shimig@DravidianLangTech2025: Stratification of Abusive content on Women in Social Media

Gersome Shimi
Madras Christian College,
Chennai, India
gshimi2022@gmail.com

C.Jerin Mahibha
Meenakshi Sundararajan
Engineering College,
Chennai, India
jerinmahibha@msec.edu.in

Durairaj Thenmozhi
Sri Sivasubramaniya Nadar
College of Engineering,
Chennai, India
theni_d@ssn.edu.in

Abstract

The social network is a trending medium for interaction and sharing content globally. The content is sensitive since it can create an impact and change the trends of stakeholder's thought as well as behavior. When the content is targeted towards women, it may be abusive or non-abusive and the identification is a tedious task. The content posted on social networks can be in English, code mix, or any low-resource language. The shared task Abusive Tamil and Malayalam Text targeting Women on Social Media was conducted as part of DravidianLangTech@NAACL 2025 organized by DravidianLangTech. The task is to identify the content given in Tamil or Malayalam or code mix as abusive or non-abusive. The task is accomplished for the South Indian languages Tamil and Malayalam using pretrained transformer model, BERT base multilingual cased and achieved the accuracy measure of 0.765 and 0.677.

1 Introduction

According to Statista, a Statistics portal for market data, market research, and market studies, the number of Internet users in the year 2024 is 5.44 billion. It is one third of the world's population and the number of YouTube users is approximately 2504 million as of April 2024[1]. People use social media platforms such as YouTube, Twitter, Instagram, Reddit, and Facebook to share their opinions, beliefs, and interests in all the state of affairs. It can be used positively in e-commerce, information transfer, advertisements, politics, hobbies, testimonies, education and training, recent happenings, etc. Alternatively, it can lead to the spread of hate speech and on-line harassment, which is termed cyberbully. It should be identified since it will cause psychological impact for the stakeholder even to depression.(Sari et al., 2022)

The trendy digital platforms, social media impact people at various levels, even the political and business scenario can be altered within next few hours in par with the comments posted. It can also target a particular individual or a group of individuals.(Priyadharshini et al., 2022) Hate speech and offensive language can harm various groups and can end with social problems, thus makes detection an essential task to reduce crime and promote harmony. Although significant research exists for languages like English, Dravidian languages such as Tamil and Malayalam lack focus.(Mahibha et al., 2021) When abusive words are aimed at gender, particularly on women it is defined as sexism. As it is one of the alarmed need of the social media, focusing on this issue, DravidianLangTech@NAACL2025 organized by DravidianLangTech initiated a shared task to identify abusive Tamil and Malayalam Text targeting Women on Social Media,

2 Related Work

Hate speech is a complicated and multifaceted issue that creates serious and widespread implications for human rights and the rule of law on democratic society. Addressing and preventing online hate speech presents particular challenges. The ongoing nature and effects of this issue have been recorded by the oversight bodies of the Council of Europe and various international organizations[2].

Social networks are rapidly expanding with different content in various low resource languages allows stakeholder to express their opinions with few restrictions. Most social media platforms allow users to share and express their thoughts to collect user comments and posts to offer channels of personalized interest. However, they are also used for negative actions, such as spreading ru-

[1]https://www.statista.com/topics/1145/internet-usage-worldwide

[2]https://www.coe.int/en/web/combating-hate-speech/what-is-hate-speech-and-why-is-it-a-problem-

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 487–492

mors and intimidating people with offensive words. Abusive language has attracted a lot of attention as social media platforms have become more popular.(Barman and Das, 2023)

2.1 Low Resource Languages

Low resource languages own complexity in terms of variation in writing and spoken style, unavailability of resources, and corpus. Words usage have different meanings when used in different communities. LLM (Large Language Model) is capable of grasping multiple languages and adapting to different contexts (Zhong et al., 2024). Low resource language, Kannada, Malayalam, Telugu, and Tamil, obtains less attention due to unavailability of the corpus. The ensemble transformer model is applied for the classification, obtain the f1 score of 0.66 and 0.72 for Kannada code mix and Malayalam, respectively. (Roy, 2024)

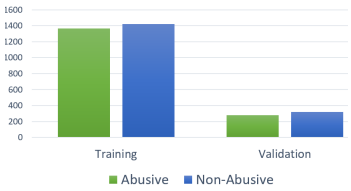

Figure 1: Distribution of Dataset-Tamil

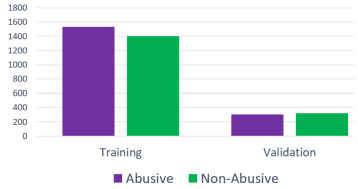

Figure 2: Distribution of Dataset-Malayalam

Language	Text	Label
Tamil	என்ன திமிர் இந்த பொண்ணுக்கு.....மக்களே இன்னும் இவளுடைய வீடியோக்களை பார்த்தால் உங்களை விட முட்டாள்கள் யாருமில்லை	Abusive
Tamil	You tube ல இப்படி எல்லாம் முன்னேற்றம் நம் சேனலையும வந்து முன்னேற்றுங்கள்!!	Non-Abusive
Malayalam	ചുമ്മാതല്ല ഇവളെ പണ്ട് ലവർ അലക്കി വിട്ടത്...വായിൽ കിടക്കുന്ന നാക്ക് അങ്ങനെ ആണ്	Abusive
Malayalam	ഒരു സിനിമയിൽ ജഗതി മദാമ്മയായിട്ട് വരുന്നുണ്ടല്ലോ. പെട്ടന്ന് അതോർമ്മ വന്നു	Non-Abusive

Figure 3: Sample Dataset

2.2 Preprocessing

Most instances are misidentified because of the presence of insignificant words. Preprocessing is an important process to feed the model with quality data in terms of size, improves the model performance.(Kumari, 2022) Text cleaning is the removal of insignificant words from the dataset sentence, it makes the content relevant for the supervised model to process the data. This can be done by eliminating stopword, noise, and encode consistency. For the prediction of multiclass classification, the transformer model outperformed with an accuracy score of 0.91.(Zerrouki and Benblidia, 2024)

2.3 mBERT Model in Classification

mBert is a modified BERT model, trained with 104 languages and takes advantage of grasping and processes several language data simultaneously.(Panchadara, 2024) Multilanguage content is vital in an electronically connected environment.The classification of offensive and non-offensive comments, the difficulty to detect it in multilingual context is discussed. Taking advantage of the multilingual BERT model, comments are classified in various languages English, Hindi, Telugu, Malayalam, Kannada, Greek, and Russian and achieved an accuracy score of 0.925.(Nandhini et al., 2024) The mBERT model for meme classification outperformed other transformer models and achieved an f1 score of 0.75, 0.95 and 0.71 for Tamil, Malayalam, and Kannada, respectively.(Ghanghor et al., 2021)

3 Proposed Approach

3.1 Problem Overview

Abusive Tamil and Malayalam Text targeting Women on Social Media is addressed as a binary classification problem. The goal is to predict unlabeled instances as Abusive (P=1) or Non-Abusive (P=0). According to the standard probability of supervised learning, the set of output P (0 or 1) is deterministically related to the input N (Text) to output P is denoted by the target function R:N->P. In addition, $N \epsilon n \; and \; P \epsilon p$. (Anthony, 2008) The model is built in such a way that

$$M(N) -> P \tag{1}$$

where P belongs to [0,1].

3.2 About Dataset

The YouTube comments dataset provided by DravidianLangTech, to perform the shared task of promoting Tamil and Malayalam Text targeting Women on Social Media-DravidianLangTech@NAACL 2025(Rajiakodi et al., 2025). The training, validation and evaluation datasets contain data in the low resource languages Tamil and Malayalam. The description

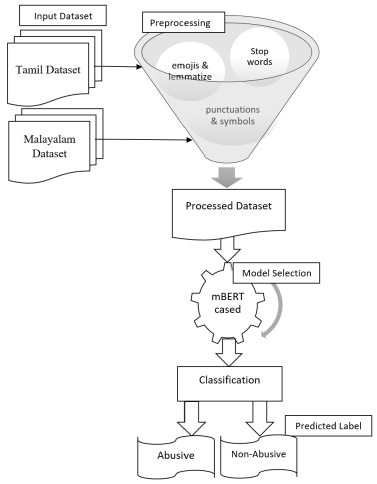

Figure 4: Methodology

Dataset	Abusive	Non-Abusive	Total
Training	1366	1423	2789
Validation	277	320	597
Testing	--	--	598

Table 1: Dataset Description-Tamil

of dataset Tamil and Malayalam is described in Table 1 and Table 2 respectively. The dataset is annotated as abusive or non-abusive according to the context of the comments targeted towards women in social media. The sample dataset of the shared task is shown in Figure 3.
Abusive- Content related to violence, abuse, mistreat, or neglect of women from intimate or dependent relationships or society.
Non-Abusive- Content without abusive context against women.
The distribution of the dataset in languages Tamil and Malayalam is shown in Figure 1 and Figure 2 respectively.

3.3 Preprocessing

Text preprocessing is the transformation of text by cleaning noise and preparing the text for further operation. The various techniques involved are removal of stop words, punctuation, special symbols, and numbers. Preprocessing helps the model to improve the performance of the model(Siino et al., 2024). The content of the dataset, and labels are converted to the lower case (codemix). The preprocessing techniques used are
Removal of noise- includes removal of special symbols, punctuations, emojis.
Tokenization- partitioning the sentences into words.
Lemmatization- Converting the original word to root word.
Removal of stop words- Removal of insignificant words from the instances.
Case conversion- Changing the comments to lowercase.

Dataset	Abusive	Non-Abusive	Total
Training	1530	1402	2932
Validation	303	325	628
Testing	--	--	629

Table 2: Dataset Description-Malayalam

3.4 Evaluation Metrics

The model is validation with the evaluation metrics accuracy, precision, and f1 score. Accuracy(Acc) measures the performance of the model, which is measured by the ratio no of instance to the correctly predicted instances.

$$Acc = \frac{No.of\,flawless\,prediction}{Total\,No.of input\,instance}$$

Prediction(Prec) measures the number of positive predictions that the model considers to be correct. It is calculated by division of the number of true positive(TP) prediction with the sum of true positive and false positive(FP) predictions.

$$Prec = \frac{TP}{TP+FP}$$

Recall(Rec) measures the number of positive prediction obtained by the model is correct. It is calculated by division of the number of true positive(TP) prediction with the sum of true positive, false negative(FN) prediction.

$$Rec = \frac{TP}{TP+FN}$$

True Positive(TP): The predicted output is Yes, and the actual output is Yes as well.
True Negatives(TN): The predicted output and the actual output is also No.
False Positives(FP): The predicted output is Yes, but the actual output is No.
False Negatives(FN): The predicted output is No, but the actual output is Yes.

Language	Accuracy	Precision	f1 score
Tamil	0.77	0.78	0.77
Malayalam	0.72	0.72	0.72

Table 3: mBERT cased-Development Results

3.5 Model selection and implementation

The BERT base multilingual cased is a powerful model, which can be used for 104 languages including Tamil and Malayalam.(Devlin et al., 2018) The articles from Wikipedia are used to train the mBERT model, its performance is based on the quality of the content of the language.(Wu and Dredze, 2020) It is enhanced with a particular training data from a single language, to another language, and enables it to work across different languages.(Nabiilah et al., 2024)

The mBERT model is used for the implementation of the shared task, the model is trained by training the parameters epoch=7, maxlength=256, batch size=16 and AdamW optimizer with the learning rate of 2e-5 and correct bias=True. The result is shown as confusion matrix in Figure 5 and 6. The result is evaluated with the evaluation metrics, accuracy, precision and recall scores of 0.77,0.78, and 0.77 respectively for Tamil. For Malayalam, we obtain the accuracy, precision and recall score of 0.72, 0.72, and 0.72 respectively for development dataset.

4 Results and Discussions

The implementation is performed in Google Colab using the Python programming language with the multilingual BERT model. The model is trained for 7 epochs by tuning the parameters maxlength=256, batch size=16, and AdamW optimizer. The AdamW parameters learning rate and correct bias are set to 2e-5 and True respectively. When the experiment is carried out by setting the correct bias to False for Malayalam language we got biased result for non-abusive label. The different runs are executed with the same parameters for the South Indian languages Tamil and Malayalam. The outperformed results can be viewed in confusion matrix in Figure 5 and 6. The result details of our team, given by the organizers are tabulated in Table 4 and our development result in Table 3. We noticed variation in the model performance related to the number of stop words also. The accuracy of the model dropped when the stopword is increased

Language	Runs	mF1 Score
Tamil	1	0.75
Tamil	2	0.765
Tamil	3	0.757
Malayalam	1	0.674
Malayalam	2	0.677

Table 4: Result Score- DravidianLangTech@NAACL 2025

when working with Tamil language and the reverse for Malayalam language. The usage of lemmatization dropped the accuracy of Tamil language.

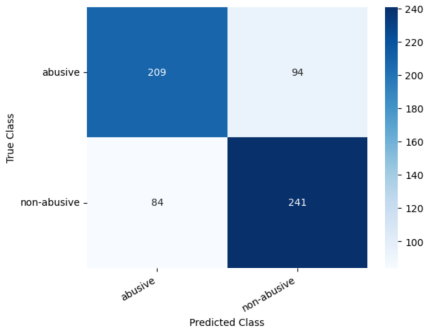

Figure 5: Confusion Matrix-Malayalam

5 Conclusion and Future Work

Usage of abusive text on social networks is a common violation and classification is a challenging task. The dataset shared by DravidianLangTech@NAACL 2025 to classify abusive Tamil and Malayalam Text targeting women in social media is used for the implementation of the model. The pretrained transformer model BERT base multilingual cased was used for the classification in all the runs. The model achieved the mF1 score of 0.765 and 0.677 for Tamil and Malayalam dataset respectively. When working with low resource languages the unavailability of stop words and dictio-

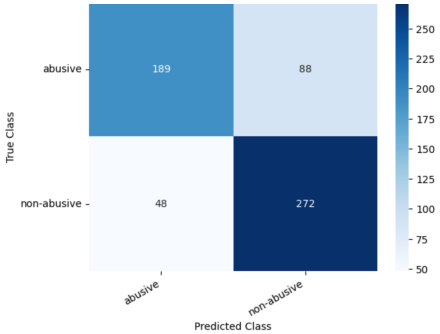

Figure 6: Confusion Matrix-Tamil

nary is another issue. Future research can focus on preprocessing with back translation of sentences, as well as the implementation of vectorization to improve the accuracy of the model.

6 Limitations

Although the implemented model performed well, it has certain limitations. The size of the training dataset is small, which limits the generalization that leads the model to struggle on an unseen dataset. To obtain an accurate result, the ambiguity of words and morphological complexity, which is fairly common in low resource languages such as Tamil and Malayalam, should be addressed.

References

Martin Anthony. 2008. Aspects of discrete mathematics and probability in the theory of machine learning. *Discrete applied mathematics*, 156(6):883–902.

Shubhankar Barman and Mithun Das. 2023. hate-alert@DravidianLangTech: Multimodal abusive language detection and sentiment analysis in Dravidian languages. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 217–224, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Nikhil Ghanghor, Parameswari Krishnamurthy, Sajeetha Thavareesan, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2021. Iiitk@ dravidianlangtech-eacl2021: Offensive language identification and meme classification in tamil, malayalam and kannada. In *Proceedings of the first workshop on speech and language technologies for dravidian languages*, pages 222–229.

Santoshi Kumari. 2022. Text mining and pre-processing methods for social media data extraction and processing. In *Handbook of research on opinion mining and text analytics on literary works and social media*, pages 22–53. IGI Global.

C Jerin Mahibha, Sampath Kayalvizhi, Durairaj Thenmozhi, and Sundar Arunima. 2021. Offensive language identification using machine learning and deep learning techniques.

Ghinaa Zain Nabiilah, Islam Nur Alam, Eko Setyo Purwanto, and Muhammad Fadlan Hidayat. 2024. Indonesian multilabel classification using indobert embedding and mbert classification. *International Journal of Electrical & Computer Engineering (2088-8708)*, 14(1).

PS Nandhini, R Karunamoorthi, P Mariappan, and S Revathi. 2024. Multilingual offensive language detection in social media content using bert-base-multilingual-cased model. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE.

Kiranmaye Panchadara. 2024. Enhancing named entity recognition in low-resource dravidian languages: A comparative analysis of multilingual learning and transfer learning techniques. *Journal of Artificial intelligence and Machine Learning*, 2(1):1–7.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. 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.

Pradeep Kumar Roy. 2024. Deep ensemble network for sentiment analysis in bi-lingual low-resource languages. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(1):1–16.

Tiara Intana Sari, Zalfa Natania Ardilla, Nur Hayatin, and Ruhaila Maskat. 2022. Abusive comment identification on indonesian social media data using hybrid deep learning. *IAES International Journal of Artificial Intelligence*, 11(3):895–904.

Marco Siino, Ilenia Tinnirello, and Marco La Cascia. 2024. Is text preprocessing still worth the time? a comparative survey on the influence of popular preprocessing methods on transformers and traditional classifiers. *Information Systems*, 121:102342.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual bert? *arXiv preprint arXiv:2005.09093*.

Khadidja Zerrouki and Nadjia Benblidia. 2024. Multilingual text preprocessing and classification for the detection of extremism and radicalization in social networks.

Tianyang Zhong, Zhenyuan Yang, Zhengliang Liu, Ruidong Zhang, Yiheng Liu, Haiyang Sun, Yi Pan,

Yiwei Li, Yifan Zhou, Hanqi Jiang, et al. 2024. Opportunities and challenges of large language models for low-resource languages in humanities research. *arXiv preprint arXiv:2412.04497*.