Zhao, Gang Zhao, Li-Cai Deng, Xiao-Wei Liu, Yi-Peng Jing, Gang Wang, Hao-Tong Zhang, Jian-Rong Shi, Xiang-Qun Cui, et al. 2015. The first data release (dr1) of the lamost regular survey. *Research in Astronomy and Astrophysics*, 15(8):1095.

Tony Mullen and Nigel Collier. 2004. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 412–418.

Hajarot Najiha and Ade Romadhony. 2023. Sentiment analysis on indonesian-sundanese code-mixed data. In *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, pages 1–7.

Pansy Nandwani and Rupali Verma. 2021. A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11(1):81.

Jose Ramon Saura, Daniel Palacios-Marqués, and Domingo Ribeiro-Soriano. 2023. Privacy concerns in social media ugc communities: Understanding user behavior sentiments in complex networks. *Information Systems and e-Business Management*, pages 1–21.

M. Shahiki Tash, Z. Ahani, A.l. Tonja, M. Gemeda, N. Hussain, and O. Kolesnikova. 2022. Word level language identification in code-mixed Kannada-English texts using traditional machine learning algorithms. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON): Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts*, pages 25–28, IIIT Delhi, New Delhi, India. Association for Computational Linguistics.

Atnafu Lambebo Tonja, Mesay Gemeda Yigezu, Olga Kolesnikova, Moein Shahiki Tash, Grigori Sidorov, and Alexander Gelbuk. 2022. Transformer-based model for word level language identification in code-mixed kannada-english texts. *arXiv preprint arXiv:2211.14459*.

Wilson Wongso, Henry Lucky, and Derwin Suhartono. 2022. Pre-trained transformer-based language models for sundanese. *Journal of Big Data*, 9(1):39.