Sentiment Analysis on Dravidian Code-Mixed YouTube Comments using Paraphrase XLM-RoBERTa Model

Yandrapati Prakash Babu, Rajagopal Eswari

Department of Computer Applications, National Institute of Technology, Tiruchirappalli, India.

Abstract

In recent days, social media users are drastically increasing, and they are very interested in participating in discussions and expressing their feelings in the form of comments. Most of the users use their native language, which is written in English(Code-Mixed Language). But the existing sentiment classification models can analyze the text sentiment if it is in English vocabulary or the script is in the native language. If the YouTube comments are in the code-mixed language, existing methodologies' performance is not promising. To solve this classification problem, we use the Paraphrase XLM-RoBERTa model. We train the model on Tamil, Malayalam, and Kannada code-Mixed language datasets, and achieve F1-scores of 71.1, 75.3, and 62.5 respectively. Our team ranks first, second and third on Tamil, Malayalam, and Kannada code-Mixed language datasets.

Keywords

XLM-RoBERTa, Paraphrase, Code Mixed, Manglish, Sentiment Analysis

1. Introduction

Nowadays, most people use the internet and express their opinions on social media platforms, blogs, e-commerce websites, health care [1] platforms, etc. India is one of the multilingual country that has 22 officially recognised languages, but according to the 2001 census report, 122 major languages and 1591 other languages were used by the Indians. People are willing to share their views in their native language, which is sometimes written in English script. Due to this reason, more research is needed to find the sentiment of code-mixing languages. In South India, significant languages are Tamil, Telugu, Malayalam, and Kannada. The Dravidian Code-mixed shared task 2021 organizers created the Tamil-English, Malayalam-English, and Kannada-English datasets[2, 3].

According to Solorio et al.[4] code-mixing is the word-level alternation of languages that often occurs by fusing words from one language with the rules of another. Words from several languages can be found in code-mixed languages. The emphasis here is solely on Code-mixed bilingual language [5]. According to Myers et al.[6] code-mixing (CM) is the process of combining an utterance of another language with linguistic units from one language, such as sentences, words, and morphemes. In a multilingual society, code-mixing is quite prevalent, and code-mixed writings are frequently produced in non-native scripts [7]. When composing the text,

FIRE 2021: Forum for Information Retrieval Evaluation, December 13-17, 2021, India.
prakash.babu23@gmail.com (Y. Prakash Babu); eswari@nitt.edu (R. Eswari)

language mixing, also known as code-mixing, occurs.
Natural language processing (NLP) is a cutting-edge technology that gives computers with the information they need to understand the languages we speak. Syntax analysis (grammatical rules) and semantic analysis are both parts of NLP. Sentiment analysis is a categorization approach that offers sentiments about a subject collectively. Sentiment analysis may be performed at the sentence, document, aspect, and phrase levels. Sentiment Analysis is a term that is frequently used to characterize a person's emotional state. To the best of our knowledge, no study on Manglish Corpora in sentiment analysis has been found. The shared task organizers produced Malayalam-English [2], Tamil-English [8] and Kannada-English [9] datasets, and they thoroughly detailed how they obtained and categorized the YouTube comments in the datasets [10]. Following the recent trend of using transformer-based pretrained language models for NLP tasks [11], our proposed system makes use of multilingual Sentence BERT model based on XLM-RoBERTa model[1] [12] for sentiment analysis of code-mixed youtube comments [13].

2. Related work

People are using code-mixed languages in online platforms which motivate the researchers to focus on sentiment analysis on code-mixed languages. Chanda et al.[14] applied the pre-trained models like BERT, DistilBERT, and fastText. Dowlagar et al. [15] used the meta embedding transformer model by using GRU and fastText deep learning models. Code-mixed languages are the combination of multiple languages Huang et al. [16] proposed the Multilingual Code Mixing Text with M-BERT and XLM-RoBERTa. Kalaivani et al. [17] employed the ULMFiT framework with AWD-LSTM model using the FastAi library dealing with the sentiment in the YouTube comments. Prakash et al. [18] combined the Malayalam sentiment features with SBERT model[19] output features to find the sentiment in the dataset and the dataset imbalance problem is solved using ClassBalancedLoss function. Lakshmanan et al.[20] proposed models based on Stochastic Gradient Descent and Logistic Regression and Soundex features for code-mixed text.

3. Methodology

3.1. Data and Pre-Processing

The provided datasets[2] are the collection of YouTube comments, and these YouTube comments are in five categories(positive, negative, {not-Tamil, not-Malayalam, and not-Kannada}, unknown_state and mixed_feelings). The statistics of the datasets are tabulated in Table 1 and class-wise statistics are tabulated in the Table 2. The datasets contain noisy text. So, we use pre-processing techniques before giving to the model. The pre-processing steps are as follows.

- removal of special characters and symbols.
- removal of repeating continuous characters in the word.
- replacing the emoticons with the suitable words.

[1]https://huggingface.co/sentence-transformers/paraphrase-xlm-r-multilingual-v1
[2]https://dravidian-codemix.github.io/2021/datasets.html

- removal of continuous words and sentences in the YouTube comment.

Datasets	Training	Validation	Test	Total
Tamil-English	35,657	3,963	4,403	44,023
Malayalam-English	15,889	1,767	1,963	19,619
Kannada-English	6213	692	768	7,673

Table 1: Statistics of Training, Validation and Test datasets

Datasets	Labels	Training	Validation	Test
Tamil-English	Positive	20070	2257	2546
	Negative	4271	480	477
	Not-Tamil	1667	176	244
	Unknown_state	5628	611	665
	Mixed_feelings	4020	438	470
Malayalam-English	Positive	6421	706	780
	Negative	2105	237	258
	Not-malayalam	1157	141	147
	Unknown_state	5279	580	643
	Mixed_feelings	926	102	134
Kannada-English	Positive	2823	321	374
	Negative	1188	139	157
	Not-Kannada	916	110	110
	Unknown_state	711	69	62
	Mixed_feelings	574	52	65

Table 2: Class-wise statistics of Training, Validation and Test datasets

3.2. Model Description

Our approach is based on Paraphrase XLM-RoBERTa model which is a multilingual sentence-transformers model. XLM-R [12] is a multilingual model obtained by pretraining on monolingual crawled data of more than 100 languages. Paraphrase XLM-RoBERTa model is obtained by distilling knowledge from Paraphrase-DistilRoBERTa model to XLM-RoBERTa model using more than parallel data from 50+ languages [19, 21]. For fine-tuning the model, following Devlin et al.[22] we consider the final hidden vector of the first special token as the aggregate input sentence representation and then pass them onto softmax layer to get the predictions.

4. Implementation Details

The Paraphrase XLM-RoBERTa model is used in this work and to train the datasets. The Paraphrase XLMRoBERTa model's hyperparameters are set as epochs=12, learning rate=3e-5, batch size=16, and dropout=0.5. The model is built with PyTorch's transformers library [23]. The

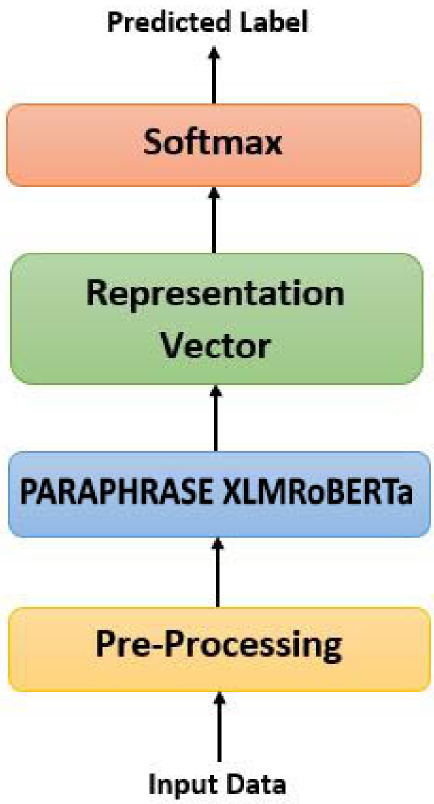


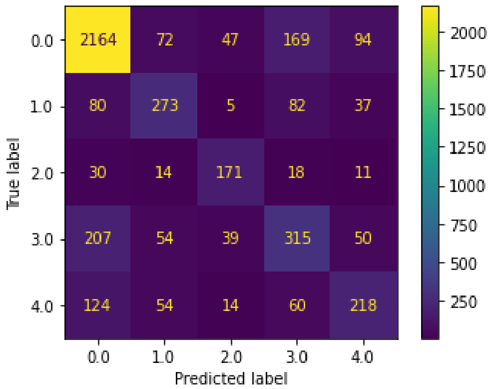
Figure 1: Overview of the Model used in this proposed work

implementation code is accessible on GitHub. [3].

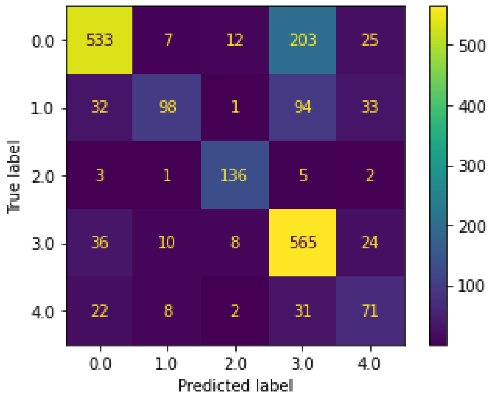
5. Results

We report precision, recall and F1-score on three datasets are shown in Table 3. The label wise precision, recall and F1-scores for Tamil-English, Malayalam-English and Kannda-English are reported in Table 4, Table 5 and Table 6 respectively. From the Tables 4,5,6, we can observe that F1-score is least for 'Mixed_feelings' instances in all the three datasets. In figures 2(a),2(b) and 2(c) dataset wise confusion matrices are given for better understanding of model predictions for the three datasets. Labels are represented as (0-Positive, 1-Negative, 2-not intended language, 3-unknown_state and 4-Mixed_feelings).

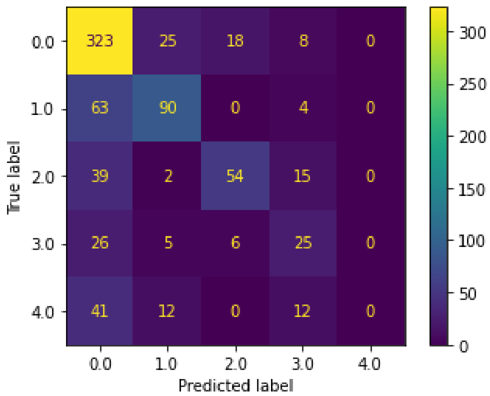
[3]https://github.com/prakashbabuy/manglish2021/



(a) Confusion Matrix for Tamil-English



(b) Confusion Matrix for Malayalam-English



(c) Confusion Matrix for Kannada-Englsih

Figure 2: Class-wise performance of the Model used in this proposed work

Dataset	**Precision**	**Recall**	**F1-Score**
Tamil-English	70.9	71.4	71.1
Malayalam-English	75.3	75.5	75.3
Kannada-English	62.7	65.5	62.5

Table 3: Precision, Recall and F1-score of evaluation sets

Label	Precision	Recall	F1-score
Positive	85.0	68.0	76.0
Negative	79.0	38.0	51.0
not-malayalam	86.0	93.0	89.0
unknown_state	63.0	88.0	73.0
Mixed_feelings	46.0	53.0	49.0

Table 4: Class-wise Precision, Recall and F1-score of Tamil dataset

Label	Precision	Recall	F1-score
Positive	83.0	85.0	84.0
Negative	58.0	57.0	58.0
not-Tamil	62.0	70.0	66.0
unknown_state	49.0	47.0	48.0
Mixed_feelings	53.0	46.0	50.0

Table 5: Class-wise Precision, Recall and F1-score of Malayalam dataset

Label	precision	Recall	F1-score
Positive	69.0	82.0	75.0
Negative	67.0	57.0	62.0
not-Kannada	69.0	49.0	57.0
unknown_state	39.0	40.0	40.0
Mixed_feelings	28.0	20.0	23.0

Table 6: Class-wise Precision, Recall and F1-score of Kannada dataset

6. Conclusion

This paper presents the system using Paraphrase XLM-RoBERTa model to identify the sentiment of Code-Mixed Tamil-English, Malayalam-English and Kannada-English YouTube comments. This shared task is treated as classification problem. The model based on Paraphrase XLM-RoBERTa model achieved promising results with the F1-score of Tamil-English->71.1, Malayalam-English->75.3, and Kannada-English->62.5. In future we will improve the model performance by identifying the sarcastic code-mixed YouTube comments.

References

[1] K. S. Kalyan, A. Rajasekharan, S. Sangeetha, AMMU - A survey of transformer-based biomedical pretrained language models, CoRR abs/2105.00827 (2021). URL: https://arxiv.

org/abs/2105.00827. arXiv:2105.00827.

[2] B. R. Chakravarthi, N. Jose, S. Suryawanshi, E. Sherly, J. P. McCrae, A sentiment analysis dataset for code-mixed Malayalam-English, 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), European Language Resources association, Marseille, France, 2020, pp. 177–184. URL: https://www.aclweb.org/anthology/2020.sltu-1.25.

[3] R. Priyadharshini, B. R. Chakravarthi, S. Thavareesan, D. Chinnappa, D. Thenmozhi, R. Ponnusamy, Overview of the DravidianCodeMix 2021 Shared Task on Sentiment Detection in Tamil, Malayalam, and Kannada, in: Forum for Information Retrieval Evaluation, FIRE 2021, Association for Computing Machinery, 2021.

[4] T. Solorio, Y. Liu, Learning to predict code-switching points, in: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, 2008, pp. 973–981.

[5] N. Jose, B. R. Chakravarthi, S. Suryawanshi, E. Sherly, J. P. McCrae, A survey of current datasets for code-switching research, in: 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020.

[6] C. Myers-Scotton, Duelling languages: Grammatical structure in codeswitching, Oxford University Press, 1997.

[7] B. R. Chakravarthi, R. Priyadharshini, B. Stearns, A. Jayapal, S. S, M. Arcan, M. Zarrouk, J. P. McCrae, Multilingual multimodal machine translation for Dravidian languages utilizing phonetic transcription, in: Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages, European Association for Machine Translation, Dublin, Ireland, 2019, pp. 56–63. URL: https://www.aclweb.org/anthology/W19-6809.

[8] B. R. Chakravarthi, V. Muralidaran, R. Priyadharshini, J. P. McCrae, 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), European Language Resources association, Marseille, France, 2020, pp. 202–210. URL: https://www.aclweb.org/anthology/2020.sltu-1.28.

[9] A. Hande, R. Priyadharshini, B. R. Chakravarthi, KanCMD: Kannada CodeMixed dataset for sentiment analysis and offensive language detection, in: Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media, Association for Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 54–63. URL: https://www.aclweb.org/anthology/2020.peoples-1.6.

[10] B. R. Chakravarthi, R. Priyadharshini, V. Muralidaran, N. Jose, S. Suryawanshi, E. Sherly, J. P. McCrae, Dravidiancodemix: Sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text, CoRR abs/2106.09460 (2021). URL: https://arxiv.org/abs/2106.09460. arXiv:2106.09460.

[11] K. S. Kalyan, A. Rajasekharan, S. Sangeetha, AMMUS : A survey of transformer-based pretrained models in natural language processing, CoRR abs/2108.05542 (2021). URL: https://arxiv.org/abs/2108.05542. arXiv:2108.05542.

[12] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116.

arXiv:1911.02116.
[13] B. R. Chakravarthi, R. Priyadharshini, S. Thavareesan, D. Chinnappa, D. Thenmozhi, E. Sherly, J. P. McCrae, A. Hande, R. Ponnusamy, S. Banerjee, C. Vasantharajan, Findings of the Sentiment Analysis of Dravidian Languages in Code-Mixed Text, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.
[14] S. Chanda, S. Pal, Irlab@ iitbhu@ dravidian-codemix-fire2020: Sentiment analysis for dravidian languages in code-mixed text., in: FIRE (Working Notes), 2020, pp. 535–540.
[15] S. Dowlagar, R. Mamidi, Cmsaone@ dravidian-codemix-fire2020: A meta embedding and transformer model for code-mixed sentiment analysis on social media text, arXiv preprint arXiv:2101.09004 (2021).
[16] B. Huang, Y. Bai, Lucashub@ dravidian-codemix-fire2020: Sentiment analysis on multilingual code mixing text with m-bert and xlm-roberta., in: FIRE (Working Notes), 2020, pp. 574–581.
[17] A. Kalaivani, D. Thenmozhi, Ssn_nlp_mlrg@ dravidian-codemix-fire2020: Sentiment code-mixed text classification in tamil and malayalam using ulmfit., in: FIRE (Working Notes), 2020, pp. 528–534.
[18] Y. P. Babu, R. Eswari, K. Nimmi, Cia_nitt@ dravidian-codemix-fire2020: Malayalam-english code mixed sentiment analysis using sentence bert and sentiment features., in: FIRE (Working Notes), 2020, pp. 566–573.
[19] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. URL: https://www.aclweb.org/anthology/D19-1410. doi:10.18653/v1/D19-1410.
[20] B. Lakshmanan, S. K. Ravindranath, Theedhum nandrum@dravidian-codemix-fire2020: A sentiment polarity classifier for youtube commentswith code-switching between tamil, malayalam and english, CoRR abs/2010.03189 (2020). URL: https://arxiv.org/abs/2010.03189. arXiv:2010.03189.
[21] N. Reimers, I. Gurevych, Making monolingual sentence embeddings multilingual using knowledge distillation, CoRR abs/2004.09813 (2020). URL: https://arxiv.org/abs/2004.09813. arXiv:2004.09813.
[22] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.
[23] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Transformers: State-of-the-art natural language processing, arXiv preprint arXiv:1910.03771 (2019).