SSN_NLP_MLRG@Dravidian-CodeMix-FIRE2020: Sentiment Code-Mixed Text Classification in Tamil and Malayalam using ULMFiT

A. Kalaivani[a], D. Thenmozhi[a]

[a]*Department of CSE, SSN College of Engineering, OMR, Kalavakkam, Tamil Nadu 603110*

Abstract

Sentiment analysis is the task of determining the subjective opinion, polarity, target, valence and of detecting and classifying the sentiment in the given text. Code-mixed multilingual language analysis plays a crucial role in research community. This paper describes the shared task of Sentiment Analysis of Dravidian Code-mixed of Tamil-English and Malayalam-English languages to identify the sentiment message polarity from social media comments. We have employed the AWD-LSTM model with ULMFiT framework using the FastAi library dealing with the detection and classification of sentiment from the Dravidian–CodeMix-FIRE2020 Dataset. Our model achieved F1 weighted scores of 0.6 for both the Tamil and Malayalam code-mixed languages for this task respectively.

Keywords

Sentiment Analysis, Code-mixed analysis, Language Modeling, Transfer learning

1. Introduction

There is an increasingly rapid growth of social communication between millions of peoples through internet that shows huge challenges in the social media platforms. Sentiment analysis plays a major role in the field of natural language processing research [1]. Sentiment analysis is the process of identifying the sentiments like emotions, affectionate to others in the given text or sentence or paragraph. Monolingual code-mixed language structure differs from the multilingual code-mixed language due to lack of data inconsistency. Usually, code-mixed texts are written in non-native scripts. Therefore, social media users used roman script for typing the non-native languages [2, 3, 4, 5]. Sentiment multilingual code-mixed language is an important challenge research area in sentiment analysis research. The organizers proposed the shared task of Dravidian–CodeMix-FIRE2020 [6] to classify the sentiment polarity as positive, negative, neutral, mixed emotions, or not in the Tamil-English and Malayalam-English languages. Based on the sentiment analysis, the task is to detect the mixed feelings of online users and to prevent the unusual activities, depression, criminal activities. Much research is going on in this field. The dataset for the Tamil [7] and Malayalam [8] languages was generated from YouTube, Twitter and Facebook.

The shared task of Dravidian–CodeMix-FIRE2020 consists of two tasks. One task is to identify

FIRE 2020: Forum for Information Retrieval Evaluation, December 16–20, 2020, Hyderabad, India

kalaiwind@gmail.com (A. Kalaivani); theni_d@ssn.edu.in (D. Thenmozhi)

the sentiments from the given native code-mixed Tamil-English text. The second task is to identify if the sentiments falls under a gives category from the given code-mixed Malayalam-English language. In the Dravidian–CodeMix-FIRE2020 [9], we have used the pre-trained models for NLP – Universal Language Model Fine-tuning for Text Classification (ULMFiT) framework which can be fine-tuned and categorize the sentiment as positive, negative, neutral, mixed emotions or not from the given comments using the FastAi [1] Library. We focused on the concept of transfer learning that is the state of art of deep learning to build accurate models using the popular FastAi library.

The challenges of this shared task include: a) It is difficult to transfer the code-mixed native languages b) Small datasets are hard to build and train the complex models c) Imbalanced datasets and removing the noisy data d) Social media forum difficulties like ungrammatical sentences, non-vocabulary words, usage of slangs, URLs, native languages written with English in Roman format, words separated as letters, misspelled words. The structure of the paper as follows: we discussed some related work based on multilingual code-mixed sentiment analysis in Section 2, we explained the data description and methodology to build the model in Section 3. We discussed and analyzed the results in Section 4. In Section 5, we concluded the work and discussed about future enhancements.

2. Related Work

Sentiment Analysis of Code-Mixed Text [10] utilizes Siamese networks to map the text and used the twin Bi-LSTM's method to classify the Hindi-English text into sentiment. The main objective of the Shared Task on Sentiment Analysis in Indian Languages (SAIL) Tweets [11] to classify the sentiment in Indian languages and they achieved good accuracy for Hindi, Bengali, and Tamil Languages. The sub-word level LSTM architecture [12] is used to analyzed the sentiments from code-mixed language namely Hindi-English. A language identification and sentiment mining approach [13] enhanced the performance of the sentiment analysis in multilingual Indian languages namely Hindi, Tamil, Telugu and Bengali.

The sentiments of code-mixed multilingual language [14, 15] are analyzed based on Domain specific, linguistic code switching and grammatical transition for the Hindi and English languages. The identification of sarcasm from the user conservation context is done by BERT [16], comparing the results with the machine learning and deep learning approach. The collective and specific encoder [17], built over LSTM to analyze the sentiment at sub-word level to process in neural networks. The sentiment experimentation [18] includes fasttext, doc2vec, SVM Classifier, bi-LSTM and CNN in the Bengali-English, Hindi-English code-mixed test corpus.

3. Data and Methodology

We have used the YouTube dataset given by Dravidian–CodeMix-FIRE2020 shared task. Our team SSN_NLP_MLRG has participated in the Tamil-English and Malayalam-English languages. The dataset contains the Tamil code-mixed language and Malayalam code-mixed language.

[1] https://docs.fast.ai/

Table 1
Tamil and Malayalam Annotated YouTube Comments

Text	Category
thala mass u bgm vera level	Positive
ivara pathta death vadi madiri irukku	Negative
ikka de carter hit aavm maamankam ennathil oru samsyavum vndaaa	Positive
pulimurgan adyam oralde thallu pinne same kattil ottakkulla adi	Mixed_feelings
chandu vine orma vannavar ivde common	Unknown_state

The dataset is given in .tsv format for both the Tamil and Malayalam language with columns named, 'Text' and 'Category' where 'Text' represents the YouTube comments from the social media and 'Category' represents sets of labels which are positive, negative, neutral, mixed emotions, unknown, not Tamil, not Malayalam . We have focused on both the Tamil-English and Malayalam-English languages. The dataset contains inter-sentential switching, intra-sentential switching and tag switching code-mixed sentences. The input of YouTube comments in Tamil-English were written in Roman script with the English Lexicon and Tamil language. The input of YouTube comments of Malayalam-English were written in Roman script with the English Lexicon and Malayalam language as shown in Table 1.

Tamil-English dataset contains totally 15,744 comments. The train data has 11,335, validation data has 1,260 comments and test data has 3,149 comments. Malayalam-English dataset contains 6,739 comments and the train data has 4,717 comments, validation data has 674 comments and test data has 1,348 comments. Table 2 shows the category-wise representations of YouTube comments in the Tamil-English and Malayalam-English dataset. We have employed the ULMFiT framework using the FastAi Library to classify the sentiment. ULMFiT achieves state-of-the-art result in language modeling and transfer learning in field of natural language processing.

We have pre-processed by using NLTK [2] libraries to remove the numerals, punctuation, and replace the noisy strings. We have created the language model with AWD-LSTM (Average-SGD Weight-Dropped LSTM) architecture model for the multi-label text classification to predict the sentiment. The FastAi library provides functions to create classification data bunch and Language model data bunch. In language modeling, The RNN learns about the next word from the previous word. We have set the batch size is 32 for learning and the data for classification. We have got F-score of 0.6 with 3 epochs for both the Tamil and Malayalam languages.

We have used different variations of learning rate to achieve better performance by gradually freezing and unfreezing the weights. We have set the learning rate to 3e-02, 3e-04, 3e-03, 1e-03, 5e-03, 5e-04 and the epochs to 1, 15, 3, 2 and 5 for both the Tamil and Malayalam languages. We have fine-tuned a pre-trained language model. We created a model and downloaded the pre-trained weights and fine-tuned by using the learner object. Figure 1 represents the gradually increase of learning rate for Tamil-English language that improves the performance in that 1 represents the learning before fine- tuning the pre-trained model, 2 represents the unfreezing and adding the weights, 3 represents the gradual freezing after fine-tuning the pre-trained language and classification model. Figure 2 shows that gradually increasing the learning rate

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

Table 2
Dataset - YouTube Comments

Category	Tamil	Malayalam
Positive	8,484	2,246
Negative	1,613	1,505
Mixed_feelings	1,424	707
Unknown_state	677	600
Not-Tamil	397	333

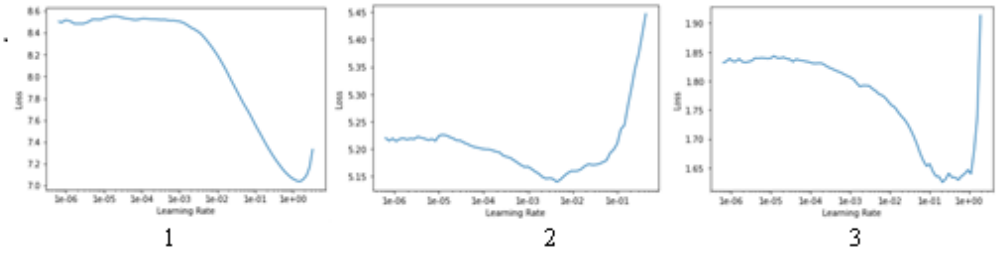

Figure 1: Variations in learning rate – Tamil

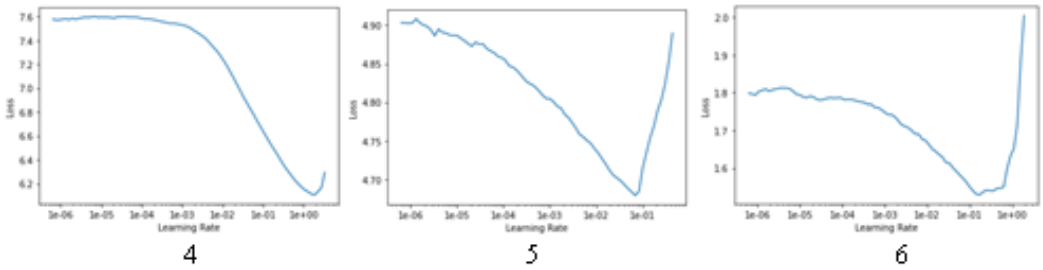

Figure 2: Variations in learning rate – Malayalam

of Malayalam-English language shows improvement in the performance and 4 represents the learning of the language model, 5 represents the unfreezing of the weights of the learning rate, 6 represents the fine-tuning of the pre-trained classification and language model by using gradual freezing. The validation and variation results of our model for the Tamil and Malayalam code-mixed languages are presented in Table 3. Accuracy of Tamil and Malayalam code-mixed language increases and training loss and valid loss decreases gradually. Comparatively, validation loss is less than the training loss. Our team obtained the best scores in both the Tamil-English and Malayalam-English languages by transfer learning and fine-tuning the language and classification model. The source code is available in the GitHub link [3].

[3]https://github.com/kalaiwind/NLP-ML

Table 3
Validation results of Tamil and Malayalam code-mixed Languages

Variations	Accuracy of Tamil	Accuracy of Malayalam
Before Fine-tune	0.255	0.326
unfreeze	0.304	0.369
After Fine-tune	0.694	0.571
Gradual freezing	0.697	0.655

Table 4
Results of Tamil and Malayalam code-mixed Languages

Metrics	Tamil	Malayalam
Precision	0.6	0.61
Recall	0.68	0.61
F-score	0.6	0.6

4. Results

We have evaluated the test data of Dravidian–CodeMix-FIRE2020 shared task for the tasks of Tamil and Malayalam languages. The performance was analyzed using the metrics namely precision, recall, F1 weighted score and accuracy. The results of our model of the Tamil and Malayalam code-mixed languages are presented in Table 4. We have obtained the best results for Tamil task and Malayalam task using ULMFiT framework, AWD-LSTM model, FastAi Library. We have achieved the 6th rank in Tamil code-mixed language task and 12th rank in the Malayalam code-mixed language task. Comparatively, the Tamil code-mixed language performs better than the Malayalam code-mixed language due to the fact that the amount of Tamil data is large than Malayalam. ULMFiT is very effective and improves the performance and accuracy in small datasets. We got F1 weighted scores as 0.6 for both the Tamil and Malayalam code-mixed language task respectively. ULMFiT achieved the state of art in natural language processing.

5. Conclusions

In this paper, we presented a AWD-LSTM to identify the sentiments for Tamil-English and Malayalam-English language, and also used the model ULMFiT framework using the FastAi library to improve the ability of the model, and fine-tune the pre-trained model in NLP and achieved the best results for Tamil-English and Malayalam-English Language. We have obtained the F1 weighted score as 0.6 for both the Tamil and Malayalam Languages. We achieved the 6th rank in Tamil code-mixed language and 12th rank in the Malayalam code-mixed language. In future research, the performance can be improved further by changing the value of parameter. We will consider the handling of data imbalances and try to enhance the performance of the model using different pre-trained models. Further, the performance can be improved by huge datasets.

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] R. Priyadharshini, B. Chakravarthi, M. Vegupatti, J. McCrae, Named entity recognition for code-mixed indian corpus using meta embedding, 2020, pp. 68–72. doi:10.1109/ICACCS48705.2020.9074379.

[4] B. R. Chakravarthi, M. Arcan, J. P. McCrae, Improving Wordnets for Under-Resourced Languages Using Machine Translation information, in: Proceedings of the 9th Global WordNet Conference, Zenodo, 2018. URL: https://doi.org/10.5281/zenodo.2599952. doi:10.5281/zenodo.2599952.

[5] B. R. Chakravarthi, M. Arcan, J. P. McCrae, Comparison of Different Orthographies for Machine Translation of Under-Resourced Dravidian Languages, in: M. Eskevich, G. de Melo, C. Fäth, J. P. McCrae, P. Buitelaar, C. Chiarcos, B. Klimek, M. Dojchinovski (Eds.), 2nd Conference on Language, Data and Knowledge (LDK 2019), volume 70 of *OpenAccess Series in Informatics (OASIcs)*, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2019, pp. 6:1–6:14. URL: http://drops.dagstuhl.de/opus/volltexte/2019/10370. doi:10.4230/OASIcs.LDK.2019.6.

[6] 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: Proceedings of the 12th Forum for Information Retrieval Evaluation, FIRE '20, 2020.

[7] 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.

[8] 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.

[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: Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2020). CEUR Workshop Proceedings. In: CEUR-WS. org, Hyderabad, India, 2020.

[10] 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.

[11] B. Patra, D. Das, A. Das, R. Prasath, Shared Task on Sentiment Analysis in Indian Languages (SAIL) Tweets - An Overview, 2015. doi:10.1007/978-3-319-26832-3_61.

[12] 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.

[13] 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.

[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] 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.

[17] 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.

[18] 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.