WordWizards@DravidianLangTech 2024: Sentiment Analysis in Tamil and Tulu using Sentence Embedding

Shreedevi Seluka Balaji[1], Akshatha Anbalagan[1], Priyadharshini T[1], Niranjana A[1], Durairaj Thenmozhi[1]

[1]**Sri Sivasubramaniya Nadar College of Engineering, Tamil Nadu, India**

shreedevi2210389@ssn.edu.in , akshatha2210397@ssn.edu.in ,
priyadharshini2210228@ssn.edu.in , niranjana2210379@ssn.edu.in , theni_d@ssn.edu.in

Abstract

Sentiment Analysis of Dravidian Languages has begun to garner attention recently as there is more need to analyze emotional responses and subjective opinions present in social media text. As this data is code-mixed and there are not many solutions to code-mixed text out there, we present to you a stellar solution to DravidianLangTech 2024: Sentiment Analysis in Tamil and Tulu task [1]. To understand the sentiment of social media text, we used pre trained transformer models and feature extraction vectorizers to classify the data with results that placed us 11th in the rankings for the Tamil task and 8th for the Tulu task with a accuracy F1 score of 0.12 and 0.30 which shows the efficiency of our approach.

1 Introduction

Social media transcends borders, fostering global communication with user-generated content expressing emotions. Research, like Wang et al.'s study (Wang et al., 2022), uses ensemble models to detect depression signs from social media. The surge in platforms underscores the importance of sentiment analysis in local languages, as seen in the study on Dravidian languages like Tamil (Chakravarthi et al., 2021). English intertwining with native scripts poses challenges, and Natural Language Processing techniques are suggested for effective sentiment analysis, focusing on specific subjects. The task involves classifying YouTube comments into labels for Tulu and Tamil, aligning with findings that sentiment analysis in code-mixed text requires careful consideration of linguistic nuances in multilingual contexts. The rest of the paper is organized as follows. Section 2 outlines the related works emphasizing Sentiment Analysis in Dravidian languages. Section 3 presents a description of the dataset. Section 4 describes the methodology used for the shared task. Section 5 discusses the result and analysis of the task assigned. In Section 6 concludes the paper followed by ethics statement in Section 7.

[1]https://sites.google.com/view/dravidianlangtech-2024/home

2 Related Works

Sentiment analysis is vital amid the vast data on social media. Approaches include lexicon-based methods, using dictionaries like Sentiwordnet and TF-IDF (Term Frequency-Inverse Document Frequency), and machine learning methods employing SVM and Naïve Bayes models. Combining these methods enhances efficiency, addressing unstructured data effectively (Jada et al., 2021; Drus and Khalid, 2019). Research on sentiment analysis of Dravidian languages, exemplified by Chakravarthi et al.'s study (Chakravarthi et al., 2021), explores challenges in linguistic diversity and code-mixing, emphasizing the need for tools handling code-mixed text (S. K. et al., 2024). The findings underscore the importance of linguistic considerations in capturing sentiments in Dravidian languages and English on social media (Chakravarthi et al., 2021; Hegde et al., 2023), providing valuable insights into public opinion, cultural discourse, and community dynamics. This highlights the necessity of language-specific sentiment analysis tools for understanding sentiments in diverse linguistic communities.

3 Dataset

The datasets provided were comments from social media and each comment corresponded to a particular label. The labels were Positive for comments that were appreciative, Negative for hate speech and other detected signs on aggression, Mixed feeling/feelings for comments on the fence and Neutral/unknown_state for comments that were neither positive nor negative. The dataset was divided into three parts namely train,test and dev. The train and dev datasets had *text* and *category* as labels while the test dataset had *id* and *text* as labels for

Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 218–222

Tamil (Chakravarthi et al., 2020). The train and dev datasets had *Text* and *Annotations* as labels while the test dataset had *ID* and *Text* as labels (Hegde et al., 2022) for Tulu. The test data for both languages was not labeled and we had to find and classify the test data as part of the task. The train dataset consisted of 33,988 rows of which
Positive: 20,070
Unknown State: 5,628
Negative: 4,271
Mixed Feelings: 4,020 for Tamil and 6945 rows of which
Positive: 3,352
Neutral: 1,854
Mixed Feeling: 1,041
Negative: 698 for Tulu. The dev dataset consisted of 3,785 rows for Tamil and 500 rows for Tulu. The test dataset consisted of 650 rows for Tamil and 502 rows for Tulu.

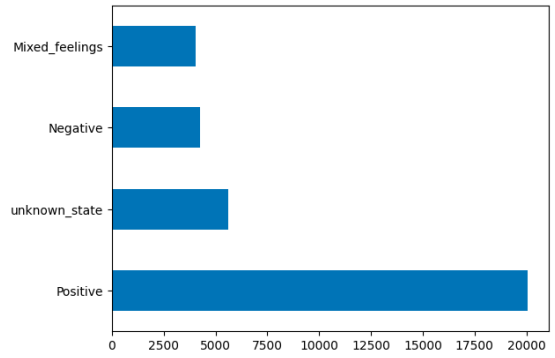

Figure 1: Category Metrics

4 Methodology

4.1 Preprocessing

Preprocessing refers to the ways in which we can clean data and remove noise (inconsistencies). These methods are applied to raw data so we may process them better later on. Preprocessing involves many steps which include:

1.Tokenisation to breakdown sentences into words or subwords.

2.Converting all text to lowercase.

3.Removing stopwords which are words that do not add any value to the meaning of the sentence. This is done using a list of stopwords, provided by the Python package NLTK and enhancing the list with our own Tamil and Tulu stopwords, which can be iterated through each row and removed.

4.De-emojify using the re package to remove emoticons, flags in iOS, symbols and pictographs, transport and map symbols.

4.2 Feature Extraction

LaBSE, or Language-agnostic BERT Sentence Embedding, is a multilingual variant of BERT designed for cross-lingual natural language processing tasks. Unlike traditional BERT models that are trained on individual languages, LaBSE is trained to generate language-agnostic sentence embeddings, making it effective for applications where text spans multiple languages(Feng et al., 2020).

The key innovation in LaBSE lies in its training approach. It utilizes a parallel data mining strategy, leveraging publicly available parallel sentences in multiple languages. This allows LaBSE to learn cross-lingual representations by aligning sentence embeddings across languages in a shared embedding space. The model is trained to map semantically similar sentences in different languages to nearby points in the embedding space.

The training process involves encoding sentences into fixed-dimensional vectors, ensuring that semantically equivalent sentences in different languages are close to each other in the embedding space. This shared embedding space enables LaBSE to capture universal semantic features across languages, facilitating effective cross-lingual understanding.

The feature extraction process with LaBSE involves encoding a given sentence into a dense vector representation. This vector, often referred to as a sentence embedding, encapsulates the semantic meaning of the input sentence. These embeddings can then be used for various downstream tasks, such as cross-lingual document retrieval, sentiment analysis, or machine translation.

LaBSE's effectiveness stems from its ability to generate language-agnostic representations, making it particularly useful for scenarios where language boundaries are fluid or when dealing with multilingual datasets. It allows practitioners to perform cross-lingual tasks without the need for language-specific models, offering a unified approach for diverse linguistic contexts.

4.3 Models Used

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a powerful machine learning algorithm employed for classification and regression tasks by finding a hyperplane

that maximizes the margin between data points of different classes. Its effectiveness in handling high-dimensional data and capability to identify complex decision boundaries make SVM widely applied across diverse domains (Cortes and Vapnik, 1995).

K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a versatile algorithm utilized for classification and regression, making predictions by considering the majority class or average of the k-nearest data points in the feature space. Renowned for its simplicity and ease of implementation, KNN is widely adopted in applications such as pattern recognition and recommendation systems (Altman, 1992).

Naive Bayes

Naive Bayes, a probabilistic classification algorithm founded on Bayes' theorem with the assumption of feature independence, has demonstrated remarkable effectiveness in tasks like text classification, spam filtering, and sentiment analysis. Its computational efficiency, minimal training data requirement, and suitability for high-dimensional datasets make Naive Bayes widely employed (Rish et al., 2001).

4.4 KNN

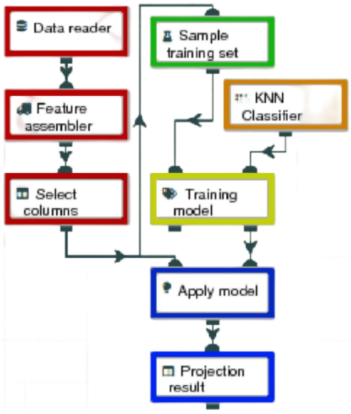

Figure 2: Workflow of KNN Model

The KNN workflow, illustrated in the figure, encompasses a sequence of operations. This includes a step for reading and preprocessing data, a subsequent step for sampling data intended for training a K Nearest Neighbors classifier, and a final step for applying the model to the dataset. In the end, the model that prevailed, with the best results was the K-Nearest Neighbors model with an overall macro F1 score of 0.62 for the Tamil train and dev datasets with an accuracy of 0.12 and a macro F1 score of 0.07 with the test dataset, placing us 11th on the ranklist and for the Tulu dataset, a macro F1 score of 0.66 and 0.67 on the training and dev datasets and a macro F1 score of 0.26 and an accuracy 0.30 with the test dataset, placing us 8th on the ranklist.

5 Results

5.1 Analysis

Here are the comparison metrics of the models applied to the training dataset for Tamil.

Metric	0	1	2	3
Precision	0.56	0.80	0.74	0.73
Recall	0.71	0.90	0.32	0.45

Table 1: Precision and Recall on Tamil Training Data

Label 0- unknown_state: Precision score 0.56 means that whenever the model predicts Unknown state, it is correct around 56% of the time. Recall 0.71 means that it identify 71% of the actual instances belonging to class 0.

Label 1- Positive: Precision score 0.80 means that whenever the model predicts Positive, it is correct around 80% of the time. Recall 0.90 means that it identify 90% of the actual instances belonging to class 1.

Label 2- Mixed_feelings: Precision score 0.74 means that whenever the model predicts Positive, it is correct around 74% of the time. Recall 0.83 means that it identify 83% of the actual instances belonging to class 2.

Label 3- Negative: Precision score 0.73 means that whenever the model predicts Positive, it is correct around 73% of the time. Recall 0.45 means that it identify 45% of the actual instances belonging to class 0.

Metric	Support Vector Machine	Naive Bayes	KNN
Accuracy	0.61	0.56	0.74
Macro Average F1 score	0.30	0.44	0.62

Table 2: Comparison of metrics on Tamil Training Data

The model demonstrates an overall performance, as indicated by the weighted average F1-score of 0.73, considering the support (number of instances)

for each class. The accuracy, standing at 0.74, reflects the percentage of correctly classified instances across all classes. The macro average F1-score, measuring 0.62, provides an assessment of the model's performance across all classes while treating them equally, irrespective of class imbalances.

Metric	Support Vector Machine	Naive Bayes	KNN
Accuracy	0.65	0.35	0.77
Macro Average F1 score	0.54	0.32	0.66

Table 3: Comparison of metrics on Tulu Training Data

Tests conclude that the KNN model has an overall F1 score accuracy of 0.30 with the test data for Tulu along with a 0.70 precision for the Positive label.

6 Conclusion

This study explores sentiment analysis in Dravidian languages, Tamil and Tulu, using LaBSE for feature extraction and KNN for classification. Achieving competitive rankings in the DravidianLangTech 2024 competition (11th for Tamil, 8th for Tulu), our model demonstrated promising precision and recall, particularly excelling in handling uncertain sentiments in the "Unknown State" class for Tamil. While acknowledging limitations and scalability challenges, we emphasize ethical considerations, including privacy and cultural sensitivity. In conclusion, our research provides insights into effective strategies for sentiment analysis in code-mixed social media text in Dravidian languages, showcasing the viability of LaBSE and KNN in culturally diverse contexts.

Label	Precision	Recall	F1 score
Mixed Feeling	0.15	0.23	0.18
Negative	0.10	0.21	0.14
Neutral	0.27	0.39	0.32
Positive	0.70	0.28	0.40

Table 4: Classification Report of Tulu Test Set

7 Limitations

Performing sentiment analysis on YouTube comments faces challenges due to their diverse language styles and informal expressions, making traditional models like K-Nearest Neighbors (KNN) and TF-IDF less effective in capturing nuanced semantics. The scalability of KNN becomes a concern with large datasets, and the presence of multilingual text further complicates sentiment representation. These limitations highlight the need for more sophisticated, context-aware approaches in future research to better handle the diverse nature of YouTube comments.

Label	Precision	Recall	F1 score
Mixed_feelings	0.14	0.01	0.02
Negative	0.42	0.01	0.03
Positive	0.11	0.92	0.20
unknown_state	0.18	0.03	0.05

Table 5: Classification Report of Tamil Test Set

8 Ethics Statement

Ethical considerations in conducting sentiment analysis on YouTube comments are crucial for maintaining responsible research practices. Privacy, consent, and the handling of sensitive content should be prioritized throughout the research process. Transparent data collection practices, coupled with efforts to mitigate biases, are paramount to ensure the ethical treatment of user-generated content.

Respecting privacy rights involves anonymizing and securing user information, especially when dealing with publicly available but personally identifiable data. Seeking consent for data usage is essential, and researchers should be transparent about the purpose and scope of their analysis when working with comments from public platforms. Sensitivity to cultural nuances and potential biases in sentiment analysis algorithms is crucial to avoid perpetuating stereotypes or misrepresenting sentiments across diverse communities.

Fairness should be a guiding principle in sentiment analysis, ensuring that the models do not disproportionately favor or disadvantage any particular group. Acknowledging the broader societal implications of sentiment analysis on YouTube comments is vital, as the findings may impact public opinion, shape narratives, and influence decision-making. Upholding ethical standards in this research involves navigating privacy, consent, sensi-

tivity, fairness, and cultural considerations to contribute responsibly to the evolving field of sentiment analysis on social media platforms.

References

Naomi S Altman. 1992. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician*, 46(3):175–185.

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, Sajeetha Thavareesan, Dhivya Chinnappa, Durairaj Thenmozhi, Elizabeth Sherly, John P McCrae, Adeep Hande, Rahul Ponnusamy, Shubhanker Banerjee, et al. 2021. Findings of the Sentiment Analysis of Dravidian Languages in Code-Mixed Text. *arXiv preprint arXiv:2111.09811*.

Corinna Cortes and Vladimir Vapnik. 1995. Support-Vector Networks. *Machine learning*, 20:273–297.

Zulfadzli Drus and Haliyana Khalid. 2019. Sentiment analysis in social media and its application: Systematic literature review. *Procedia Computer Science*, 161:707–714.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *arXiv preprint arXiv:2007.01852*.

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, Durairaj Thenmozhi, 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.

Pawan Kalyan Jada, D Sashidhar Reddy, Konthala Yasaswini, Arunaggiri Pandian K, Prabakaran Chandran, Anbukkarasi Sampath, and Sathiyaraj Thangasamy. 2021. Transformer based Sentiment Analysis in Dravidian Languages. In *FIRE (Working Notes)*, pages 926–938.

Irina Rish et al. 2001. An empirical study of the Naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46.

Lavanya S. K., Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rajeswari Natarajan, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Durairaj Thenmozhi, 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.

Wei-Yao Wang, Yu-Chien Tang, Wei-Wei Du, and Wen-Chih Peng. 2022. NYCU_TWD@ LT-EDI-ACL2022: Ensemble models with VADER and contrastive learning for detecting signs of depression from social media. In *Proceedings of the second workshop on language technology for equality, diversity and inclusion*, pages 136–139.