HTMS@DravidianLangTech 2025: Fusing TF-IDF and BERT with Dimensionality Reduction for Abusive Language Detection in Tamil and Malayalam

Bachu Naga Sri Harini[1], Kankipati Venkata Meghana [1], Kondakindi Supriya[1], S Tara Samiksha [1] Premjith B[1],
[1]Amrita School of Artificial Intelligence, Coimbatore,
Amrita Vishwa Vidyapeetham, India,
{cb.sc.u4aie24010,cb.sc.u4aie24022,cb.sc.u4aie24025}@cb.students.amrita.edu,
cb.sc.u4aie24045@cb.students.amrita.edu, b_premjith@cb.amrita.edu

Abstract

Detecting abusive and similarly toxic content posted on a social media platform is challenging due to the complexities of the language, data imbalance, and the code-mixed nature of the text. In this paper, we present our submissions for the shared task on abusive Tamil and Malayalam texts targeting women on social media—DravidianLangTech@NAACL 2025. We propose a hybrid embedding model that integrates embeddings generated using term frequency-inverse document frequency (TF-IDF) and BERT. To get rid of the differences in the embedding dimensions, we used a dimensionality reduction method with TF-IDF embedding. We submitted two more runs to the shared task, which involve a model based on TF-IDF embedding and another based on BERT-based embedding. The code for the submissions is available at https://github.com/Tarrruh/NLP_HTMS.

1 Introduction

Social media platforms are prevalent, and each of them has its own properties. People use these platforms to articulate their opinions on various topics. YouTube is a popular platform, which has a limited set of rules and regulations to control the comments posted by the users for videos. More freedom in expressing the ideas allowed the users to comment and interact using toxic content (Chakravarthi et al., 2023; Kavitha et al., 2020). Abusive language is one of the ways to spread toxic content, and it affects the mental well-being of others. Therefore, it is critical to maintain a healthy online environment.

Detecting abusive content from comments posted on social media and related interactions is a challenging task, mostly due to the properties of the language (Justen et al., 2022). Imbalance in data (Muzakir et al., 2022) and the presence of multilingual and code-mixed data make the detection more challenging (Kogilavani et al., 2023; Aporna et al., 2022). Researchers have utilised diverse methodologies to identify and alleviate harmful information on social media platforms. Transformer-based models, especially those based on BERT, have become extensively utilised for this task (Kalraa et al., 2021). Additionally, other deep learning algorithms like recurrent neural networks (RNNs) and their variants have gained widespread use (Mahmud et al., 2024; Darmawan et al., 2023).

In this paper, we discuss our submission to the shared task on abusive Tamil and Malayalam text targeting women on social media—DravidianLangTech@NAACL 2025 (Rajiakodi et al., 2025). We proposed a machine learning model by fusing embeddings generated using term frequency-inverse document frequency (TF-IDF) and BERT. To balance the dimensionality of TF-IDF and BERT embeddings, we reduced the TF-IDF embeddings, which are generally very high in dimension, to lower dimension using the random kitchen sink (RKS) algorithm (Sathyan et al., 2018).

The paper is structured outlined as follows. Section 2 reviews pertinent literature, Section 3 delineates the datasets employed, Section 4 elucidates the methodology, Section 5 discusses the analysis of experiments and results, and Section 6 concludes by summarizing findings.

2 Literature Review

Various machine learning and deep learning methods were developed to determine whether a comment in Dravidian language texts contains abusive content or not. In (Prasanth et al., 2022), the authors presented a Support Vector Machine (SVM) model that used a feature vector made with Term Frequency-Inverse Document Frequency (TF-IDF) along with character-level analysis and the Random Kitchen Sink (RKS) algorithm. The authors of (Subramanian et al., 2023) proposed adapter-

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 239–243

based multilingual transformer models based on MuRIL, XLM-RoBERTa, and mBERT to classify abusive comments in Tamil. The authors developed both conventional fine-tuned and adapter-based versions of the aforementioned models for classification. They observed that MuRIL had outperformed other models in this task with a detection accuracy of 74.7%. In (Chakravarthi et al., 2023), the authors employed machine learning algorithms such as naive Bayes, SVM, decision trees, random forests, and logistic regression with TF-IDF, Bag of Words (BoW), and FastText features to detect abusive comments in low-resource languages. In addition, the authors proposed deep learning models such as BiLSTM, BiLSTM with attention, mBERT, and XLM for this task. (Priyadharshini et al., 2022a) reported the submissions of the participants of the shared task. The participants used various machine learning and deep learning models for this task. The machine learning models include logistic regression, SVM, gradient boost classifiers, K-nearest neighbours, and ensemble models. Multi-layered perceptron (MLP), recurrent neural network (RNN), and Long Short-Term Memory (LSTM) are the examples of deep learning models submitted by the participants. In addition, the efficacy of transformer models such as mBERT, MuRIL, and XLM-R0BERTa was also investigated. The paper (Priyadharshini et al., 2022a) describes the systems developed for detecting abusive comments in Tamil, Tamil-English, and Telugu-English data. The authors reported that the majority of the systems used BERT-based models for feature extraction. There are models based on LSTM and traditional machine learning classifiers using TF-IDF features and word2vec embeddings. (Priyadharshini et al., 2023) and (Priyadharshini et al., 2022b) discuss different machine learning and deep learning models developed for detecting abusive content in Dravidian languages.

3 Dataset Description

The training dataset consists of the classification of abusive and non-abusive comments in Tamil and Malayalam. The organisers provided not only the training data, but also development (validation) data and test data without labels. We combined the training and development sets to train the model. Table 1 describes the dataset used for building the model.

Data Split	Number of Data Points	
	Tamil	**Malayalam**
Train	2,790	2,933
Validation	598	629
Test	598	629

Table 1: Dataset statistics showing the number of data points in each split

4 Methodology

In this work, we experimented with three models. These three models differ in the embeddings used. Following are the models we considered.

1. **TF-IDF embedding-based method**: In this model, we used TF-IDF embeddings. Here, we considered a maximum of 5000 unique words in the feature set. Therefore, the embedding model considered 3,000 most frequently occurring words for the study. We used the extracted features to train the classifier model, which we built using logistic regression. This is our Run 3 submission.

2. **A fused TF-IDF and BERT embeddings**: In this model, we computed both TF-IDF and BERT embeddings for the input text. We restricted the number of feature words to 5,000 based on their importance, similar to Run 3. TF-IDF learns the importance of a word in a sentence, whereas BERT generates an embedding by considering the context of the words. We generated the BERT embeddings using the 'bert-base-uncased' model. Because the TF-IDF and BERT embeddings have very different dimensions, the random kitchen sink (RKS) algorithm was used to reduce the size of the TF-IDF embeddings so that they are the same size as the BERT embeddings. RKS is a feature-mapping technique based on the Radial Basis Function that turns data into a space with desired dimensions. We concatenated these embeddings to obtain the representation of the input text. Machine learning classifiers were employed for training the models using these embeddings. This is our Run 2 submission.

3. **BERT-based method**: Here, we used the 'bert-base-uncased' model to generate the embeddings of the next text, which was further

fed into the machine learning classifiers to train the classification model. This is our Run 1 submission.

The flow-diagram of the model is illustrated in Figure 1.

5 Experimentation and Results

Based on the methodology, we divided the experiments into three.

5.1 TF-IDF embedding-based method

In this method, we used TF-IDF embeddings as feature representations. The TF-IDF algorithm converts text data into embeddings based on word frequency while reducing the influence of commonly occurring words. To control complexity and computational cost, we limit the number of features to 5,000 (max features = 5,000). We performed a 5-fold cross-validation (n splits = 5) with shuffling enabled (shuffle = True) to ensure a robust and reproducible evaluation of the model. The TF-IDF embeddings of the training data were used to train a logistic regression classifier. The maximum number of iterations was set to 1000 to ensure convergence, and the random state was kept constant to get stable results.

5.2 A fused TF-IDF and BERT embeddings

In this method, for feature extraction, we employed two methods: TF-IDF and BERT embeddings. We used the RKS algorithm with a gamma value of 1.0 to reduce the dimensionality of the TF-IDF embeddings from 5,000 features to 512. Simultaneously, we used the pre-trained bert-base-uncased model to generate contextual sentence embeddings. To get contextual representations, the input text was broken up into tokens that could be up to 512 characters long and processed in batches of 32. This was done on the T4 GPU of Google Colab. We obtained the final sentence embeddings by averaging the token representations from the last hidden layer. We then fused these two embeddings by concatenating the reduced TF-IDF and BERT embeddings horizontally. To make sure of the model's generalization capability, we did a 5-fold cross-validation with data. We trained a random forest classifier on the fused embeddings, using default hyperparameters.

5.3 BERT-based method:

We used deep contextual representations from the pre-trained BERT model, bert-base-uncased (with 768 dimensions), in this method. The input comments were processed to generate sentence embeddings using the BERT tokenizer, with truncation and padding applied for a maximum sequence length of 512 tokens. We processed the data efficiently with a batch size of 32 and passed the text through the BERT model. We computed the mean of the token representations from the last hidden layer to obtain a fixed-size embedding for each sentence. To train the Random Forest classifier, we used a 5-fold cross-validation method that included stratified K-Fold with shuffling and a random seed of 42 to make sure the results would be the same every time. The classifier was initialized with 100 decision trees (n estimators) in the forest and a maximum tree depth set to None, ensuring that the trees grew until all leaves were pure (or contained less than the minimum number of samples, which was set to 1).

5.4 Results

The training of the model with BERT embeddings and a random forest classifier demonstrated the potential for deep contextualisation from the BERT model. The performance of this approach varied depending on the dataset and the fold in the cross-validation. The BERT embeddings helped the model understand the hidden meanings of words and their context, which made it better at dealing with complicated language structures and subtle textual relationships. However, the computational cost of BERT made this method slower, particularly for larger datasets.

When combined with logistic regression, the TF-IDF approach trained significantly faster and required less computational power. Statistically, this method obtained the best results for the training set (almost the same as the method that used fused embedding).

The approach, which combined TF-IDF and BERT embeddings, yielded the most promising results by leveraging the strengths of both methods. TF-IDF captured word frequency and distribution information, while BERT provided semantic and contextual understanding. The fused embeddings allowed the model to capture both shallow and deep linguistic features, leading to more accurate predictions. This method performed well compared to

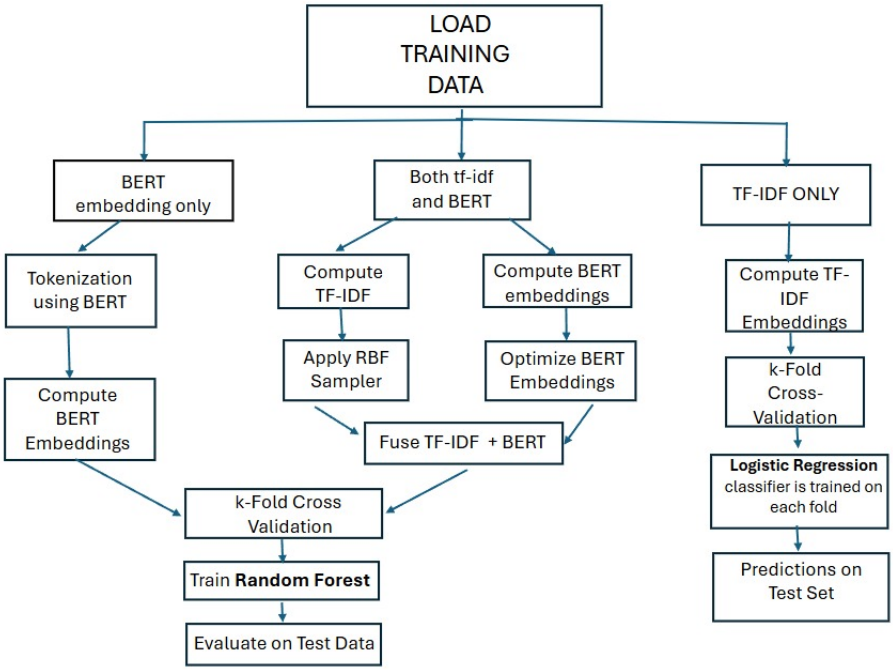

Figure 1: Block diagram explaining proposed methodology used for the task

Model	Precision	Recall	F1-score
TF-IDF	0.62	0.63	0.62
BERT	0.59	0.59	0.59
TF-IDF + BERT	0.62	0.62	0.62

Table 2: Macro-Averaged Precision, Recall, and F1-score for Different Models

Tamil	Malayalam
0.50	0.49

Table 3: Macro F1-scores of BERT embedding for predicted dataset

both individual models.

Test Data Predictions:

All three models made predictions on a separate test dataset after training. The predicted dataset had different results for the three models. The BERT-trained model had the most evenly distributed dataset. The hybrid model found very few abusive comments and labeled most of them as not abusive. In contrast, the TF-IDF model predominantly labels the entire test dataset for both Tamil and Malayalam as non-abusive. The performance of the proposed methodologies are shown in Tables 2 and 3.

6 Limitations

These languages are low-resource languages, hence we had lesser data points to train the model. We used fewer embeddings, so the model couldn't uncover all of the patterns required to classify the comments, which affected results. The approach for generating the feature representation was one of the reasons for obtaining low performance scores. Fine-tuning a pretrained model or using more data points to enable the model to uncover more hidden patterns could improve the model performance.

7 Conclusion

This paper presents the system description of the HTMS team for detecting abusive comments in Malayalam and Tamil against women. We experimented with three distinct methods for classifying abusive and non-abusive comments in Dravidian languages. We trained a model with fused embedding, which outperformed other models. After training and testing the datasets with these models, we observed that the BERT-only model provided excellent results. However, the TF-IDF model classified all comments as non-abusive, leading to highly inaccurate results.

References

Amena Akter Aporna, Istinub Azad, Nibraj Safwan Amlan, Md Humaion Kabir Mehedi, Mohammed Julfikar Ali Mahbub, and Annajiat Alim Rasel. 2022. Classifying offensive speech of bangla text and analysis using explainable ai. In *International Conference on Advances in Computing and Data Sciences*, pages 133–144. Springer.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Shubanker Banerjee, Manoj Balaji Jagadeeshan, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Sean Benhur, and John Philip McCrae. 2023. Detecting abusive comments at a fine-grained level in a low-resource language. *Natural Language Processing Journal*, 3:100006.

Bagus Tri Yulianto Darmawan, Bassamtiano Renaufalgi Irnawan, and Yoshimi Suzuki. 2023. Indonesian hate speech and abusive tweets classification with deep learning pre-trained language models. In *2023 6th International Conference of Computer and Informatics Engineering (IC2IE)*, pages 30–35. IEEE.

Lennart Justen, Kilian Müller, Marco Niemann, and Jörg Becker. 2022. No time like the present: Effects of language change on automated comment moderation. In *2022 IEEE 24th Conference on Business Informatics (CBI)*, volume 1, pages 40–49. IEEE.

Sakshi Kalraa, Mehul Agrawala, and Yashvardhan Sharmaa. 2021. Detection of threat records by analyzing the tweets in urdu language exploring deep learning transformer-based models. In *Proc. CEUR Workshop*, pages 1–7.

KM Kavitha, Asha Shetty, Bryan Abreo, Adline D'Souza, and Akarsha Kondana. 2020. Analysis and classification of user comments on youtube videos. *Procedia Computer Science*, 177:593–598.

Shanmuga V Kogilavani, Subramanian Malliga, KR Jaiabinaya, M Malini, and M Manisha Kokila. 2023. Characterization and mechanical properties of offensive language taxonomy and detection techniques. *Materials Today: Proceedings*, 81:630–633.

Tanjim Mahmud, Tahmina Akter, Mohammad Kamal Uddin, Mohammad Tarek Aziz, Mohammad Shahadat Hossain, and Karl Andersson. 2024. Machine learning techniques for identifying child abusive texts in online platforms. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE.

Ari Muzakir, Kusworo Adi, and Retno Kusumaningrum. 2022. Classification of hate speech language detection on social media: Preliminary study for improvement. In *International Conference on Networking, Intelligent Systems and Security*, pages 146–156. Springer.

SN Prasanth, R Aswin Raj, P Adhithan, B Premjith, and Soman Kp. 2022. Cen-tamil@ dravidianlangtech-acl2022: Abusive comment detection in tamil using tf-idf and random kitchen sink algorithm. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 70–74.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Malliga Subramanian, Kogilavani Shanmugavadivel, Premjith B, Abirami Murugappan, Sai Prashanth Karnati, Rishith, Chandu Janakiram, and Prasanna Kumar Kumaresan. 2023. Findings of the shared task on Abusive Comment Detection in Tamil and Telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages DravidianLangTech 2023*. Recent Advances in Natural Language Processing.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022a. Overview of abusive comment detection in tamil-acl 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022b. Overview of abusive comment detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Ruba Priyadharshini, Rajameenakshi J, Kathiravan Pannerselvam, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Paul Buitelaar, Bhavanimeena K, Jananayagam V, and Kishore Kumar Ponnusamy. 2025. Findings of the Shared Task on Abusive Tamil and Malayalam Text Targeting Women on Social Media: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Dhanya Sathyan, Kalpathy Balakrishnan Anand, Aravind Jaya Prakash, and Bhavukam Premjith. 2018. Modeling the fresh and hardened stage properties of self-compacting concrete using random kitchen sink algorithm. *International journal of concrete structures and materials*, 12:1–10.

Malliga Subramanian, Kogilavani Shanmugavadivel, Nandhini Subbarayan, Adhithiya Ganesan, Deepti Ravi, Vasanth Palanikumar, and Bharathi Chakravarthi. 2023. On finetuning adapter-based transformer models for classifying abusive social media tamil comments.