Multilingual Sentiment Analysis in Tamil, Malayalam, and Kannada code-mixed social media posts using MBERT

Adaikkan Kalaivani[1,2], Durairaj Thenmozhi[3]

[1] *Department of Information and Communication Engineering, Anna University, Chennai*
[2] *Research Centre, Department of CSE, Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam, TamilNadu*
[3] *Department of CSE, Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam, TamilNadu*

Abstract

This paper presents the submitted runs to Dravidian-CodeMix-FIRE2021: Sentiment Analysis for Dravidian Languages in Code-Mixed Text. The identification of sentiment polarity in code-mixed text from social media has paid much attention in recent studies. Moreover, the sentiment analysis in multilingual posts moves forward in the field of natural language processing multilingual community. We have participated in Tamil-English, Malayalam-English, and Kannada-English languages. The shared task of sentiment analysis is the message-level sentiment polarity text classification from the social media posts. We have adapted and fine-tuned the pre-trained Multilingual BERT models for the three languages. We applied the adaptive transliteration and translation technique to enrich the training data for the three languages. Our team SSN_NLP_MLRG achieved the F1-scores of 0.603, 0.698, and 0.595 in the shared task for the Tamil, Malayalam, and Kannada code-mixed languages, respectively.

Keywords

Dravidian language, Code-mixed text classification, Transfer learning, Sentiment Analysis

1. Introduction

In recent years, there is a continuous growth of online user posts in social network forums such as Twitter, Facebook, YouTube, Instagram, etc. Social Media forums will be the biggest sources of data available largely in the upcoming years. Sentimental analysis has received much attention in the research community recently. So, the sentiment analyses of social media posts are very important to regularize them. Sentiment analysis is the task of determining the polarity, subjective opinion, target, and valence and of classifying the sentiments in the given code-mixed text [1].

Tamil is the Dravidian language that is officially spoken by the state of Tamil Nadu in India, Sri Lanka, and Singapore. Malayalam and Kannada are the Dravidian languages spoken by the state of Kerala and the state of Karnataka in India. Code-mixing is a phenomenon where native speakers switch between two or more languages and are also written in Roman script in a single utterance [2]. So, analysis of the Code-mixed bilingual [3] or multilingual post [4] from

FIRE 2021: Forum for Information Retrieval Evaluation, December 13-17,2021, India.

kalaiwind@gmail.com, kalaivania@ssn.edu.in, (A. Kalaivani); theni_d@ssn.edu.in (D. Thenmozhi)
0000-0002-1497-5605 (A. Kalaivani); 0000-0003-0681-6628 (D. Thenmozhi)

online social media plays a crucial role in the recent research community. The identification of sentiments [5] in indirect comments like sarcasm [6], metaphors are challenging to annotate manually. Therefore, Automatic identification of sentiments in various multilingual languages is a challenging task.

The Dravidian-CodeMix-FIRE2021 [7] competition aims to build systems capable of identifying sentiment polarity in social networks forum for the Tamil, Malayalam, and Kannada code-mixed languages [8]. The Dravidian-CodeMix-FIRE2020 [9] organizers defined the shared task of identifying sentiments for the Tamil [10], Malayalam code-mixed languages [11]. Furthermore, we classify whether the post into positive, negative, neutral, mixed emotions or not in the intended language [12]. This year, Dravidian-CodeMix-FIRE2021 offers datasets with three languages include Tamil-English, Malayalam-English, and Kannada-English.

This article presents our approaches to Dravidian-CodeMix-FIRE2021. We have participated in the shared task for the three languages. We performed selective transliteration and translation for these languages. We used the NLTK library for pre-processing the training and test data for all languages. The goal of the shared task is to determine and categorize if a message is positive, negative, unknown state, mixed feelings, or not in the intended language. We adapt and fine-tune Multilingual BERT (MBERT) pre-trained model with ktrain library for the Tamil-English, Malayalam-English, and Kannada-English languages[1]. The outline of the paper is as follows. Section 2 reviews the work related to sentiment analysis. Section 3 presents the detailed data description and methodology of our model. In section 4, we analyze the experiment results. Finally, Section 5 discusses the conclusion of our work and further improvement.

2. Related Work

The researchers used the Bi-LSTM's model to determine the sentiments in the Hindi-English posts [13]. The goal of the shared task of sentiment analysis is to identify the sentiment scores and achieved the best results in the Indian languages (SAIL). The sentiments identified in the code-mixed multilingual language are analyzed based on the linguistic code-switching, domain-specific and grammatical transition for the Hindi and English languages [14, 15].

The author used the fastText word embedding, doc2vec features, SVM Classifier, bi-LSTM model, and Conventional neural network (CNN) to classify sentiments in the Bengali-English, Hindi-English code-mixed test corpus [16]. The sub-word level LSTM architecture is used to analyze the sentiments from Hindi-English code-mixed language [17]. The neural network architecture is used to analyze the sentiments and build the system over LSTM [18]. They used the sentiment mining approach to classify sentiments in a multilingual environment namely Hindi, Tamil, Telugu, and Bengali languages [19].

From the observation, most of the research is going on multilingual community. Still, we have to face the challenges in the code-mixed bilingual and multilingual languages in the online social media comments, problems in handling the imbalanced dataset in the low resourced languages, and detecting the sentiments from sarcastic comments. These problems open to researchers in the industry and academia in the different native low resourced languages other than high resourced languages. Therefore automation of detecting the sentiments is a

[1] https://github.com/kalaiwind/Dravidian-2021

Table 1
Dravidian-CodeMix-FIRE2021 dataset

Category	Tamil	Malayalam	Kannada
Positive	19892	6238	2644
Negative	4256	2018	1186
Mixed_feelings	3986	847	560
Unknown_state	5589	5003	696
Not	1658	1107	812

crucial task. The organizers of the Dravidian-CodeMix-FIRE2021 provided the resources for the Dravidian languages.

3. Data and Methodology

This section presents the data preprocessing techniques, data descriptions, models experimented with for the Dravidian code-mixed data.

3.1. Data Description

Typically the Dravidian-CodeMix-FIRE2021 dataset offers posts from YouTube social media forums. The shared dataset involves posts written in Tamil, Malayalam, and Kannada code-mixed languages. For Tamil language, the training data size is 39336 posts and the size of the test data is 4402 posts. For Malayalam language, the size of the training data is 16970 posts and the test data size is 1962 posts. For Kannada language, the training data size is 6578 comments and the test data size is 768 posts. Table 1 presents the category-wise description of Tamil, Malayalam, and Kannada languages. The shared task of the sentiment analysis in Dravidian languages is a Multiclass classification task. It aims to build systems that can classify the YouTube comments into a positive, negative, unknown state, mixed feelings or (not-Tamil, not-Malayalam, not-Kannada) not in the intended language.

3.2. Data Preprocessing

The data preprocessing were minimal to make that flexible for all the shared task of the Dravidian languages. We perform data preprocessing by using NLTK[2] for the Tamil, Malayalam, and Kannada languages. First, the duplication in the training data has been clean because it affects the performance. The strings start with @ symbols has cleared because it denoted as the author name or user id. After that, we remove the hashtag, punctuations, URLs, numerals which don't have semantic meaning. Finally, we cleared the emoji's then converted all the upper case English text and Native language in the Roman script into lower case text.

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

Table 2
Validation Results of MBERT model

Language	Precision	Recall	weighted F1
Tamil	0.59	0.60	0.60
Malayalam	0.72	0.72	0.72
Kannada	0.61	0.61	0.61

3.3. Methodology

We used the pre-trained models MBERT[3] (Multilingual Bidirectional Encoder Representations from Transformers) [20] with the ktrain library for the Tamil, Malayalam, and Kannada languages. ktrain is a lightweight wrapper of TensorFlow Keras to help build, deploy, and train neural networks, machine learning models, and deep learning models more accessible. We take 20% of the data from the training set for the validation process.

We used the Multilingual BERT model to build the system and predict the code-mixed comments for all the three languages. Firstly, we set the sequence length as 512 and batch size as 6. We fine-tuned the pre-trained weights to predict the sentiment polarity. The different learning rates like 1e-5, 2e-5, 3e-5, 5e-5, and the epochs as 5, 6, 7, and 10 were analyzed to improve the performance. Finally, we used the learning rate of 2e-5, 2e-5, 2e-5, and the epochs of 7, 10, and 10 for the Tamil, Malayalam, and Kannada languages of the MBERT model. We have performed selectively transliterate the Roman script and translated the other language like English into particular Tamil, Malayalam, and Kannada code-mixed languages using Google API. The validation results of the three languages of the MBERT model are present in Table 2. In the final test results, we got a weighted-average F1-score of 0.603, 0.698, and 0.595 in the shared task for the Tamil, Malayalam, and Kannada code-mixed languages.

4. Results

In this section, we present the evaluation of our model and submitted results for the Tamil, Malayalam, and Kannada code-mixed languages.

4.1. Experimental Results

The evaluation metrics like precision, recall, macro averaged F1-score, and weighted average F1-score are used to analyze the performance of the model. The Dravidian-CodeMix-FIRE2021 organizers provided the test data for the Dravidian languages. Based on the performance of the validation process, we fine-tuned the MBERT model to build the system and predict the sentiment polarity for the Tamil-English, Malayalam-English, and Kannada-English languages. The performance of the test results of the MBERT model for the three languages is presented in Table 3.

For Tamil language, the MBERT model achieved an accuracy of 0.61, and Precision, Recall, and an F1-score for positive comments are 0.74, 0.80, and 0.77 respectively. The Precision, Recall, and

[3] https://github.com/google-research/bert/blob/master/multilingual.md

Table 3
Test Results of MBERT model

Language	Precision	Recall	weighted F1	macro F1
Tamil	0.597	0.613	0.603	0.47
Malayalam	0.698	0.706	0.698	0.64
Kannada	0.601	0.592	0.595	0.48

F1-score for the negative comments are 0.64, 0.51, and 0.57 respectively. The Precision, Recall, and F1-score for the mixed-feeling comments are 0.42, 0.42, and 0.42 respectively. The Precision, Recall, and F1-score for the unknown-state comments are 0.39, 0.29, and 0.33 respectively. The Precision, Recall, and F1-score for the not-Tamil comments are 0.27, 0.24, and 0.26 respectively. Comparatively, the positive comments of the Tamil language perform well because the number of positive comments is more (difference 15,000 comments) than the other comments.

For Malayalam Language, the MBERT model achieved an accuracy of 0.71, and Precision, Recall, and an F1-score for positive comments are 0.73, 0.81, and 0.77 respectively. The Precision, Recall, and F1-score for the negative comments are 0.82, 0.76, and 0.88 respectively. The Precision, Recall, and F1-score for the mixed-feeling comments are 0.70, 0.72, and 0.71 respectively. The Precision, Recall, and F1-score for the unknown-state comments are 0.60, 0.53, and 0.56 respectively. The Precision, Recall, and F1-score for the not-Malayalam comments are 0.54, 0.28, and 0.37 respectively. Comparatively, the negative, positive, not-Malayalam, unknown-state comments of the Malayalam language perform well even though the number of negative comments is less than positive comments because it is a balanced training set.

For Kannada Language, the MBERT model achieved an accuracy of 0.61, and Precision, Recall, and an F1-score for positive comments are 0.71, 0.71, and 0.71 respectively. The Precision, Recall, and F1-score for the negative comments are 0.62, 0.66, and 0.64 respectively. The Precision, Recall, and F1-score for the mixed-feeling comments are 0.26, 0.32, and 0.29 respectively. The Precision, Recall, and F1-score for the unknown-state comments are 0.62, 0.53, and 0.57 respectively. The Precision, Recall, and F1-score for the not-Kannada comments are 0.20, 0.20, and 0.20 respectively. Comparatively, the positive, negative, not-Kannada comments of the Kannada language perform well.

4.2. Submitted Results

We present the results of the evaluation of our submissions for all the three languages. The task organizers provided the evaluation report based on weighted-average F1-scores. Our team SSN_NLP_MLRG submission had a Weighted-average F1-score of 0.603, 0.698, and 0.595 in the shared task for Tamil, Malayalam, and Kannada code-mixed languages respectively. Our team SSN_NLP_MLRG submission got the 8^{th}, 7^{th}, 10^{th}rank in the shared task for Tamil, Malayalam, and Kannada code-mixed languages respectively.

Furthermore, the F1-scores for the three languages have improved when compared with the baseline F1-scores. For further analysis, we represent our results of the MBERT model by using the confusion matrix is shown in the Figure 1 for Tamil language, Figure 2 for the Malayalam language, and Figure 3 for the Kannada language. From the confusion matrix, we observed that

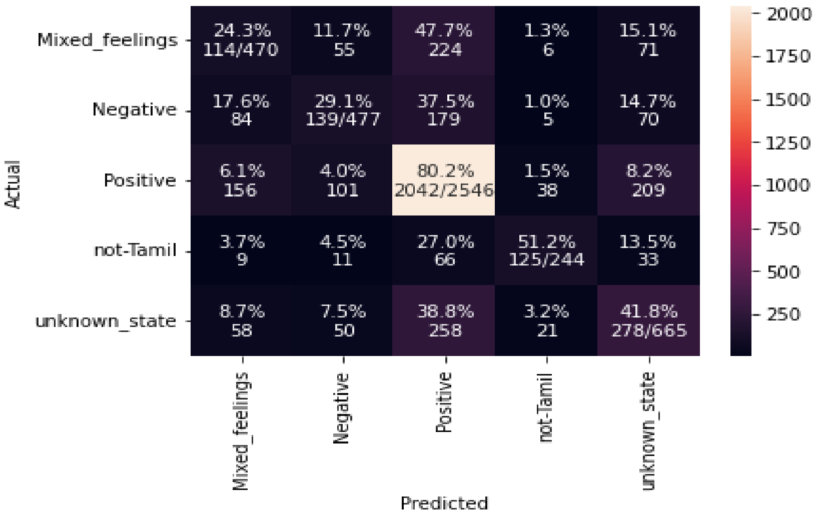

Figure 1: Tamil - Confusion matrix of MBERT

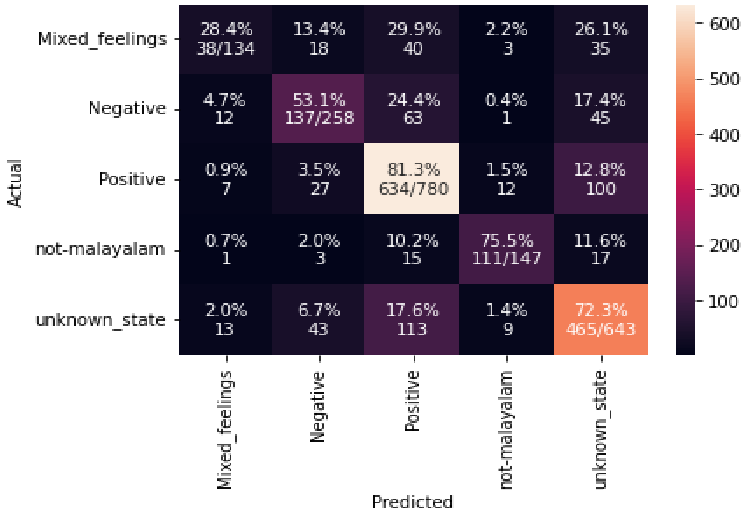

Figure 2: Malayalam - Confusion matrix of MBERT

the positive comments perform well by the MBERT model for all the three languages.

5. Conclusion

This paper presents the methodology for identifying the sentiment polarities from YouTube social media comments in Tamil, Malayalam, and Kannada code-mixed languages. Our team used minimal preprocessing techniques. We experimented with a pre-trained Multilingual BERT transformer model with the variations of inputs for the shared task in three languages.

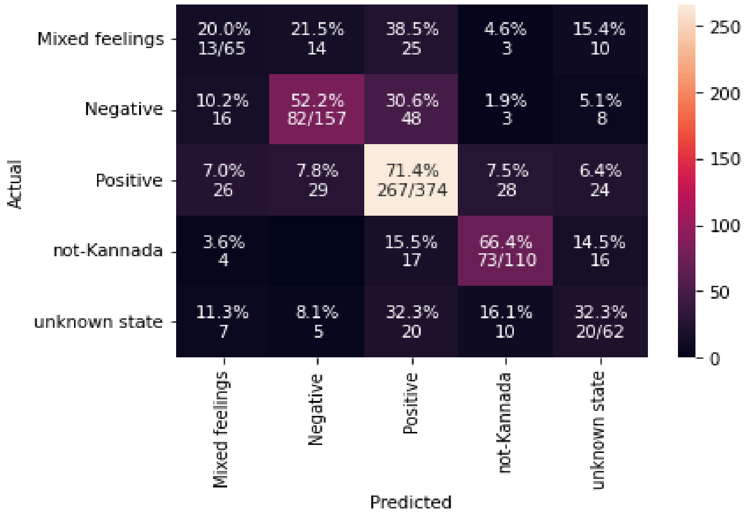

Figure 3: Kannada - Confusion matrix of MBERT

According to evaluation, it is clear that fine-tuning MBERT architecture scores well. Our model MBERT performs well in all the three languages. Our team F1 scores have improved when compared with the baseline F1 scores. Due to non-language-specific preprocessing, we applied adaptive transliteration and translation techniques for better performance in the Tamil, Malayalam, and Kannada code-mixed languages. In future research, we can improve the performance of code-mixed comments by using different deep learning algorithms. Further, we will extend this work to other languages and improve the performance by handling the indirect code-mixed comments to avoid misclassification.

References

[1] A. Kalaivani, D. Thenmozhi, Sentimental analysis using deep learning techniques, International Journal of Recent Technology and Engineering (IJRTE) 7 (2019) 600–606.

[2] 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, pp. 136–141. doi:10.1109/ICACCS48705.2020.9074205.

[3] A. Kalaivani, D. Thenmozhi, SSN_NLP_MLRG@HASOC-FIRE2020: Multilingual Hate Speech and Offensive Content Detection in Indo-European Languages using ALBERT, in: Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020, volume 2826 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 188–194. URL: http://ceur-ws.org/Vol-2826/T2-12.pdf.

[4] A. Kalaivani, D. Thenmozhi, SSN_NLP_MLRG at SemEval-2020 task 12: Offensive language identification in English, Danish, Greek using BERT and machine learning approach, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee

for Computational Linguistics, Barcelona (online), 2020, pp. 2161–2170. URL: https://aclanthology.org/2020.semeval-1.287.

[5] 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.

[6] A. Kalaivani, D. Thenmozhi, Sarcasm identification and detection in conversion context using BERT, in: Proceedings of the Second Workshop on Figurative Language Processing, Association for Computational Linguistics, Online, 2020, pp. 72–76. URL: https://www.aclweb.org/anthology/2020.figlang-1.10. doi:10.18653/v1/2020.figlang-1.10.

[7] R. Priyadharshini, B. R. Chakravarthi, S. Thavareesan, D. Chinnappa, T. Durairaj, E. Sherly, 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.

[8] 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://aclanthology.org/2020.peoples-1.6.

[9] B. R. Chakravarthi, R. Priyadharshini, V. Muralidaran, S. Suryawanshi, N. Jose, E. Sherly, J. P. McCrae, Overview of the track on Sentiment analysis for Dravidian languages in code-mixed text, in: Forum for Information Retrieval Evaluation, FIRE 2020, Association for Computing Machinery, New York, NY, USA, 2020, p. 21–24. URL: https://doi.org/10.1145/3441501.3441515. doi:10.1145/3441501.3441515.

[10] 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.

[11] 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.

[12] 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.

[13] N. Choudhary, R. Singh, I. Bindlish, M. Shrivastava, Sentiment analysis of Code-Mixed languages leveraging resource rich languages, CoRR abs/1804.00806 (2018). URL: http://arxiv.org/abs/1804.00806. arXiv:1804.00806.

[14] A. Pravalika, V. Oza, N. P. Meghana, S. S. Kamath, Domain-specific sentiment analysis approaches for code-mixed social network data, in: 2017 8th International Conference on

Computing, Communication and Networking Technologies (ICCCNT), 2017, pp. 1–6.

[15] B. R. Chakravarthi, Leveraging orthographic information to improve machine translation of under-resourced languages, Ph.D. thesis, NUI Galway, 2020.

[16] K. Shalini, H. B. Ganesh, M. A. Kumar, K. P. Soman, Sentiment analysis for code-mixed indian social media text with distributed representation, in: 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2018, pp. 1126–1131.

[17] A. Joshi, A. Prabhu, M. Shrivastava, V. Varma, Towards sub-word level compositions for Sentiment analysis of Hindi-English code mixed text, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 2482–2491. URL: https://www.aclweb.org/anthology/C16-1234.

[18] Y. K. Lal, V. Kumar, M. Dhar, M. Shrivastava, P. Koehn, De-mixing sentiment from code-mixed text, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Association for Computational Linguistics, Florence, Italy, 2019, pp. 371–377. URL: https://www.aclweb.org/anthology/P19-2052. doi:10.18653/v1/P19-2052.

[19] R. Bhargava, Y. Sharma, S. Sharma, Sentiment analysis for mixed script indic sentences, in: 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2016, pp. 524–529.

[20] A. Kalaivani, D. Thenmozhi, C. Aravindan, SSN_NLP_MLRG@Dravidian-CodeMix-HASOC2021: TOLD: Tamil Offensive Language Detection in Code-Mixed Social Media Comments, in: Forum for Information Retrieval Evaluation, FIRE 2021, 2021.