Necto@DravidianLangTech 2025: Fine-tuning Multilingual MiniLM for Text Classification in Dravidian Languages

Livin Nector Dhasan
IIT Madras / BS Degree
livin@study.iitm.ac.in

Abstract

This paper explores the application of a fine-tuned Multilingual MiniLM model for various binary text classification tasks, including AI-generated product review detection, abusive language targeting woman detection, and fake news detection in the Dravidian languages Tamil and Malayalam. This work was done as part of submissions to shared tasks organized by DravidianLangTech@NAACL 2025. The model was fine-tuned using both Tamil and Malayalam datasets, and its performance was evaluated across different tasks using macro F1-score. The results indicate that this model produces performance very close to the best F1 score reported by other teams. An investigation is conducted on the AI-generated product review dataset and the findings are reported.

1 Introduction

The advancement of natural language processing (NLP) models has significantly improved text classification capabilities. However, Dravidian languages, such as Tamil and Malayalam, remain underrepresented in NLP research. BERT-based models like mBERT, XLM-Roberta, and IndicBERT, have been demonstrating significant results in different classification tasks in the context of fake news detection(Luo and Wang, 2023; Tabassum et al., 2024) and abusive content detection(Hegde et al., 2023). To reduce the computational costs in the fine-tuning and the inference, smaller models of the BERT family with a lesser number of parameters such as DistilBERT(Sanh et al., 2020), MobileBERT(Sun et al., 2020) and TinyBERT(Jiao et al., 2019) are pre-trained using different knowledge distillation methods with a larger BERT based model as the Teacher and then fine-tuned for downstream tasks. These distilled models show near-comparable performance with fewer number of parameters.

The model used in this study, Multilingual MiniLM, was pre-trained using the Deep Self-Attention Distillation method by distilling the XLM-Roberta model (Wang et al., 2020). It outperforms similar distilled models such as DistilBERT, TinyBERT, and MobileBERT on various benchmarks. The multilingual nature of MiniLM made it a suitable choice for fine-tuning with Tamil and Malayalam data. This paper investigates the effectiveness of the Multilingual MiniLM model for diverse text classification tasks in these languages. By fine-tuning the model, specific challenges such as detecting AI-generated content, abusive language targeting women, and fake news are addressed.

2 Task Description

2.1 AI-Generated Product Review Detection (ai-gen)

This task addresses the growing concern of AI-generated product reviews, particularly in Tamil and Malayalam. As AI tools for content generation become more sophisticated, distinguishing between human-written and AI-generated reviews has become essential to ensure authenticity and reliability in consumer decision-making. The dataset poses it as a binary sentence classification problem, classifying the given product review text as "HUMAN" or "AI" containing data splits for both Tamil and Malayalam languages (Premjith et al., 2025). The data set does not include a development split for both languages, thus a development set is created from the training set using a stratified 80-20 train-dev split, which is used for fine-tuning.

2.2 Abusive Text Targeting Women on Social Media (abusive-woman)

This task focuses on classifying social media texts, particularly comments on YouTube, that are directed at women in a derogatory manner. Previ-

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 398–402

ously, abusive content classification in Tamil and Telugu languages are explored as Multi-class classification problems with the labels Homophobia, Misandry, Counter-speech, Misogyny, Xenophobia, and Transphobic in the shared tasks on RANLP-2023 and ACL-2022(Priyadharshini et al., 2023, 2022). The current dataset includes Tamil and Malayalam text, often containing code-mixed content. It is framed as a binary classification problem to detect the presence of abusive content targeting women with the labels "Abusive" and "Non-Abusive".

2.3 Fake News Detection (fake-news)

This task aims to identify fake news in Malayalam texts. Given the rapid spread of misinformation, the ability to detect fake news in regional languages is crucial for maintaining information integrity. The shared task consists of two datasets (Subramanian et al., 2024, 2025), one with binary classification labels as "Fake" and "Original"(**Task A**)(Subramanian et al., 2023) and the other dataset with multi-class classification labels as "Half-True", "False", "Partly-False" and "Mostly-False"(**Task B**)(Devika et al., 2024). Only the binary classification **Task A** is explored in this work.

3 Methodology

3.1 Data Preprocessing

The text data was preprocessed using the XLM-Roberta tokenizer to generate token embeddings and attention masks. The Multilingual MiniLM model uses the XLM-Roberta tokenizer as the former is a distilled version of the later model. Tokens were truncated and padded to a maximum length of 256. Labels were encoded as binary values for each task(Table 1).

Task	Negative (0)	Positive (1)
ai-gen-review	HUMAN	AI
abusive-woman	Non-Abusive	Abusive
fake-news	Original	Fake

Table 1: Task and Label Mapping

3.2 Model

The pre-trained checkpoint from Hugging Face, `microsoft/Multilingual-MiniLM-L12-H384`, is used as the base model for fine-tuning.

The MiniLM model architecture consists of 12 hidden layers, each with a hidden layer size of 384, totaling 21M[1] parameters.

3.3 Fine-Tuning

A classification head with a fully connected layer and softmax activation function was added on top of the base Multilingual MiniLM model using the `AutoModelForSequenceClassification` class from the transformers library by Hugging Face(Wolf et al., 2020). The model is trained using the `Trainer` API from the transformers library. Three models **ai-gen-review**, **abusive-woman** and **fake-news** were created as the result of the fine-tuning. Both the Tamil and Malayalam datasets are jointly used to fine-tune the models **ai-gen-review** and **abusive-woman** and the Malayalam dataset is used for fine-tuning the model **fake-news**. The best model was selected based on the f1-score evaluated during fine-tuning.

Models	Batch Size	No. of Epochs
ai-gen-review	128	6
abusive-woman	128	6
fake-news	256	9

Table 2: Hyperparameter configuration used for fine-tuning the model on different tasks.

4 Results

The fine-tuned models are then evaluated in Tamil and Malayalam for different tasks using F1-Score with macro averaging as the evaluation metric. The evaluation results on Tamil and Malayalam language Tasks are presented in Table 3 and Table 4 respectively.

The fine-tuned checkpoints of the models for AI-generated product review detection (Tamil & Malayalam), abusive text targeted at woman detection (Tamil & Malayalam) and Fake News Detection (Malayalam) are made available as a collection in Hugging Face[2].

Task	F1-Score	Rank	v/s Best
ai-gen-review	0.6745	24	-0.2955
abusive-woman	0.7821	5	**-0.0062**

Table 3: Evaluation Metrics for Tamil Tasks.

[1]This only includes the transformer parameters and does not include the embedding parameters

[2]https://huggingface.co/collections/livinNector/multilinugal-minilm-dravidianlangtech-679b3d894e207e2844c4d637

Task	F1-Score	Rank	v/s Best
ai-gen-review	0.8997	6	**-0.0202**
abusive-woman	0.6915	7	**-0.0656**
fake-news	0.8320	11	**-0.0660**

Table 4: Evaluation Metrics for Malayalam Tasks.

5 Investigating the results of AI-Generated Review detection in Tamil

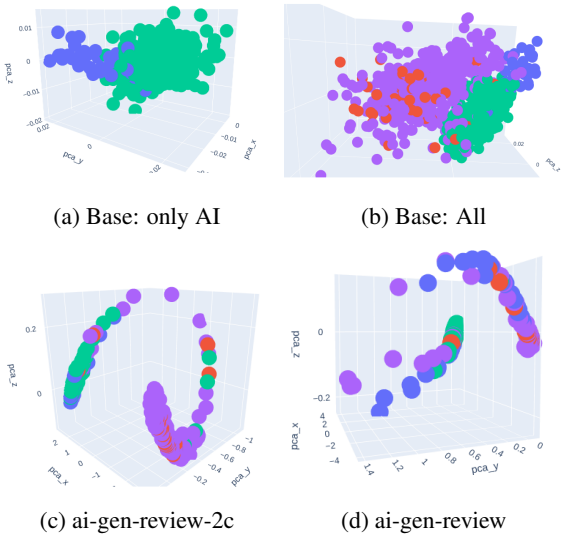
After the declaration of the result of the shared tasks, the reason for the significant variation in the performance of the **ai-gen review** model in the Tamil language test data is explored. Even though the model achieved an F1 score of 0.9816 in the development set during the initial fine-tuning, it had a lower F1 score in the test set. Three more fine-tuning runs are conducted with the same dataset, despite the F1 score with the development set (Tamil-Malayalam) being consistent in the range of 0.97-0.98, the F1 score in the test set varied significantly for the Tamil language, although the development set had a consistently high F1 score (Table 5). These different fine-tunings of the Multilingual MiniLM are available in Hugging Face.

Model	Tam-Test	Mal-Test
ai-gen-review	**0.6745**	0.8997
ai-gen-review-2a	0.8996	0.9147
ai-gen-review-2b	0.9095	0.8942
ai-gen-review-2c	**0.9800**	0.8749

Table 5: Macro F1-score evaluated on the Tamil and Malayalam test sets individually.

To study this in detail, a 3-dimensional visualization of the embedding space of the Multilingual MiniLM model is created using PCA transformation of the pooling layer outputs to study the embeddings of reviews in train and test sets. The visualizations show that the embeddings of AI-generated text in both train and test sets varied significantly. Also, visualizing the embeddings of the **ai-gen-review** and **ai-gen-review-2c** suggests that the model is capable of achieving a higher performance. The visualization of the embeddings is presented in Figure 1.

These explorations make it clear that the train and test set vary significantly in the embedding space of AI-generated reviews. This makes the training set insufficient to capture the entire region of AI-generated reviews. To overcome this issue in the data, more diverse AI-generated reviews from different AI models can be added to the training set so that the training and evaluation objectives of AI-generated reviews align closer and can be captured by the model.



(a) Base: only AI (b) Base: All

(c) ai-gen-review-2c (d) ai-gen-review

Figure 1: 3d visualization of PCA transformation of pooling layer outputs of the base and fine-tuned MiniLM models. green=AI-Train, blue=AI-Test, violet=HUMAN-Train, red=HUMAN-Test.

6 Conclusion

The results indicate that the performance of Multilingual MiniLM on the downstream tasks AI-generated review detection, abusive text detection, and fake news detection is comparable to the other models while having significantly fewer parameters than the other BERT-based models. The misalignment in the train and test sets of the Tamil AI-generated review data set is identified. The results of the fine-tuned models are also justified using visualization of the output layer.

7 Limitations

This study has the following limitations.

- The AI-generated review detection task exhibited significant variability in model performance, particularly for Tamil, suggesting limitations in the training dataset's representativeness.
- Though this study compares the performance of this model to the best performance reported by others from the shared task, it doesn't compare it with the performance of a model with a

similar architecture and more parameters like XLM-Roberta.

- Explainability and interpretability of the model's performance are not analyzed in detail. A detailed study on the intermediate attentions of the model might give hints on the tokens that contribute more to the results.
- The effects of different data augmentation and regularization techniques on the performance of the model have not been explored in this work.

References

K Devika, B Haripriya, E Vigneshwar, B Premjith, Bharathi Raja Chakravarthi, et al. 2024. From dataset to detection: A comprehensive approach to combating malayalam fake news. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 16–23.

Asha Hegde, Kavya G, Sharal Coelho, and Hosahalli Lakshmaiah Shashirekha. 2023. MUCS@DravidianLangTech2023: Leveraging learning models to identify abusive comments in code-mixed Dravidian languages. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 266–274, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*.

Zhipeng Luo and Jiahui Wang. 2023. DeepBlueAI@DravidianLangTech-RANLP 2023. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 171–175, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

B Premjith, Nandhini K, Bharathi Raja Chakravarthi, Thenmozhi Durairaj, Balasubramanian Palani, and Kumaresan Prasanna Kumar Thavareesan, Sajeetha. 2025. Overview of the Shared Task on Detecting AI Generated Product Reviews in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

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. 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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Malliga Subramanian, , B Premjith, Kogilavani Shanmugavadivel, Santhia Pandiyan, Balasubramanian Palani, and Bharathi Raja Chakravarthi. 2025. Overview of the Shared Task on Fake News Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, B Premjith, K Vanaja, S Mithunja, K Devika, et al. 2024. Overview of the second shared task on fake news detection in dravidian languages: Dravidianlangtech@ eacl 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 71–78.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, Muskaan Singh, Sandhiya Raja, Vanaja, and Mithunajha S. 2023. Overview of the shared task on fake news detection from social media text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. MobileBERT: a compact task-agnostic BERT for resource-limited devices. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2158–2170, Online. Association for Computational Linguistics.

Nafisa Tabassum, Sumaiya Aodhora, Rowshon Akter, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. Punny_Punctuators@DravidianLangTech-EACL2024: Transformer-based approach for

detection and classification of fake news in Malayalam social media text. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 180–186, St. Julian's, Malta. Association for Computational Linguistics.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.