Transformers@DravidianLangTech-EACL2024: Sentiment Analysis of Code-Mixed Tamil Using RoBERTa

Kriti Singhal, Jatin Bedi
Computer Science and Engineering Department
Thapar Institute of Engineering and Technology
kritisinghal711@gmail.com, jatin.bedi@thapar.edu

Abstract

Sentiment analysis has been an active field of research for over 20 years and has gained immense popularity due to its applications in both academia and industry. Sentiment Analysis of code-mixed posts and comments on social media, especially in Dravidian languages, is gaining more and more traction. This paper describes the team Transformers' submission to the Sentiment Analysis in Tamil shared task organized by DravidianLangTech 2024 workshop at EACL 2024. A BERT-based architecture, RoBERTa was used for the shared task. The best macro average F1-score achieved was 0.212. We secured the 5th rank in the Sentiment Analysis shared task in Tamil.

1 Introduction

Sentiment analysis can be defined as the task of classifying text on the basis of the subjective ideas presented in it. The shared task[1] facilitated by DravidianLangTech aimed to identify the sentiment polarity of code-mixed dataset of comments and posts in Tamil-English from various social media platforms (S. K. et al., 2024).

With millions of people across the world gaining access to the internet, freedom of speech and expression now knows no borders. Posts and comments shared on social media platforms like Twitter and YouTube can be accessed by anyone, anywhere in the world, in just a few milliseconds (Shanmugavadivel et al., 2022). The amount of user-generated content available online has set new records. Many scholars have, hence, directed their efforts to identify the sentiments expressed in the content shared online (Yue et al., 2019).

Tamil holds the status of being one of the twenty-two scheduled languages recognized by the Constitution of India (Ghanghor et al., 2021b). Tamil is also is a part of the Dravidian languages' (Chakravarthi and Raja, 2020), dating back over 4500 years. However, Tamil remains under-resourced (Ghanghor et al., 2021a). Most of the resources available for Tamil are code-mixed in nature, i.e., the text comprises of different languages.

Recent advances in the field of Natural Language Processing (NLP) have helped overcome many of the challenges presented by long texts, under-resourced languages and code-mixed data. Some of these include Long Short Term Memory (Hochreiter and Schmidhuber, 1997) and Gated Recurrent Units (Chung et al., 2014). But transformers (Vaswani et al., 2017), have helped researchers reach new heights which was not possible earlier.

In this paper, we discuss our use of a transformer-based model, RoBERTa in the shared task of Sentiment Analysis in Tamil organized by DravidianLangTech at EACL 2024.

2 Related Work

In the past multiple researchers have proposed various approaches for sentiment analysis. Special efforts have been directed towards performing sentiment analysis on code-mixed and under-resourced languages such as Tamil.

Varsha et al. (2022) experimented with different tokenizers on various models, including Random Forest, Support Vector Machine, Adaboost, etc., on Tamil-English, Malayalam-English, and Kannada-English. They also tested how different feature extraction techniques, such as Count Vectorizer, TF-IDF, XLM feature extraction, etc., affected the performance of these models. They found that for Tamil-English data, the Count Vectorizer with the Random Forest model gave the best performance and achieved an F1-score of 0.61.

A 6-layer deep learning model was proposed by Ugursandi and Anand Kumar (2022). The first layer was an embedding layer, which used one-hot

[1] https://codalab.lisn.upsaclay.fr/competitions/16088

Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 151–155
March 22, 2024 ©2024 Association for Computational Linguistics

Table 1: Dataset Distribution for Sentiment Analysis Task

Dataset	**Label**				**Total**
	Positive	**unknown_state**	**Negative**	**Mixed_feelings**	
Dev	2257	611	480	438	3768
Train	20070	5628	4271	4020	33989

encoding followed by a convolutional layer, which created a new vector over a specific geographic dimension. The third layer was a Max Pooling layer, which returned the maximum values for each feature in the vector returned by the convolutional layer. This was followed by a dropout layer to eliminate the contribution from some of the neurons in the subsequent dense layer.

Three Bidirectional Long Short Term Memory (Bi-LSTM) networks were concatenated together for feature extraction in the approach adopted by Mishra et al. (2021) for sentiment analysis in Dravidian languages. They found that on the Tamil dataset, the performance of the traditional machine learning classifiers was not at par with the deep learning approaches. A hybrid model of word2vec, random word embedding, and random char embeddings with three parallel BiLSTM models gave the best weighted F1 of 0.55.

Jada et al. (2021) used a soft voting approach from the results derived from various transformer models. They used multiple pre-trained models including MuRIL (Khanuja et al., 2021), mBERT (Devlin et al., 2019), DistilmBERT (Sanh et al., 2019) and XLM Roberta (Conneau et al., 2019). After obtaining the prediction from all the models, soft voting was performed by taking the weighted average for each class label and assigning the label with the highest probability. This approach achieved an F1 score of 0.626 on the Tamil text.

3 Dataset Description

The code-mixed Tamil dataset was provided by the organizers of the shared task (Chakravarthi et al., 2020; Hegde et al., 2022, 2023). The train and dev dataset comprised of three columns: id, text, and label. The test set comprised of only columns, i.e., id and text. The distributions of the dev and train datasets have been shown in the table 1.

The labels provided for the text were, 'Positive', 'unknown_state', 'Negative', and 'Mixed_feelings'. These labels were assigned to the text based on the polarity of sentiment expressed in the comment or post.

4 Methodology

Sentiment analysis is a text classification problem. This is one of the most important problems in NLP that researches are working on. Text classification can be described as a task where the given texts need to be categorized based upon context. Sentiment analysis makes use of the sentiment polarity to determine what is the sentiment expressed in a given piece of text.

After concatenating the dev and the train dataset, the procedure shown in 1 was used to fine-tune RoBERTA. Then the fine-tuned RoBERTa was used to classify the sentiment of an unseen text into one of the four possible classes, which are 'Positive', 'unknown_state', 'Negative', and 'Mixed_feelings', as represented in Figure 2.

4.1 Data Preprocessing

The data provided in the dataset had been collected from comments and posts on social media. Naturally, there was use of emojis, numbers, and other special characters. Emojis, numbers, and other special characters usually do not convey much about the sentiment and were hence removed from the text.

Table 1 shows the number of comments or posts for the various classes. It was observed that there was a data imbalance problem in the dataset, i.e., the number of samples for one class was much greater than the number of samples for another class. To tackle the issue of data imbalance, undersampling was performed on the data to randomly select samples from all the classes such that the number of samples for all the classes is same. Since 'Mixed_feelings' has the least number of samples, random sampling was performed on 'Positive', 'unknown_state', and 'Negative' to select samples such that the total number of samples from each class were equal. After performing undersampling, 4458 samples were present for each class in the dataset.

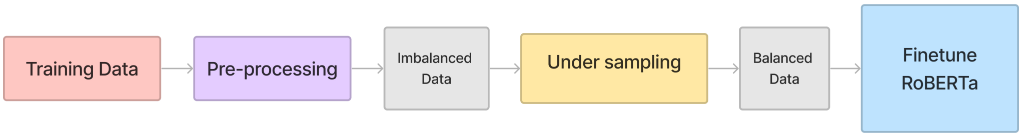

Figure 1: Proposed Methodology

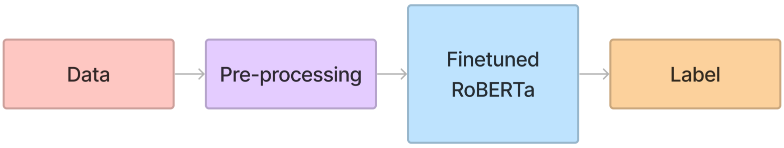

Figure 2: Label Generation for Unseen Data

4.2 Model Building

The XLM RoBERTa model is a transformer model trained using the unsupervised learning approach. It was based on the 2019 RoBERTa architecture by Facebook. This is a large multi-lingual language model that was trained on 2.5TB of data obtained from CommonCrawl after filtering. The model was trained for 100 different languages and then fine-tuned for various downstream tasks like sentiment analysis.

After preprocessing was completed, the text was tokenized using the XLM RoBERTA (Conneau et al., 2019) Tokenizer. The tokenized text was then used to fine-tune an XLM RoBERTA Large model, as shown in Figure 1.

After performing tokenization, the sentences were padded to the maximum length during the encoding procedure, where the maximum length was chosen as 512. Truncation was performed if the length exceeded the maximum limit of 512. The encoded sentences were then passed through the XLM Roberta Large to fine-tune the model. The performance of the model was tested from 5 to 40 epochs while performing hyperparameter tuning. Adam optimizer and cross entropy loss were chosen as the optimizer and the loss function, respectively. The highest performance was achieved at 20 epochs.

After the model was fine-tuned, the unseen or the test data was passed to the model to predict the labels as illustrated in Figure 2.

5 Results and Discussion

A transformer-based approach, XLM RoBERTa was discussed to perform sentiment analysis on code-mixed Tamil posts and comments.

The data imbalance issue was addressed by undersampling the majority class randomly. This was followed by text pre-processing to remove any special symbols, numbers, and emojis. The pre-processed text was used for fine-tuning various transformer based models for performing sentiment analysis.

At 20 epochs, the performance of the XLM RoBERTa model gave the highest F1 score compared to the other transformer-based models. The proposed methodology achieved an F1 score of 0.212.

6 Conclusion and Future Work

Sentiment analysis is the process of classifying text based on the subjective ideas it represents. The shared task by DravidianLangTech at EACL 2024 was focused on finding the sentiment of code-mixed dataset of comments and posts in Tamil-English on different social media platforms.

In this paper, we discussed our use of a BERT-based architecture, XML RoBERTa, in the Sentiment Analysis in Tamil shared task. We achieved a highest F1-score of 0.212 with the discussed approach.

Ensembling techniques using different multilingual transformers such as IndicBert and other deep learning-based techniques may help further improve the performance. Also, since Tamil is an under-resourced language, fine-tuning the model on different datasets may give better results.

References

Asoka Chakravarthi and Bharathi Raja. 2020. *Leveraging orthographic information to improve machine translation of under-resourced languages*. Ph.D. thesis, NUI Galway.

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.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nikhil Ghanghor, Parameswari Krishnamurthy, Sajeetha Thavareesan, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2021a. IIITK@DravidianLangTech-EACL2021: Offensive language identification and meme classification in Tamil, Malayalam and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 222–229, Kyiv. Association for Computational Linguistics.

Nikhil Ghanghor, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021b. IIITK@LT-EDI-EACL2021: Hope speech detection for equality, diversity, and inclusion in Tamil , Malayalam and English. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 197–203, Kyiv. 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, Marseille, France. European Language Resources Association.

Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rahul Ponnusamy, SUBALALITHA CN, Lavanya S K, Thenmozhi D, 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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

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.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages.

Ankit Kumar Mishra, Sunil Saumya, and Abhinav Kumar. 2021. Sentiment analysis of dravidian-codemix language. In *Working Notes of FIRE 2021-Forum for Information Retrieval Evaluation (Online). CEUR*.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

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.

Sushil Ugursandi and M Anand Kumar. 2022. Sentiment analysis and homophobia detection of youtube comments. In *Working Notes of FIRE 2022-Forum for Information Retrieval Evaluation (Hybrid). CEUR*.

Josephine Varsha, B Bharathi, and A Meenakshi. 2022. Sentiment analysis and homophobia detection of youtube comments in code-mixed dravidian languages using machine learning and transformer models. In *Working Notes of FIRE 2022-Forum for Information Retrieval Evaluation (Hybrid). CEUR*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Lin Yue, Weitong Chen, Xue Li, Wanli Zuo, and Minghao Yin. 2019. A survey of sentiment analysis in social media. *Knowledge and Information Systems*, 60:617–663.