Pretrained Transformers for Offensive Language Identification in Tanglish

Sean Benhur[1], Kanchana Sivanraju[1]

[1]PSG College of Arts and Science, Civil Aerodrome Post, Coimbatore, India

Abstract

This paper describes the system submitted to Dravidian-Codemix-HASOC2021: Hate Speech and Offensive Language Identification in Dravidian Languages (Tamil-English and Malayalam-English). This task aims to identify offensive content in code-mixed comments/posts in Dravidian Languages collected from social media. Our approach utilizes pooling the last layers of pretrained transformer multilingual BERT for this task which helped us achieve rank nine on the leaderboard with a weighted average score of 0.61 for the Tamil-English dataset in subtask B. After the task deadline, we sampled the dataset uniformly and used the MuRIL pretrained model, which helped us achieve a weighted average score of 0.67, the top score in the leaderboard. Furthermore, our approach to utilizing the pretrained models helps reuse our models for the same task with a different dataset. Our code and models are available in GitHub [1]

Keywords

Hate Speech, Offensive Content, BERT, Transformer

1. Introduction

In the era of the internet, people from various age groups engage in social media, it has become a one-stop shop for all activities from learning to entertainment, but it is also filled with offensive and disturbing content, which is potentially harmful to everyone [1]. To prevent this, an automated system of identifying and flagging offensive content should be developed. Though there is a substantial amount of work done on major languages like English to identify offensive content [2], it is a challenging task to identify and flag offensive content in low resource languages, since many users tend to write their language in English script, which is called code-switching or code-mixing [3, 4, 5]. Developing NLP systems on code-mixed text is challenging since the number of datasets is scarce[6, 7, 8, 9] and there are no clear patterns on these texts. The spelling and context might vary depending upon the user.

Dravidian languages are under-resourced in natural language processing [10]. Dravidian name was derived from Tamil, Dravidian means Tamil [11], Dravidian languages are Tamili languages [12]. Tamil is a language spoken by Tamils in Tamil Nadu, India, Sri Lanka, and the Tamil diaspora worldwide, with official recognition in India, Sri Lanka, and Singapore [13, 14, 15]. Current Tamil script was developed from the Tamili script, the Vatteluttu alphabet, and the Chola-Pallava script. There are 12 vowels, 18 consonants, and 1 ytam in this word

[1]https://github.com/seanbenhur/tanglish-offensive-language-identification

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

seanbenhur@gmail.com (S. Benhur); kanchana@psgcas.ac.in (K. Sivanraju)

https://seanbenhur.github.io/ (S. Benhur)

(voiceless velar fricative) [16, 17, 18, 19]. The Tamil script is also used to write minority languages including Saurashtra, Badaga, Irula, and Paniya. Tamil Eluttu "means" sound, letter, phoneme" in Tolkappiyam (about 5,320 BCE), and thus includes the sounds of the Tamil language, as well as how they are created (phonology) [20, 21, 22, 23]. All the Tamili (Dravidian) languages evolved from Tamil language [24, 25].

HASOC2021: Hate Speech and Offensive Content Identification is a competition that helps increase research in offensive language identification in code mixed languages such as Tamil-English and Malayalam-English [6]. The dataset consists of comments/posts that were collected from Youtube and social media. Each comment/post is annotated with an offensive language label at the comment/post level. This dataset also has class imbalance problems depicting real-world scenarios.

In this paper, we present our system developed for HASOC 2021; the rest of the paper is organized as follows. Section 2 discusses the research work on offensive language identification and natural language processing in under-resourced languages. Following this, in section 3, we present the methodology of our work, from preprocessing the dataset, our model architecture, and training procedures. In section 4, we discuss our results. Finally, in section 6, we conclude with a summary and future work.

2. Related Work

Offensive Language identification has been widely made across many people in multiple languages. Shared tasks like HASOC-19 [26] dealt with hate speech and offensive language identification in Indo-European languages. HASOC-Dravidian-CodeMix - FIRE 2020 [27][28] is the first shared task for identifying offensive content in Tamili languages. Previous work on Tamili languages on hope speech [29, 30], troll meme detection [31], multimodal sentiment analysis [9] have paved the way to research in Tamili languages.

Researchers have used a wide variety of techniques for the identification of offensive language. There have been previous work [32] in using classical machine learning models with efficient feature generation. Other researchers in [33] [34] have used an ULMFit model [35] and pretrained XLM-Roberta model with translated and transliterated texts for this task.

3. Methodology

This section briefly describes our methodology for this task, including data preparation, model architecture, and training strategies. For this HASOC 2021 competition, we only use the datasets that were provided for the HASOC task. Table 1 shows the statistics of the train and dev distribution.

3.1. Dataset

The dataset given for subtask, Offensive Language Identification in Tamil-English, consists of Youtube comments, present in code-mixed data containing text written in both native and roman scripts in English.

Table 1
Distribution of Tamil English Dataset

Distribution	Data
Train	4937
Test	1000

For training our model, we concatenate both training and dev sets; we remove the URLs, English stopwords, @username mentions, NAN values, emojis, and also punctuations; this preprocessing method is applied to all the train, dev, and test sets. After the task deadline, we sample the dataset uniformly to handle the class imbalance problem in this dataset, which helps us improve our score. Table 1 shows the statistics of the given dataset after preprocessing.

3.2. Model Architecture

We use pretrained transformer models with custom pooled output for this task of identifying offensive content. We have used mBERT and MuRIL pretrained models from huggingface checkpoints. In this section, we describe our pooling operations on the pretrained models and the pretrained models.

Attention Pooler: In this method, the attention operation described in the below equation is applied to the last hidden state of the pretrained transformer; empirically, this helps the model learn the contribution of individual tokens. Finally, the returned pooled output from the transformer is further passed to a linear layer for the prediction of the label.

$$o = W_h^T softmax(q h_{CLS}^T) h_{CLS} \tag{1}$$

where W_h^T and q are learnable weights.

$$y = softmax(W_o^T + bo) \tag{2}$$

Mean Pooler: In this method, the average of the last layer of the pretrained transformer is taken. This acts like a pooling layer in a convolutional neural net. An alternative to this method is to use max pooling, but max-pooling selects only the words with essential features rather than utilizing all the words. Since our dataset is code-mixed and the spelling of the tokens are not precise, we choose to go with mean pooling approach.

mBERT Multilingual models of BERT [36]. This model was pre-trained using the same pretraining strategy that was employed to BERT, which is Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). It was pretrained on the Wikipedia dump of top 104 languages. To account for the data imbalance due to the size of Wikipedia for a given language, exponentially smoothed weighting of data was performed during data creation and word piece vocabulary creation. This results in high resource languages being under-sampled while low resourced languages being over-sampled.

MuRIL MuRIL [37], pretrained model is trained on 16 different Indian Languages; the model was pretrained on Masked Language Modeling(MLM) and Translated Language Modelling(TLM). This model outperforms mBERT on all the tasks in XTREME [38]

3.3. Training

Though finetuning transformers gives better results and is dominant across leaderboards of various NLP competitions. Transformer models are still unstable due to catastrophic forgetting [39]. For this offensive language identification task, we carefully choose our hyperparameters for experimentation. We finetune our custom models with binary-cross-entropy loss and AdamW optimizer, which decouples the weight decay from the optimization step. Linear scheduler for learning rate scheduling with 2e-5 as an initial step is used with this training strategy. The training hyperparameters are listed in Table 2.

Table 2
Hyperparameters used across experiments

Hyperparameters	Values
Learning Rate	2e-5
Maximim Sequence Length	512
Batch Size	8
Epochs	5
Weight Decay	0.01
Dropout	0.5
AdamW epsilon	1e-06

4. Results and Discussion

In this HASOC 2021 competition, the teams were ranked by the weighted F1-score of their classification system. This section discusses our experimental results; since we have used both training and dev sets for training, the train set in the dataset distribution refers to the concatenated given train and dev sets. The W-Precision, W-Recall, and W-F1-Score refer to the Weighted precision, weighted recall, and weighted F1-Score. Table 3 shows our results obtained before the task deadline using Attention Pooler and mBERT without sampling the dataset. After the task deadline, we uniformly sample our dataset and run or experiments on MuRIL and mBERT with Attention Pooling and Mean Pooling. The results are provided in Table 4 and Table 5. The W-precision, W-Recall and W-F1 Score stands for Weighted Precision, Weighted Recall and Weighted F1-Score.

Table 3
Results before task deadline

Dataset Distribution	W-Precision	W-Recall	W-F1 Score
Train	0.90	0.91	0.92
Test	0.61	0.60	0.61

From the above results, we conclude that the pretrained model MuRIL with MeanPooler performs best than others. Also, one can infer that the difference between training and test scores shows that the model is suffering from overfitting, and also sampling the dataset uniformly is a crucial step to increasing the score.

Table 4
Results on Train dataset

Model	W-Precision	W-Recall	W-F1 Score
mBERT with AttentionPooler	0.93	0.90	0.93
mBERT with MeanPooler	0.90	0.92	0.91
MuRIL with AttentionPooler	0.88	0.88	0.88
MuRIL with MeanPooler	0.93	0.93	0.93

Table 5
Results on Test data

Model	W-Precision	W-Recall	W-F1 Score
mBERT with AttentionPooler	0.65	0.65	0.65
mBERT with MeanPooler	0.61	0.61	0.61
MuRIL with AttentionPooler	0.63	0.63	0.63
MuRIL with MeanPooler	0.67	0.67	0.67

5. Conclusion

In this paper, we have presented our solution for the Offensive Language Identification system, which uses pretrained transformers mBERT and MuRIL. As a result, we achieve Rank 9 on the leaderboard and a 0.67 f1-score after the task deadline. For future research, we will consider improving the results by using any external dataset and other pretrained models and reducing the generalization error of the model.

References

[1] S. U. Hegde, A. Hande, R. Priyadharshini, S. Thavareesan, R. Sakuntharaj, S. Thangasamy, B. Bharathi, B. R. Chakravarthi, Do Images really do the Talking? Analysing the significance of Images in Tamil Troll meme classification, arXiv preprint arXiv:2108.03886 (2021).

[2] B. R. Chakravarthi, R. Priyadharshini, R. Ponnusamy, P. K. Kumaresan, K. Sampath, D. Thenmozhi, S. Thangasamy, R. Nallathambi, J. P. McCrae, Dataset for Identification of Homophobia and Transophobia in Multilingual YouTube Comments, arXiv preprint arXiv:2109.00227 (2021).

[3] S. Suryawanshi, B. R. Chakravarthi, M. Arcan, S. Little, P. Buitelaar, TrollsWithOpinion: A Dataset for Predicting Domain-specific Opinion Manipulation in Troll Memes, arXiv preprint arXiv:2109.03571 (2021).

[4] A. Hande, K. Puranik, K. Yasaswini, R. Priyadharshini, S. Thavareesan, A. Sampath, K. Shanmugavadivel, D. Thenmozhi, B. R. Chakravarthi, Offensive Language Identification in Low-resourced Code-mixed Dravidian languages using Pseudo-labeling, arXiv preprint arXiv:2108.12177 (2021).

[5] A. Hande, S. U. Hegde, R. Priyadharshini, R. Ponnusamy, P. K. Kumaresan, S. Thavareesan,

B. R. Chakravarthi, Benchmarking multi-task learning for sentiment analysis and offensive language identification in under-resourced Dravidian languages, arXiv preprint arXiv:2108.03867 (2021).

[6] 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-DravidianCodeMix Shared Task on Offensive Language Detection in Tamil and Malayalam, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.

[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] 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, arXiv preprint arXiv:2106.09460 (2021).

[9] B. R. Chakravarthi, K. Soman, R. Ponnusamy, P. K. Kumaresan, K. P. Thamburaj, J. P. McCrae, et al., DravidianMultiModality: A Dataset for Multi-modal Sentiment Analysis in Tamil and Malayalam, arXiv preprint arXiv:2106.04853 (2021).

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

[11] D. Chinnappa, P. Dhandapani, Tamil lyrics corpus: Analysis and experiments, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Kyiv, 2021, pp. 1–9. URL: https://aclanthology.org/2021.dravidianlangtech-1.1.

[12] J. J. Andrew, JudithJeyafreedaAndrew@DravidianLangTech-EACL2021:offensive language detection for Dravidian code-mixed YouTube comments, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Kyiv, 2021, pp. 169–174. URL: https://aclanthology.org/2021.dravidianlangtech-1.22.

[13] K. Sarveswaran, G. Dias, M. Butt, ThamizhiMorph: A morphological parser for the Tamil language, Machine Translation 35 (2021) 37–70.

[14] B. Bharathi, A. S. A, SSNCSE_NLP@DravidianLangTech-EACL2021: Offensive language identification on multilingual code mixing text, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Kyiv, 2021, pp. 313–318. URL: https://aclanthology.org/2021.dravidianlangtech-1.45.

[15] B. Bharathi, A. S. A, SSNCSE_NLP@DravidianLangTech-EACL2021: Meme classification for Tamil using machine learning approach, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Kyiv, 2021, pp. 336–339. URL: https://aclanthology.org/2021.dravidianlangtech-1.49.

[16] S. Thavareesan, S. Mahesan, Word embedding-based Part of Speech tagging in Tamil texts, in: 2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS),

2020, pp. 478–482. doi:10.1109/ICIIS51140.2020.9342640.
[17] S. Thavareesan, S. Mahesan, Sentiment Lexicon Expansion using Word2vec and fastText for Sentiment Prediction in Tamil texts, in: 2020 Moratuwa Engineering Research Conference (MERCon), 2020, pp. 272–276. doi:10.1109/MERCon50084.2020.9185369.
[18] S. Thavareesan, S. Mahesan, Sentiment Analysis in Tamil Texts: A Study on Machine Learning Techniques and Feature Representation, in: 2019 14th Conference on Industrial and Information Systems (ICIIS), 2019, pp. 320–325. doi:10.1109/ICIIS47346.2019.9063341.
[19] D. Thenmozhi, C. Aravindan, Ontology-based Tamil–English cross-lingual information retrieval system, Sādhanā 43 (2018) 1–14.
[20] R. Sakuntharaj, S. Mahesan, A novel hybrid approach to detect and correct spelling in Tamil text, in: 2016 IEEE International Conference on Information and Automation for Sustainability (ICIAfS), IEEE, 2016, pp. 1–6.
[21] R. Sakuntharaj, S. Mahesan, Use of a novel hash-table for speeding-up suggestions for misspelt Tamil words, in: 2017 IEEE International Conference on Industrial and Information Systems (ICIIS), IEEE, 2017, pp. 1–5.
[22] R. Sakuntharaj, S. Mahesan, Detecting and correcting real-word errors in Tamil sentences, Ruhuna Journal of Science 9 (2018).
[23] R. Sakuntharaj, S. Mahesan, A refined pos tag sequence finder for Tamil sentences, in: 2018 IEEE International Conference on Information and Automation for Sustainability (ICIAfS), IEEE, 2018, pp. 1–6.
[24] A. R, S. C N, Building discourse parser for Thirukkural, in: Proceedings of the 16th International Conference on Natural Language Processing, NLP Association of India, International Institute of Information Technology, Hyderabad, India, 2019, pp. 18–25. URL: https://aclanthology.org/2019.icon-1.3.
[25] C. Subalalitha, Information extraction framework for Kurunthogai, Sādhanā 44 (2019) 1–6.
[26] P. Majumder, D. Patel, S. Modha, T. Mandl, Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages, 2019. doi:10.1145/3368567.3368584.
[27] 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, 2020, pp. 21–24.
[28] T. Mandl, S. Modha, A. Kumar M, B. R. Chakravarthi, 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, 2020, pp. 29–32.
[29] B. R. Chakravarthi, 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, Association for Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 41–53. URL: https://aclanthology.org/2020.peoples-1.5.
[30] B. R. Chakravarthi, V. Muralidaran, Findings of the shared task on hope speech detection for equality, diversity, and inclusion, in: Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion, Association for Computational Linguistics, Kyiv, 2021, pp. 61–72. URL: https://aclanthology.org/2021.ltedi-1.8.
[31] S. Suryawanshi, B. R. Chakravarthi, Findings of the shared task on troll meme classification

in Tamil, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Kyiv, 2021, pp. 126–132. URL: https://aclanthology.org/2021.dravidianlangtech-1.16.

[32] B. Lakshmanan, S. K. Ravindranath, Theedhum nandrum@dravidian-codemix-fire2020: A sentiment polarity classifier for youtube comments with code-switching between tamil, malayalam and english, 2020. arXiv:2010.03189.

[33] C. Vasantharajan, U. Thayasivam, Hypers@DravidianLangTech-EACL2021: Offensive language identification in Dravidian code-mixed YouTube comments and posts, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Kyiv, 2021, pp. 195–202. URL: https://aclanthology.org/2021.dravidianlangtech-1.26.

[34] S. Sai, Y. Sharma, Towards offensive language identification for Dravidian languages, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Kyiv, 2021, pp. 18–27. URL: https://aclanthology.org/2021.dravidianlangtech-1.3.

[35] J. Howard, S. Ruder, Fine-tuned language models for text classification, CoRR abs/1801.06146 (2018). URL: http://arxiv.org/abs/1801.06146. arXiv:1801.06146.

[36] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[37] S. Khanuja, D. Bansal, S. Mehtani, S. Khosla, A. Dey, B. Gopalan, D. K. Margam, P. Aggarwal, R. T. Nagipogu, S. Dave, S. Gupta, S. C. B. Gali, V. Subramanian, P. Talukdar, Muril: Multilingual representations for indian languages, 2021. arXiv:2103.10730.

[38] J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, M. Johnson, Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization, 2020. arXiv:2003.11080.

[39] M. Mosbach, M. Andriushchenko, D. Klakow, On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines, 2021. arXiv:2006.04884.