IndicBERT based approach for Sentiment Analysis on Code-Mixed Tamil Tweets

R.Ramesh Kannan, Ratnavel Rajalakshmi and Lokesh Kumar

School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, TamilNadu, India

Abstract

Nowadays, Social media networks have made a huge impact in the lifestyle. Many people prefer to express their opinions on various topics in the social media platforms such as Facebook, Twitter etc. Even though, English is predominantly used by most of the people across the world to express their views, the technological advancements have paved a way for people to use their native language also to post their opinions. As many of the social media users are bilingual in nature, the trend of using a combination of English and native language has become a common scenario. Sentiment Analysis, the task of identifying the correct opinion from these Code-Mixed social media posts, is a challenging one, as the existing architectures and algorithms are designed to handle uni-lingual posts. The diversity and the rich linguistic nature of Indian languages demand highly sophisticated systems to address the above issues. In this work, we have conducted an experimental study to handle the challenges in Code-Mixed Tamil tweets and proposed a transformer based Indic-BERT approach. From the experimental results, we have shown that, an F_1 score of 61.73% can be achieved, which is a significant improvement over the other traditional methods. This work has been submitted to the shared task on [1] Dravidian-CodeMix-FIRE 2021.

Keywords

Code-Mixed, Sentiment Analysis, Dravidian Language, Tanglish, Tamil,

1. Introduction

Sentiment analysis is the process of analyzing emotions or opinions of a given topic. It uses Natural Language Processing(NLP), text analysis and statistics to monitor the people opinions/reviews. In recent years, it is been an active area of research in both academia and industry. The best sentiment analysis system reveals how people are saying, what people are trying to mean on reviews/opinions. There is an increasing demand for sentiment analysis on social media texts which are largely Code-Mixed for Dravidian languages. Code-Mixing is a common marvel in a multilingual community and the texts are written in non-native scripts[2, 3]. Monolingual systems fails due to the complexity of the Code-Switching at different levels. Dravidian Code-Mixed shared task[3] contains data for sentiment analysis on Code-Mixed text in Dravidian language Tanglish (Tamil English).

Tamil, Telugu, Kannada and Malayalam are four of the 22 official languages of India and very few of the Dravidian Languages of India spoken in India. In particular, Kannada, Malayalam and

FIRE'21: Forum for Information Retrieval Evaluation, December 13-17, 2021,India
kannanrameshr@gmail.com (R.Ramesh Kannan); rajalakshmi.r@vit.ac.in (R. Rajalakshmi)
0000-0002-6220-1217 (R.Ramesh Kannan); 0000-0002-6570-483X (R. Rajalakshmi)

Telugu are the Dravidian Languages spoken by people from Karnataka, Kerala and Andrapradesh. Tamil is a Dravidian Languages which is spoken by people from India, Srilanka, and Tamil diaspora around the world. Tamil is the official language in Singapore and Srilanka. These languages are used by people for Various purpose like administration, education and business media. However, people often their native language with Roman script for typing because it is easy for the user to type the contents. Hence, the majority of the under-resourced languages in social media are Code-Mixed in nature.

Regional languages are used to share people opinions on the social media. Many of the resources are developed/generated in Arabic [4],English and other regional languages. In the technological world, people have ease to access internet and share Code-Mixed texts on the internet platform. Texts needs to be understood at linguistic level and the lack of Code-Mixed data to train the model is the challenging part during analysis. Monolingual trained system might not be suitable for Code-Mixed data, since the linguistic structure is different for Code-Mixed data.

Shared task [3] released in Tanglish (Tamil+English) language with social media comments for sentiment analysis on Code-Mixed data. Our proposed system reveals, how sentiment is expressed in Code-Mixed scenarios on social media by applying transformer based approach Indic-BERT[5] and we have obtained an $F1_1$ score of 61.73%. This paper is organized as follows: Section 2 shows about the Related works that are carried out on the same domain, Section 3 discusses about the proposed methodology for the shared task, Section 4 deals with the results obtained using the proposed methodology. Section 5 focuses on conclusion part of the study.

2. Related Works

Sentiment polarity analysis on online medium like YouTube comments is an important problem in analyzing people opinion on public, product, sports or on movies etc. Analyzing the polarity on the online medium contents is a challenging task. Various authors [6, 7] carried out their research on under resources languages. Sentiment analysis on online media contents[8],[9] and social media contents [10] had been studied by various authors. Sentiment analysis on movie reviews were studied by [8, 9]. Online movie review is done by [9]combining Convolution Neural Network and Bidirectional Long Short Memory to identify the opinions on movie contents as Hybrid approach. [8] shows the work implementation of feature weighting method on online movie reviews. New Relevance Factor (NRF) weighting method [11] for text classification using Naive Bayes classifier. [12] proposed universal dictionary method for text classification on Uniform Resource Locator(URL)using Linear SVM. Text classification on legal documents[13], context aware solution based on Cosine similarity approach and Term Frequency - Inverse Document Frequency(TF-IDF) to obtain the similarity between the documents. Attention mechanism is proposed with Recurrent Convolutional Neural Network(RCNN) [14] for effective learning of text features on uniform resource locator. Deep learning architecture of Convolutional Neural Network (CNN) is combined with Bidirectional Gated Recurrent Unit (BGRU) [15] to extract the features for web page classifications. Sentiment movie reviews is analysed with Long Short-term Memory (LSTM) with word embedding to extract the polarity of the reviews with self attention based approach [16]. Sentiment analysis on Tweet contents

were analysed [10] by applying Maximum Entropy supervised approach and obtained 74% cross validation accuracy score. A detailed survey on sentiment analysis was presented in work by [17].

The task on sentiment polarity identification on Code-Mixed data is challenging and recent days works are reported on the Code-Mixed data sets. The authors in [18] proposed an ensemble based machine learning approach on Code-Mixed data set. The authors proposed n-gram features with machine learning to perform classification on Hindi- English and Bengali-English data set and obtained a F1 score of 58% and 69% respectively. Ensemble classifier approach proposed using CHI square feature selection approach[19] on Code-Mixed Hindi-German language using Random Forest Classifier. Rajalakshmi. et al, [20] proposed BERT based approach on Code-Mixed data set for offensive language identification by capturing linguistic features. The authors obtained a validation F1 Score of 65% and testing F1 Score of 64%. Hate Speech analysis on Code-Mixed Marathi, Hindi data were analysed using Ensembled approach [21] Extreme Gradient Boosting Code-Mixed Hindi, English were analysed for Hate Offensive detection using Indic-BERT [22] with Majority voting approach for HASOC2021. To process multi-lingual queries Code-Mixing and Code-Borrowing were studied in recent days [23, 24, 25]. Relevance metric[26] based approach is proposed for borrowing likeliness of Hindi-English tweets for ranking. [5] proposed a new multilingual ALBERT model based approach for some of the Indian languages. Indic-BERT can be applied to various downstream tasks in Natural Language Processing. In this study, we have applied Indic-BERT on Dravidian Code-Mixed data set for sentiment polarity identifications.

3. Data Set Description

Dravidian Code-Mixed data set is a collection of YouTube video comments, which contains code mixed sentences and the types of Code-Mixed sentences are Inter-Sentential switch, Intra-Sentential switch and Tag switch[27] . Almost all the comments were written in Tamil grammar with English lexicon or English grammar with Tamil lexicon in native script and Roman scripts. Few of the comments were in Tamil script with English expressions. Data set contains ID,text and Label for each of the comments. Id contains unique number to identify particular row, text contains YouTube comments and label shows the category of the text, which contains five categories like Positive, Negative, not-Tamil, unknown state and mixed feelings.

Example from Data set:

Original Text : Yarayellam FDFS paga ippove ready agitinga

Meaning : Who are all now ready for FDFS(FIrst Day First Show)- Positive category

Original Text : Ennada viswasam mersal sarkar madhri time la likes and views create pannalayae - Negative Category

Meaning : Why likes and views are not created for the films like viswasam, mersal,sarkar. - Negative Category

The objective of the task is to identify sentiment polarity of the Code-Mixed data set of comments or posts in Tamil+English collected from social media that contains any of the following 5 category labels viz., Positive(Po), Negative(Ne), Mixed_feelings(Mf), not-Tamil(Nt), unknown_state(Us). The data distribution is tabulated in Table 1. 56% of the comments are

Table 1
Data set Distribution(percentage in category)

Category	Training	Validation
Positive(Po)	20070 (56%)	2257 (57%)
Negative(Ne)	4271 (12%)	480 (12%)
Not Tamil(Nt)	1667 (5%)	176 (5%)
Mixed Feelings(Mf)	4020 (11%)	438 (11%)
Unknown State(Us)	5628 (16%)	611 (16%)
Total	35656	3962

positive and other remaining 44% of the comments are in other four categories. The percentage of category values are as follows: Ne with 12%, Nt with 5%, Mf with 11% and Us with 16%. As part of the sentiment analysis task, the training and validation set were released with 35656 and 3962 labelled social media comments. Both the training and validation set follows the same distribution.

4. Proposed Methodology

The Code-Mixed comments contains Tamil, English and other language phrases and words in the context. Instead of converting the text into any of the common language, a Multilingual pretrained model[5] Indic-BERT is used, that has been pretrained on 12 indian languages. Indic-BERT pretrained model is based on ALBERT(A Lite BERT for Self-Supervised Learning of Language Representations) model, which is a recent derivative of BERT(Bidirectional Encoder Representations from Transformers), which is pretrained on 12 indian languages like Assamese, Bengali, English, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil, Telugu. The proposed BERT has less parameters than other public models like mBERT and XLM-R while it manages to give state of the art performance on several tasks.

Since it is a pretrained multilingual transformer model, the data needs to converted into corresponding embeddings for classifications. As a pre processing step, Autotokenizer tokenizes all the sentences into tokens. In tokens, Class[CLS] token is added at the beginning of the sentence and seperation[SEP] token is added at the end of the each sentences. Padding [Pad] token is padded with all the sentences till the maximum length of the sentence. Assign unique id to each token for further processing. Attention mask is also generated for each input sentences and it tells which tokens should be attempted and which should not be attempted by the model during training. This will be useful when input is fed into transformer based Indic-BERT model.

To determine the sentiments expressed in the Code-Mixed YouTube comments/posts, Indic-BERT model is proposed with the fine tuned parameters [5]. Indic-BERT is a multilingual representation model that extracts the context from different language input representations in both the directions. To capture the semantic and linguistic features of a multilingual sentence, Indic-BERT is applied. YouTube comments/Posts may have more than one sentences. Indic-BERT has the ability to consider these multilingual inputs sentences into a single sequence for input representations. Indic-BERT embeddings combines the token embedding,segment embedding

Table 2
Performance of the proposed approach on Training set

Epoch	Training Acc	Val Acc	Training Loss	Val Loss
1	0.6725	0.6012	1.1724	1.0428
2	0.6421	0.5959	1.0027	1.0234
3	0.6896	0.6327	0.8877	1.0013
4	0.6893	0.6297	0.7425	0.9687
5	0.6991	0.6329	0.6939	0.9681

Table 3
Comparison of results

Team	Precision	Recall	F1
AIML	59.6	60.3	59.9
SSN_NLP_MLRG	59.7	61.3	60.3
Ryzer	59.7	61.4	60.4
Proposed	61.27	64.54	**61.73**

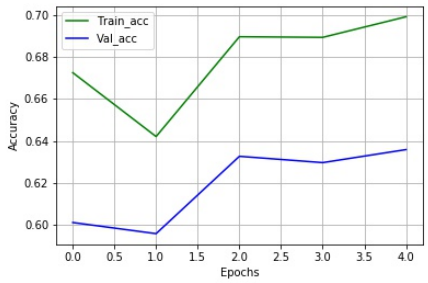

Figure 1: Training and Validation Accuracy

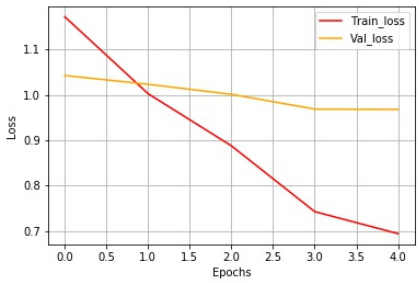

Figure 2: Training and Validation Loss

and positional embeddings. Pretrained model can be fine-tuned to suit the downstream tasks by adding classification layer at the bottom of the model. Indic-BERT can be used for this task of how sentiment is expressed in Code-Mixed scenarios on social media.

5. Results and Discussion

To study the performance of the sentiment polarity of the system, we have conducted experiment based on Indic-BERT approach on Code-Mixed data set. The experiment was conducted on workstation with Intel Xeon Quad Core Processor, 32 GB RAM, NVIDIA Quadro P4000 GPU 8GB. To capture the sentiment polarity on the Code-Mixed data set, we have tried transformer based approach of Indic-BERT. To attain better performance of the BERT model, we have fine-tuned the parameters and obtained learning rate=3e-5,batch size=64, epochs=5. Figure 1, shows the accuracy graph on training data and validation data. For the 5th epoch accuracy score reached a

maximum level. Figure 2, plotted with loss values on training data and validation data. Obtained a training accuracy of 69.91% and loss of 0.6939. For the validation set, obtained a loss of 0.9681 and accuracy of 63.29%. Here the classifier is able to classify all the categories, even the data set is not balanced set. Even the data set contains very less number of Not Tamil categories are classified correctly. From Table 3, Our proposed model out performs on Weighted average F1 score of 61.73%. The model is able to classify all the categories irrespective of the specific category.

6. Conclusion

There is an increase in social media contents in recent days. The goal of the Dravidian-CodeMix-FIRE 2021 is to identify the subjective opinions or emotional responses of the social media comments. In this work, we have presented the challenges involved in extracting the key terms to identify the opinion from the Code-Mixed tweets. A detailed experimental study has been performed using different architectures and we found that, the sentiment in the social media contents are better captured using the Indic-BERT language model. We have obtained a weighted F_1 score of 61.73% with the proposed model. We observed that, the data set is skewed and the lack of enough samples for every category has impacted the performance of the classifier. In our future work, we planned to address the class imbalance problem for Code-Mixed sentiment analysis.

Acknowledgments

The authors would like to thank the management of Vellore Institute of Technology, Chennai for providing the support to carry out this work. We would like to thank the Department of Science and Engineering Research Board (SERB),Government of India for their financial grant (Award Number: ECR/2016/00484) for this research work.

References

[1] B. R. Chakravarthi, R. Priyadharshini, S. Thavareesan, D. Chinnappa, T. Durairaj, 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 2021, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.

[2] B. R. Chakravarthi, P. K. Kumaresan, R. Sakuntharaj, A. K. Madasamy, S. Thavareesan, P. B, S. Chinnaudayar Navaneethakrishnan, J. P. McCrae, T. Mandl, Overview of the HASOC-Dravidian CodeMix Shared Task on Offensive Language Detection in Tamil and Malayalam, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.

[3] 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.

[4] H. Mubarak, K. Darwish, W. Magdy, Abusive language detection on Arabic social media, in: Proceedings of the First Workshop on Abusive Language Online, Association for Computational Linguistics, Vancouver, BC, Canada, 2017, pp. 52–56. URL: https://aclanthology.org/W17-3008.

[5] D. Kakwani, A. Kunchukuttan, S. Golla, G. N.C., A. Bhattacharyya, M. M. Khapra, P. Kumar, IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 4948–4961. URL: https://aclanthology.org/2020.findings-emnlp.445.

[6] S. Thavareesan, S. Mahesan, Sentiment lexicon expansion using word2vec and fasttext for sentiment prediction in tamil texts, 2020. doi:10.1109/MERCon50084.2020.9185369.

[7] S. Thavareesan, S. Mahesan, Sentiment analysis in tamil texts: A study on machine learning techniques and feature representation, 2019 14th Conference on Industrial and Information Systems (ICIIS) (2019) 320–325.

[8] S. Sivakumar, R. Rajalakshmi, Comparative evaluation of various feature weighting methods on movie reviews, in: H. S. Behera, J. Nayak, B. Naik, A. Abraham (Eds.), Computational Intelligence in Data Mining, Springer Singapore, Singapore, 2019, pp. 721–730.

[9] S. Soubraylu, R. Rajalakshmi, Hybrid convolutional bidirectional recurrent neural network based sentiment analysis on movie reviews, Computational Intelligence 37 (2021) 735–757. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/coin.12400. doi:https://doi.org/10.1111/coin.12400. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/coin.12400.

[10] A. Samuels, J. Mcgonical, Sentiment analysis on social media content, CoRR abs/2007.02144 (2020). URL: https://arxiv.org/abs/2007.02144. arXiv:2007.02144.

[11] R. R., Supervised term weighting methods for url classification, Journal of Computer Science 10 (2014). doi:10.3844/jcssp.2014.1969.1976.

[12] R. R., C. Aravindan, An effective and discriminative feature learning for url based web page classification, in: 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2018, pp. 1374–1379. doi:10.1109/SMC.2018.00240.

[13] R. R. Kannan, R. Rajalakshmi, Dlrg@aila 2019: Context - aware legal assistance system, in: P. Mehta, P. Rosso, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019, volume 2517 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019, pp. 58–63. URL: http://ceur-ws.org/Vol-2517/T1-10.pdf.

[14] R. R., H. Tiwari, J. Patel, R. R., K. Ramamurthy, Bidirectional GRU-Based Attention Model for Kid-Specific URL Classification, 2020, pp. 78–90. doi:10.4018/978-1-7998-1192-3.ch005.

[15] R. Rajalakshmi, H. Tiwari, J. Patel, A. Kumar, R. Karthik., Design of kids-specific url classifier using recurrent convolutional neural network, Procedia Computer Science 167 (2020) 2124–2131. URL: https://www.sciencedirect.com/science/article/pii/S1877050920307262. doi:https://doi.org/10.1016/j.procs.2020.03.260, international Conference on Computational Intelligence and Data Science.

[16] S. Soubraylu, R. Rajalakshmi, Analysis of sentiment on movie reviews using word embed-

ding self-attentive lstm, International Journal of Ambient Computing and Intelligence 12 (2021) 33–52. doi:10.4018/IJACI.2021040103.

[17] V. Ganganwar, R. Rajalakshmi, Implicit aspect extraction for sentiment analysis: A survey of recent approaches, Procedia Computer Science 165 (2019) 485–491.

[18] P. Mishra, P. Danda, P. Dhakras, Code-mixed sentiment analysis using machine learning and neural network approaches, CoRR abs/1808.03299 (2018). URL: http://arxiv.org/abs/1808.03299. arXiv:1808.03299.

[19] R. Rajalakshmi, B. Y. Reddy, Dlrg@hasoc 2019: An enhanced ensemble classifier for hate and offensive content identification, in: P. Mehta, P. Rosso, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019, volume 2517 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019, pp. 370–379. URL: http://ceur-ws.org/Vol-2517/T3-26.pdf.

[20] R. Rajalakshmi, Y. Reddy, L. Kumar, DLRG@DravidianLangTech-EACL2021: Transformer based approachfor offensive language identification on code-mixed Tamil, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Kyiv, 2021, pp. 357–362. URL: https://aclanthology.org/2021.dravidianlangtech-1.53.

[21] R. Rajalakshmi, S. Srivarshan, M. L. P. R. Faerie, M. Faerie, K. E, S. Prithvi, K. M. Anand, Conversational hate-offensive detection in code-mixed hindi-english tweets, Association for Computing Machinery, 2021.

[22] R. Rajalakshmi, L. P. Reddy, M. Faerie, S. Srivarshan, K. M. Anand, Hate speech and offensive content identification in hindi and marathi languages using ensemble techniques, Association for Computing Machinery, 2021.

[23] B. R. Chakravarthi, M. Arcan, J. P. McCrae, Improving wordnets for under-resourced languages using machine translation, in: Proceedings of the 9th Global Wordnet Conference, Global Wordnet Association, Nanyang Technological University (NTU), Singapore, 2018, pp. 77–86. URL: https://aclanthology.org/2018.gwc-1.10.

[24] WordNet Gloss Translation for Under-resourced Languages using Multilingual Neural Machine Translation, Zenodo, 2019. URL: https://doi.org/10.18653/v1/w19-7101.

[25] 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.

[26] R. Rajalakshmi, R. Agrawal, Borrowing likeliness ranking based on relevance factor, in: Proceedings of the Fourth ACM IKDD Conferences on Data Sciences, CODS '17, Association for Computing Machinery, New York, NY, USA, 2017. URL: https://doi.org/10.1145/3041823.3067694. doi:10.1145/3041823.3067694.

[27] 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.