# JUNLP@DravidianLangTech-EACL2021: Offensive Language Identification in Dravidian Langauges

**Avishek Garain**[1]
Jadavpur University,
India

**Atanu Mandal**[2]
Jadavpur University,
India

**Sudip Kumar Naskar**[3]
Jadavpur University,
India

{[1]avishekgarain, [2]atanumandal0491}@gmail.com,
[3]sudipkumar.naskar@jadavpuruniversity.in

## Abstract

Offensive language identification has been an active area of research in natural language processing. With the emergence of multiple social media platforms offensive language identification has emerged as a need of the hour. Traditional offensive language identification models fail to deliver acceptable results as social media contents are largely in multilingual and are code-mixed in nature. This paper tries to resolve this problem by using IndicBERT and BERT architectures, to facilitate identification of offensive languages for Kannada-English, Malayalam-English, and Tamil-English code-mixed language pairs extracted from social media. The presented approach when evaluated on the test corpus provided precision, recall, and F1 score for language pair Kannada-English as 0.62, 0.71, and 0.66, respectively, for language pair Malayalam-English as 0.77, 0.43, and 0.53, respectively, and for Tamil-English as 0.71, 0.74, and 0.72, respectively.

## 1 Introduction

Social media platforms like question answering platforms, collaborative projects, social networks, news platforms provide discussion area for users, where content moderators are engaged to keep respectful conversations (Thavareesan and Mahesan, 2019, 2020a,b). Moderators assure that the platform's discussion rules are adhered to, including the prohibition of offensive languages. Moderators implement these rules by partly or entirely removing user comments.

Typically, platform rules are available to the user in the form of guidelines. However, all user doesn't follow the rules while commenting on a post. An increase of end-users in social platforms leads to the increasing number of comments that make the moderator restless to identify offensive and non-offensive. To intercept the circumstances recent trends of identification of offensive language have become a scientific asset (Chakravarthi et al., 2020c; Chakravarthi, 2020).

In the context of Natural Language Processing (NLP), offensive language identification is a classification task that is aimed to identify and minimize offensive contents in social media (Mandl et al., 2020). There have been advancements in this domain of research both in industrial and academia with increasing access to larger and richer social media data over the years. India is a linguistically diverse country with 22 official languages with common languages used being English and Hindi. There has been a mixing of cultures and languages over the years thus leading to an increasing demand for offensive language identification on social media texts which are largely code-mixed.

Code-mixing refers to a prevalent phenomenon which exists in a multilingual community and the code-mixed texts are sometimes written in non-native scripts. Any System which is trained on monolingual datasets miserably fails when exposed to code-mixed data due to the complexity of code-switching at different linguistic levels in the text (Jose et al., 2020; Priyadharshini et al., 2020).

Many researcher has proposed many different approaches to achieve the state-of-the-art results in code-mixing scenarios during the recent years. Author (Mathur et al., 2018) has solved the problem of classification of the tweets in Hindi-English Offensive Tweet (HEOT) dataset using transfer learning in which author has employed Convolutional Neural Networks as pre-trained on tweets in English followed by retraining on Hinglish tweets. HEOT dataset consists of Hindi-English code switched language into three classes nonoffensive, abusive and hate-speech. The author (Bohra et al., 2018) proposes a supervised classification system that detects hate speech in the text using various character level, word level, and lexicon based features

*Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 319–322
April 20, 2021 ©2021 Association for Computational Linguistics

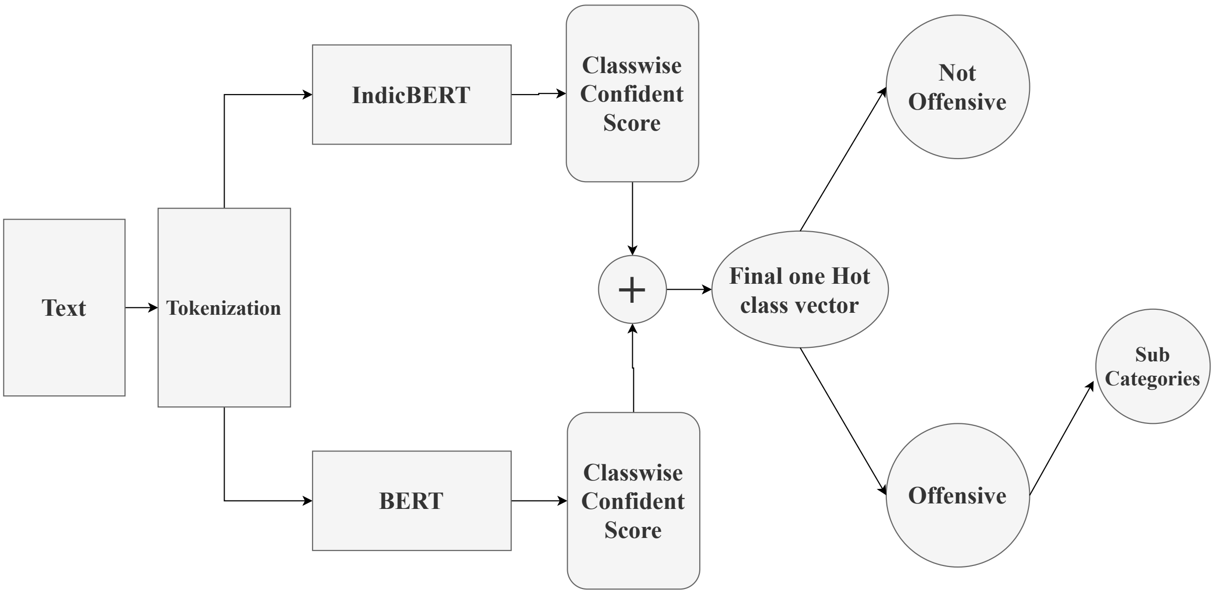
Text
Tokenization
IndicBERT
BERT
Classwise Confident Score
Classwise Confident Score
+
Final one Hot class vector
Not Offensive
Offensive
Sub Categories

Figure 1: Framework for Offensive Language Identification

whereas (Santosh and Aravind, 2019) dealt the task of identification of hate speech and offensive language from code-mixed social media text using two architecture namely sub-word level LSTM model and Hierarchical LSTM model with attention based on phonemic sub-words. Another real-world issue dataset have been released by the authors (Hande et al., 2020) who have introduced Kannada CodeMixed Dataset (KanCMD). The dataset is a multi-task learning dataset for sentiment analysis and offensive language identification. The KanCMD dataset highlights two real-world issues from the social media text. Firstly, it contains actual comments in code mixed text posted by users on YouTube social media, rather than in monolingual text from the textbook. Secondly, it has been annotated for two tasks, namely sentiment analysis and offensive language detection for under-resourced Kannada language. Hence, KanCMD was meant to stimulate research in under-resourced Kannada language on real-world code-mixed social media text and multi-task learning. Our proposed system is yet to explore the dataset.

This paper aims to solve this research problem by using generalized Deep learning architectures named IndicBERT (Kakwani et al., 2020) and BERT (Devlin et al., 2019). The goal of this task is to identify offensive language content of the code-mixed dataset of comments or posts in Dravidian Languages collected from social media. The code-mixed language for the task was Tamil-English (Chakravarthi et al., 2020b), Malayalam-English (Chakravarthi et al., 2020a), and Kannada-English (Hande et al., 2020) provided by DravidianLangTech-2021 (Chakravarthi et al., 2021). The working system is available in GitHub[1].

The rest of the paper has been organized as follows. Section 2 describes the data that was used to build the proposed Offensive Language Identification system. Section 3 describes the proposed model used to build the system and will be followed by the evaluation of the model in section 4.

## 2 Data

The shared task organized by DravidianLangTech-2021 provided the gold standard corpus for offensive language identification of code-mixed text for three different sets Tamil-English, Malayalam-English, and Kannada-English in Dravidian languages.

The dataset was collected from social media. The average sentence length of comment or post of the corpora is 1. Each comment or post is annotated at the comment or post level. Depicting the real-life scenarios the corpora was provided with class imbalance problem. The corpora has 6 output labels: Not offensive, Offensive untargeted, Offensive targeted individual, Offensive targeted group, Offensive targeted other, or Not in indented language.

[1] https://github.com/garain/EACL21-system

## 3 Framework

Every sentence is a sequence of words. For a code-mixed scenarios, these words from different languages might have no proper boundary to separate them. Our goal was to first classifying the texts to the primary category which are either Offensive or not offensive. Thereafter, if the texts were identified to be offensive, then further classification was done for sub categories which are targeted insult individual, targeted insult group or targeted insult other. We considered the whole task as a multi-label classification problem. The overall methodology is shown in Figure 1.

The training corpus was first tokenized and the required tags were inserted. Then the tokenized data was fed to the IndicBERT and BERT models. We make use of pre-trained vectors for initializing the mentioned models and then fine-tuned using the training corpus. The IndicBERT model supports various Indic languages. Therefore for each language pair a seperate model had to be trained with corresponding language settings. The confidence scores for each class obtained from the IndicBERT model and the regular BERT model are then concatenated to give the final scores for each of the classes. The score array is converted to a 1-D array in one-hot vector format containing 1's depicting not offensive category and 0's depicting offensive category, which was further classified for rest of the sub categories.

The output vector consisted of the following labels in order:

- Not - offensive
- Offensive - Untargeted
- Offensive - Targeted Insult Individual
- Offensive - Targeted Insult Group
- Offensive - Targeted Insult Other
- Not in indented language

If the output vector had '1' for the first class, then rest of the classes were converted to '0' else the maximum among the classes leaving the first class was converted to '1'. Finally the corresponding label is given as output.

## 4 Evaluation

Offensive language identification was evaluated using Sklearn Classification Report (Pedregosa et al., 2011). Performance was measured in terms of Precision, Recall and F1-Score across all the classes. The result of the evaluation are shown in Table 1.

| Language Pair | Precision | Recall | F1 score |
|---|---|---|---|
| Kannada English | 0.62 | 0.71 | 0.66 |
| Malayalam English | 0.77 | 0.43 | 0.54 |
| Tamil English | 0.71 | 0.74 | 0.72 |

Table 1: Results for all the Language pairs by Team JUNLP

## 5 Conclusion

In the current work, we attempted to solve the problem of Offensive Language Identification, while participating in the DravidianLangTech-2021 shared task. Our system was based on ensemble of IndicBERT model and generic BERT model using multi-label classification approach. Our system when evaluated by the organizers earned F1 score of 0.66, 0.54, and 0.72 for Kannada-English, Malayalam-English, and Tamil-English Language Pair, respectively. As future work, we would like to increase this data, and use other state-of-the-art Neural Network architecture to improve the performance. Also we would like to evaluate with the dataset released by author (Hande et al., 2020).

## References

Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of hindi-english code-mixed social media text for hate speech detection. In *Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media*, pages 36–41.

Bharathi Raja Chakravarthi. 2020. HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53, Barcelona, Spain (Online). Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John Philip McCrae. 2020a. 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)*, pages 177–184, Marseille, France. European Language Resources association.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020b. 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.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Anand Kumar M, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan V, Elizabeth Sherly, and John Philip McCrae. 2021. Findings of the shared task on Offensive Language Identification in Tamil, Malayalam, and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Shardul Suryawanshi, Navya Jose, Elizabeth Sherly, and John P. McCrae. 2020c. Overview of the Track on Sentiment Analysis for Dravidian Languages in Code-Mixed Text. In *Forum for Information Retrieval Evaluation*, FIRE 2020, page 21–24, New York, NY, USA. Association for Computing Machinery.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Adeep Hande, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2020. 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*, pages 54–63, Barcelona, Spain (Online). Association for Computational Linguistics.

Navya Jose, Bharathi Raja Chakravarthi, Shardul Suryawanshi, Elizabeth Sherly, and John P. McCrae. 2020. A Survey of Current Datasets for Code-Switching Research. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 136–141.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2020. Overview of the HASOC Track at FIRE 2020: Hate Speech and Offensive Language Identification in Tamil, Malayalam, Hindi, English and German. In *Forum for Information Retrieval Evaluation*, FIRE 2020, page 29–32, New York, NY, USA. Association for Computing Machinery.

Puneet Mathur, Rajiv Shah, Ramit Sawhney, and Debanjan Mahata. 2018. Detecting offensive tweets in hindi-english code-switched language. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 18–26.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Mani Vegupatti, and John P. McCrae. 2020. Named Entity Recognition for Code-Mixed Indian Corpus using Meta Embedding. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 68–72.

TYSS Santosh and KVS Aravind. 2019. Hate speech detection in hindi-english code-mixed social media text. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, pages 310–313.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2019. Sentiment Analysis in Tamil Texts: A Study on Machine Learning Techniques and Feature Representation. In *2019 14th Conference on Industrial and Information Systems (ICIIS)*, pages 320–325.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020a. Sentiment Lexicon Expansion using Word2vec and fastText for Sentiment Prediction in Tamil texts. In *2020 Moratuwa Engineering Research Conference (MERCon)*, pages 272–276.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020b. Word embedding-based Part of Speech tagging in Tamil texts. In *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*, pages 478–482.