IIITDWD@TamilNLP-ACL2022: Transformer-based approach to classify abusive content in Dravidian Code-mixed text

Shankar Biradar and **Sunil Saumya**
Department of Computer Science and Engineering
Indian Institute of Information Technology
Dharwad, Karnatka, India
(shankar,sunil.saumya)@iiitdwd.ac.in

Abstract

Identifying abusive content or hate speech in social media text has raised the research community's interest in recent times. The major driving force behind this is the widespread use of social media websites. Further, it also leads to identifying abusive content in low-resource regional languages, which is an important research problem in computational linguistics. As part of ACL-2022, organizers of DravidianLangTech@ACL 2022 have released a shared task on abusive category identification in Tamil and Tamil-English code-mixed text to encourage further research on offensive content identification in low-resource Indic languages. This paper presents the working notes for the model submitted by IIITDWD at DravidianLangTech@ACL 2022. Our team competed in Sub-Task B and finished in 9^{th} place among the participating teams. In our proposed approach, we used a pre-trained transformer model such as Indic-bert for feature extraction, and on top of that, SVM classifier is used for stance detection. Further, our model achieved 62 % accuracy on code-mixed Tamil-English text.

1 Introduction

Many people from various demographics and linguistic backgrounds have been using social media sites to exchange information and interact with others. Further, these speakers tend to combine their mother tongue with a second language during the conversation. This leads to code-mixed text; code-mixing refers to two or more languages appearing one after another during a conversation (Poplack and Walker, 2003). Monitoring code-mixed content from social media sites has caught the research community's interest in natural language processing. Currently, many social media networks use a manual content screening method to deal with abusive content posted by the users. Here a human reviewer will go through the user posts and determine whether they violate the norms or not (Mandl et al., 2020; Biradar et al., 2022). However, as the number of social media users grows, massive amounts of data are generated, making it virtually difficult to monitor every data point personally. As a result, the manual technique of dealing with abusive content has become unsuccessful. Further, it increased the demand for automated abusive language identification models in a social media text, which is largely code-mixed.

The existing models have been trained on high-resource monolingual languages such as English and Hindi. Further, due to the complexity induced by code-mixing at different language levels, models trained with monolingual text failed to identify objectionable features in the code-mixed text. As a result, identifying abusive content in Indic languages poses a significantly greater problem for the NLP community. Hence identifying abusive content in low-resource Dravidian languages such as Tamil, Kannada, and Malayalam is made more difficult due to a lack of pre-trained models and a scarcity of training data to further train models.

To bridge this gap, DravidianLangTech@ACL 2022(Priyadharshini et al., 2022) organizers have provided a gold standard data set for abusive content identification in Dravidian languages such as Tamil and Tanglish Code-mixed text. The task's objective is to identify abusive categories from YouTube comments at the sentence/comment level. The original task is divided into two sub-Tasks: Sub-Task A involves sentence level abusive category detection from monolingual Tamil script, and Sub-Task B involves comment level abusive category identification from code-mixed Tamil-English text. Our team has participated in Sub-Task B and secured 9^{th} rank among the participating teams, and this paper presents working notes of our presented model.

The remainder of the paper is organized as follows: Section 2 provides a review of existing work,

Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, pages 100 - 104

Section 3 provides insight into the suggested model, and Section 4 concludes by offering information about outcomes.

2 Litrature review

The subject of automatic detection of hostile and harmful information from social media has attracted the interest of many researchers and practitioners from industry and academia. However, most of the past research has focused on high-resource languages. Previous attempts have been made to develop hate speech detection models in English, German (Mandl et al., 2019), and Italian (Corazza et al., 2020). But on the other hand, low-resource Indic languages are rarely explored. The first such attempt was made by (Bohra et al., 2018), they have created annotated corpus in Hindi-English code mixed text. They used traditional machine learning models to classify features extracted from the data set, such as character n-grams, word n-grams, punctuation, negation words, and hate lexicons. (Mathur et al., 2018) used a CNN-based transfer learning approach to detect abusive tweets in Hindi-English code-mixed text. They also introduced the HEOT data set and the Profanity Lexicon Set.

Abusive content identification in Indic languages is also the topic of a few shared tasks. Chakravarthi et al. created a shared task in low resource code-mixed Dravidian languages like Tamil-English, Malayalam-English, and Kannada-English (Chakravarthi et al., 2021). The objective of the task is to identify abusive content in a social media text. The shared task presents a new gold standard corpus for abusive language identification of code-mixed text in three Dravidian languages: Tamil-English (Chakravarthi et al., 2020b), Malayalam-English (Chakravarthi et al., 2020a), and Kannada-English (Hande et al., 2020). (Dowlagar and Mamidi, 2021) built a transformer-based transliteration and class balancing loss model to identify abusive content from code-mixed Dravidian languages. TIF-DNN, a transformer-based interpretation and feature extraction model to identify abusive content in Hindi-English code-mixed text, is built by (Biradar et al., 2021)

3 Data and methods

3.1 Task and data set information

We have taken the data set from Dravidian-LangTech@ACL 2022. As part of the competition, organizers have provided two sub-Tasks. Sub-Task A: comment/post-level abusive categories identification in monolingual Tamil text. Sub-Task B: Given a code-mixed text in Tamil-English, the user must identify abusive categories at the post/comment level. Our team took part in Sub-Task B, which involves the identification of abusive categories in Tamil-English code-mixed text. According to the task coordinators, Tamil-English data is gathered from YouTube comments (Priyadharshini et al., 2022). The organizers have provided train, validation, and test data sets. The train data set contained 5948 comments and labels, the validation data set contained 1488 comments, and the test data set contained 1857 comments. Details of the data set are provided in table 1.

3.2 Model description

The model architecture is divided into three steps in our proposed approach, as indicated in Fig 1. The model is comprised of an initial data pre-treatment stage, a transformer-based feature extraction layer, and an outer classification layer. The succeeding subsections will provide a complete details of each these stages.

3.3 Data pre-processing

The data collected from the organizers contains a lot of extraneous information. A few data preparation processes were performed on the text and label fields to make the data appropriate for model building. We removed digits, special characters, hyperlinks, and Twitter user handles from the data set because they are not useful for abusive content detection. Furthermore, the data provided by the user is taken from social media sites, and social media data does not follow grammatical rules. Lemmatization is carried out to convert the data to its usable basic form. Converting upper case text to lower case is also done to avoid redundant words. We used the NLTK toolbox from the python library (Bird et al., 2009) to perform these pre-processing steps.

Next, tokens are created by passing pre-processed text through a tokenizer. For this purpose, we employed the IndicBERT tokenizer[1] in our proposed approach. Additional padding and masking are applied on tokenized data to manage varying length sentences. At the end of the pre-treatment stage, we generate tokenized padded data,

[1]https://indicnlp.ai4bharat.org/indic-bert/

	None	Misandry	Xenop -hobia	Counter- speech	Hope- speech	Trans phobic	misog -yny	homop -hobia	Total
Train	3720	830	297	348	213	157	211	172	5948
Validation	919	218	70	95	53	40	50	43	1488
Test	1142	292	95	88	70	58	57	56	1857
Total	5781	1340	462	531	336	255	318	271	9293

Table 1: Data set distribution

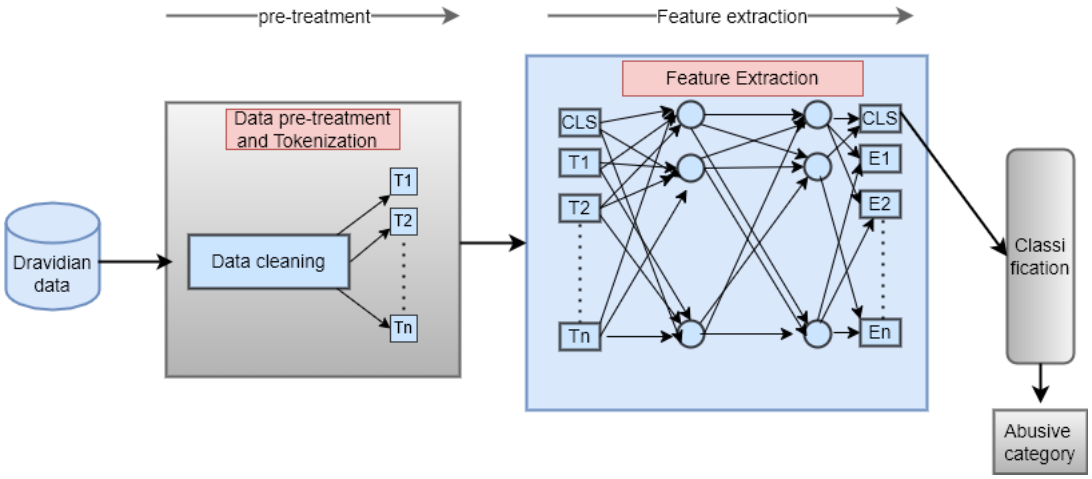

Figure 1: Overall Architecture of proposed system

which will be used as input for the feature extraction stage.

3.4 Feature extraction

Our proposed approach used the transformer-based IndicBERT (Kakwani et al., 2020) model for feature extraction. IndicBERT is a multilingual ALBERT model covering 12 main Indian languages: Assamese, Bengali, English, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil, and Telugu. It was trained on large-scale corpora. IndicBERT has fewer parameters than other public models like mBERT and XLM-R, yet it performs very well on various tasks. The architecture of IndicBERT is similar to that of the original BERT (Devlin et al., 2018). BERT is a non-regression model consisting of a transformer layer. Here transformer part of the model acts as an attention mechanism through which the model can learn contextual information from the data. We have used embedding from the CLS token in our proposed model, which gives full-sentence embedding. Embeddings are then passed through the outer classification layer for Stance detection.

3.5 Classification Layer

A conventional support vector machine (SVM) classifier is used for stance detection in the classification layer. SVM is built for binary classification problems and does not natively support multi-class classification problems. However, data provided by the organizers consist of linearly separable data with eight different classes. As a result, we utilized the One-vs-rest method from the Scikit-learn Python package[2] to transform a multi-class problem into a binary classification problem. Our experiment built a linear SVM classifier with ten-fold cross-validation using the sklearn SVM.LinearSVC model type and the OnevsRestClassifer wrapper. Experiment findings show that the penalty parameter value "1" and kernel type "linear" generate the best outcomes for the proposed model. Experimental trials determine these hyper-parameter values. The model uses embedding from IndicBERT as input and outputs one of the eight abusive categories. Implementation details of the proposed model are provided in GitHub repository[3].

[2]https://scikit-learn.org/stable/

[3]https://github.com/shankarb14/dravidian-codemix/blob/main/IndicBERT

	Precision	Recall	F1-score
Counter speech	0.55	0.32	0.40
Homo phobia	0.34	0.51	0.41
Hope speech	0.19	0.36	0.25
Misandry	0.63	0.69	0.66
Misogyny	0.05	0.27	0.09
None	0.89	0.81	0.85
Transphobic	0.21	0.40	0.27
Xenophobic	0.59	0.73	0.63
Accuracy	-	-	0.73

Table 2: Classification report for proposed model

4 Results

In the competition, teams were ranked based upon macro averaged Precision, macro averaged Recall and macro averaged F1-Score across all the classes. Our suggested model came in 9^{th} place for abusive category recognition on a code-mixed Tamil-English data set among the participating teams. Table 3 shows the top-five rated teams and the performance of our proposed model. From the table, the performance of our model is indicated in bold letters. According to the table, our model ranks second among the top-performing models with an accuracy of 62%. However, our model underperformed in identifying some of the abusive categories, and as a result of this unevenness, the overall model macro F1 score has decreased to 18.7. In addition, our model scored lower in the macro F1 score since it did not capture some abusive traits such as Trans-phobic, Hope-speech, and misogyny, as illustrated in table 2. The absence of sufficient training data in the categories mentioned contributed to our model's poor performance; nevertheless, our model's performance can be improved further by balancing the overall data set across all categories.

5 Conclusion and future enhancement

Our work presented a model proposed by team IIITDWD for detecting abusive categories in Tamil-English code mixed text as part of the shared task DravidianLangTech@ACL 2022. Our proposed model came in 9^{th} place among the participating teams, with a significant accuracy value of 62%. In the proposed model, we employed the transformer-based IndicBERT, trained on Indic languages, to extract features for classification with improved results. We can further improve the model performance by fine-tuning the model on Dravidian languages and including domain-specific embeddings.

Team name	Acc	m_F1
abusive-checker	0.65	0.41
GJG_TamilEnglish _deBERTa	0.60	0.35
umuteam	0.59	0.35
pandas	0.52	0.34
Optimize_Prime_Tamil _English_Run2	0.45	0.29
IITDWD	**0.626**	**0.187**

Table 3: Top performing models

References

Shankar Biradar, Sunil Saumya, and Arun Chauhan. 2021. Hate or non-hate: Translation based hate speech identification in code-mixed hinglish data set. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 2470–2475. IEEE.

Shankar Biradar, Sunil Saumya, and Arun Chauhan. 2022. Combating the infodemic: Covid-19 induced fake news recognition in social media networks. *Complex & Intelligent Systems*, pages 1–13.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

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, 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.

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.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Parameswari Krishnamurthy, Elizabeth Sherly, et al. 2021. Proceedings of the first workshop on speech and language technologies for dravidian languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*.

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2020. A multilingual evaluation for online hate speech detection. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–22.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Suman Dowlagar and Radhika Mamidi. 2021. Offlangone@ dravidianlangtech-eacl2021: Transformers with the class balanced loss for offensive language identification in dravidian code-mixed text. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 154–159.

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.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, 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.

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*, pages 29–32.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th forum for information retrieval evaluation*, pages 14–17.

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.

Shana Poplack and James A Walker. 2003. Pieter muysken, bilingual speech: a typology of code-mixing. cambridge: Cambridge university press, 2000. pp. xvi+ 306. *Journal of Linguistics*, 39(3):678–683.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumar Kumaresan. 2022. Findings of the shared task on Abusive Comment Detection in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.