Misogynistic Meme Detection in Dravidian Languages Using Kolmogorov Arnold-based Networks

Manasha Arunachalam[1], Navneet Krishna Chukka[1], Harish Vijay V[1]
Premjith B[1], Bharathi Raja Chakravarthi[2]
[1]Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India,
[2]School of Computer Science, University of Galway, Ireland,
manasha.arun@gmail.com, navneetkrishna.918@gmail.com
harishvijay0204@gmail.com, b_premjith@cb.amrita.edu
bharathi.raja@universityofgalway.ie

Abstract

The prevalence of misogynistic content online poses significant challenges to ensuring a safe and inclusive digital space for women. This study presents a pipeline to classify online memes as misogynistic or non misogynistic. The pipeline combines contextual image embeddings generated using the Vision Transformer Encoder (ViTE) model with text embeddings extracted from the memes using ModernBERT. These multimodal embeddings were fused and trained using three advanced types of Kolmogorov Artificial Networks (KAN): PyKAN, FastKAN, and Chebyshev KAN. The models were evaluated based on their F1 scores, demonstrating their effectiveness in addressing this issue. This research marks an important step towards reducing offensive online content, promoting safer and more respectful interactions in the digital world.

1 Introduction

Recent studies have highlighted the role of social media algorithms in amplifying such harmful content, thereby normalizing detrimental ideologies among users. Addressing this issue necessitates effective detection and mitigation strategies. In this study, we propose a comprehensive pipeline for classifying online memes as either containing misogynistic content or not. This approach integrates multimodal data by combining contextual image embeddings from the Vision Transformer Encoder (ViTE) model with text embeddings derived from ModernBERT. The fused embeddings are then processed through advanced Kolmogorov-Arnold Networks (KAN), specifically PyKAN, FastKAN, and Chebyshev KAN, to enhance classification accuracy. The efficacy of these models is evaluated using F1 scores, demonstrating their potential in identifying and mitigating offensive online content. This research helps create safer and more inclusive digital environments by providing a robust method for detecting misogynistic material.

Kolmogorov-Arnold Network (KAN): A modern neural network architecture built on the Kolmogorov-Arnold representation theorem, which asserts that any continuous multivariate function can be decomposed into a combination of single-variable functions. Using a two-layer structure, KAN approximates the target function by combining input mappings to a higher-dimensional space in the inner layer. In KAN, the activation function consists of a spline function, parameterized as a linear combination of B-splines, a base function, often referred to as the SiLU (Sigmoid Linear Unit). KAN is theoretically robust, but because of the complexity of the spline function, it can be computationally difficult to train.

Chebyshev KAN :The Kolmogorov-Arnold theorem is expanded upon by Chebyshev KAN, which uses Chebyshev polynomials for function approximation. Using a single-layer Chebyshev interpolation method, KAN models the target function through a weighted sum of Chebyshev polynomials, with the input normalized by a hyperbolic tangent (tanh) function. While following the theoretical underpinnings of the Kolmogorov-Arnold theorem, ChebyKAN takes advantage of the superior approximation properties of Chebyshev polynomials. This method seeks to improve function approximation's accuracy and efficiency in comparison to the original KAN.

FastKAN: In order to overcome the computational difficulties of KAN, FastKAN substitutes B-spline basis with Gaussian Radial Basis Functions (RBFs), greatly lowering the computational cost. B-splines can be efficiently approximated by Gaussian RBFs, which removes the computational bottleneck in the original KAN implementation. Because of this modification, FastKAN is more feasible for real-time and large-scale applications while preserving the approximation capabilities of KAN and increasing computing efficiency.

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 37–44

2 Literature Review

Previous research (Ponnusamy et al., 2024) has focused on detecting misogyny and gender bias in online spaces, with datasets in languages like English and Spanish. However, there is a lack of studies focusing on regional languages, particularly Tamil and Malayalam, where cultural context plays a significant role. Existing tools often don't address these specific issues effectively. The MDMD dataset fills this gap by providing a resource for detecting misogyny in Tamil and Malayalam memes, helping researchers understand and address gender bias in these communities more accurately.

In recent years, detecting toxic and abusive comments on social media has become crucial to maintaining a safe online environment. Several models have been developed to identify hate speech, toxicity, and bullying, primarily in high-resource languages like English. However, there is limited research on detecting such content in low-resource languages, such as Tamil. Previous work has highlighted the challenges of language-specific nuances, especially when it comes to understanding cultural contexts. This paper (Bhattacharyya, 2022) contributes to the gap by focusing on Tamil, approaching the problem of abusive comment detection as a multi-class classification task. The study compares various pre-processing and modeling techniques, evaluating their effectiveness based on weighted average accuracy.

Recent research (Shaun et al., 2024) has explored the classification of Tamil and Malayalam memes as misogynistic or non-misogynistic. One approach involved separately analyzing textual content using Multinomial Naive Bayes and visual content using the ResNet50 model. By combining the results from both modalities, researchers achieved significant success in identifying misogynistic content in memes. This work underscores the importance of multi-modal analysis in detecting harmful content, especially in low-resource languages.

Detecting misogynistic memes is challenging due to the complex interaction between image and text, where these elements often convey different meanings.Prior research (Jindal et al., 2024) has focused on individual modalities, such as text or image analysis, but these approaches overlook the need for multimodal fusion. Recent works have started exploring fusion techniques, utilizing models like Vision Transformer for images and transformer-based models like DistilBERT for text. These approaches have shown promise in improving the detection of harmful content. However, there remains a gap in combining these modalities effectively, especially in detecting misogyny. The MISTRA framework addresses this by using variational autoencoders for dimensionality reduction and large language models for fusion embeddings, enhancing classification performance on multimodal data.

Additional Studies (Sharma et al., 2024) have explored the use of deep learning models, such as recurrent neural networks (RNN), long-short term memory (LSTM), and bidirectional LSTM, for detecting various categories of hate speech, including misogyny, misandry, and xenophobia. These models are applied to Tamil and Tamil-English code-mixed comments, and results are analyzed to evaluate their effectiveness in identifying abusive content.

Social media memes, combining text and images, can sometimes contain harmful content like misogyny, affecting users' well-being. Detecting such content, especially in low-resource languages, is challenging due to the lack of suitable datasets. This work (Singh et al., 2024) introduces a Hindi-English code mixed meme dataset of 5,054 annotated memes for two tasks: misogyny detection and multi-label classification. Results show that multimodal fusion models outperform text-only and image-only models in identifying misogyny. This dataset provides a valuable resource for advancing research in detecting harmful online content.

Misogynistic memes, which target women with disrespectful language, pose a challenge to maintaining a healthy online environment. The paper (Mahesh et al., 2024) presents three models: BERT+ResNet-50, MuRIL+ResNet-50, and mBERT+ResNet-50, which combine text and image representations for meme classification. The mBERT+ResNet-50 and MuRIL+ResNet-50 models achieved impressive macro F1 scores of 0.73 and 0.87 for Tamil and Malayalam datasets, securing 1st place for both languages in the shared task.

A system for identifying abusive remarks in Tamil and Tamil-English is shown in the work (Duraphe et al., 2022) utilizing three different approaches: transformer-based modeling, deep learning, and machine learning. Classifying remarks into groups such as misogyny, misandry, homophobia, and others was their goal. For the Tamil+English dataset, the system performs best

when employing Random Forest, with a weighted average F1-score of 0.78. Furthermore, mBERT produces the best result for Tamil with an F1-score of 0.7 in Transformer-based modeling, whereas Bi-Directional LSTM performs better for Deep Learning.

A unique method for identifying inappropriate language on social media in multilingual, code-mixed, and script-mixed contexts is presented in this study (Saumya et al., 2024). The challenge makes use of a hybrid multilingual dataset that was produced by fusing bilingual and monolingual materials. The study assesses the effects of deep learning models (CNN, Bi-LSTM, Bi-LSTM-Attention, and fine-tuned BERT) and various input representations (Word2Vec, GloVe, BERT, and uniform initialization). With a macro average F1-score of 0.79 for monolingual tasks and 0.86 for code-mixed/script-mixed tasks, the results show how well fine-tuned BERT performs, improving the identification of abusive language in a variety of multilingual contexts.

3 Dataset Description

The Misogyny Meme Detection dataset consists of images, corresponding transcriptions, and labels (Ponnusamy et al., 2024).

For Tamil, the training data includes 1,135 images and a CSV file containing image IDs, transcriptions, and labels (0 or 1). Out of these, 732 images have matching entries in the CSV file based on image IDs. The development (dev) set contains approximately 282 images and a similar CSV file with image IDs, transcriptions, and labels . Among these, 252 images have matching entries in the CSV file based on image IDs. The test set comprises approximately 356 images, along with a CSV file containing transcriptions and labels. Among these, 82 images have corresponding entries in the CSV file based on their image IDs. The images and their corresponding transcriptions were used to predict the labels, and the predicted labels were evaluated against the true labels provided in the CSV file to measure accuracy.

In the Tamil dataset, some images listed in the test set were missing from the provided test data. To address this, the missing images were searched for in the training and development sets. Once identified, these images were added to the test set. It was confirmed that these images had not been used during training or validation, ensuring the test set remained unique and independent. A similar approach was applied to the training and development sets. For the training set, missing images were identified by cross-checking the data entries and were searched for in the development set. Likewise, for the development set, any missing images were located in the training set. This ensured that all data was appropriately assigned while maintaining the uniqueness and integrity of each dataset.

The Malayalam Misogyny Meme Detection dataset includes 640 images in the training set, accompanied by a CSV file containing image IDs, transcriptions, and labels (0 or 1). The development (dev) set consists of 160 images and a corresponding CSV file with the same structure. The test set contains 200 images and a CSV file with transcriptions and labels. Similar to Tamil, the images and transcriptions in the test set were used to predict the labels, and the accuracy of these predictions was evaluated by comparing them with the true labels from the CSV file.

4 Methodology

As per Figure(1), For the Tamil and Malayalam dataset, we used training data that contained images and their corresponding transcriptions, along with binary labels (0 or 1).

For each image in the dataset, we used the Vision Transformer Encoder (ViTE) model to extract high-dimensional embeddings. This allowed us to represent visual data effectively for downstream tasks.

The transcriptions associated with each image were processed using ModernBERT, which generated text embeddings representing the semantic content of the text. The image and text embeddings were fused to create a single representation for each data point. This fused embedding served as the input for the classification models. We trained three types of Kolmogorov Artificial Networks (KANs) using the fused embeddings: PyKAN, Chebyshev KAN and FastKAN and fine tuned using the hyperparameters listed in Table 1.

We utilized ModernBERT to generate text embeddings. ModernBERT is an advanced language model that enhances the original BERT architecture by extending the context length to 8,192 tokens, allowing it to process longer documents effectively. It incorporates Rotary Positional Embeddings (RoPE) for improved token position understanding and replaces traditional MLP layers

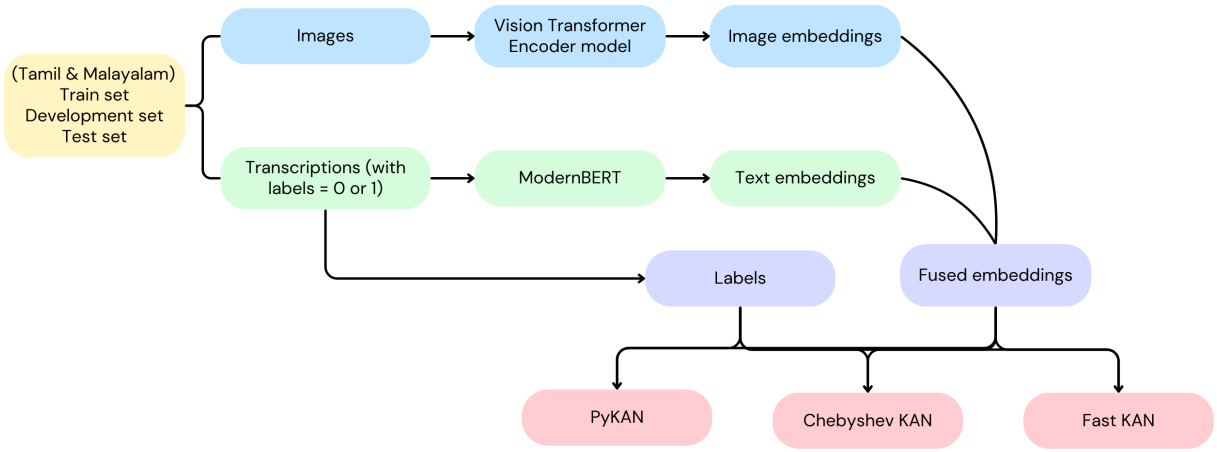

Figure 1: Overall Workflow

with GeGLU layers, enhancing model performance. These architectural improvements allow ModernBERT to produce more comprehensive and contextually rich embeddings, which is crucial for our classification task.

To generate image embeddings, we employed the Vision Transformer Encoder (Remya et al., 2024) (ViTE) model. ViTE operates by dividing an input image into fixed-size patches, each of which is linearly transformed into a vector representation. These patch embeddings are then combined with positional embeddings to retain spatial information. The resulting sequence is processed through transformer encoder layers, enabling the model to capture both local and global features of the image. This method allows ViTE to generate comprehensive embeddings that effectively represent the visual content.

5 Experiments and Discussion

5.1 Experimental Setup

The experiments were carried out on a MacBook (M4 Pro) equipped with 24GB of Unified Memory and a 512GB SSD. This setup, combined with PyTorch's support for macOS using the Metal Performance Shaders (MPS) backend, allowed for smooth model loading and faster training by efficiently utilizing the Mac hardware capabilities.

5.2 Hyperparameters

The hyperparameters used for training the KAN models have been discussed below in table 1.

Model	Learning Rate	Epochs	Batch Size	Optimiser
Chabyshev KAN (Tamil)	0.0001	20	32	Adam
PyKAN (Tamil)	0.001	20	32	Adam
FastKAN (Tamil)	0.001	55	32	Adam
Chabyshev KAN (Malayalam)	0.0001	20	32	Adam
PyKAN (Malayalam)	0.001	20	32	Adam
FastKAN (Malayalam)	0.001	55	32	Adam

Table 1: Hyperparameters and Optimisers for Different Models

5.3 Software packages

We used PyTorch and TensorFlow for model training and related tasks, while scikit-learn was employed for evaluation metrics. PyTorch and TensorFlow are prominent deep learning frameworks that facilitate the development and training of neural networks. Scikit-learn, on the other hand, is a widely-used machine learning library that provides tools for data analysis and model evaluation.

6 Results

The F1 Score—the harmonic mean of precision and recall—was the main metric we used to assess our models. Because it ensures that both false positives and false negatives are taken into account, it is especially helpful in situations where the dataset is unbalanced. In our case, we found that class 1, the misogynistic class, had less incidents. The following formula provides the F1 score:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (1)$$

The ratio of accurately predicted positive observations to all actual positives is called recall, while the ratio of properly predicted positive observations to all predicted positives is called precision.

The F1 Score is preferred in our evaluation for the following reasons: It balances precision and recall, making it suitable for imbalanced datasets where accuracy alone may be misleading. It ensures that both false positives and false negatives are accounted for, which is crucial for our classification task. Unlike accuracy, F1 Score does not get skewed when one class is significantly larger than the other. Thus, using the F1 score provides a more reliable measure of the model's effectiveness in real-world scenarios like these.

6.1 Results for Tamil

As observed in Table 1,2,3, the highest F1-score of 0.77 for ChebysevKAN, followed by 0.76 for PyKAN and 0.73 for FastKAN.

ChebysevKAN

Class	Precision	Recall	F1-Score	Support
0.0	0.88	0.90	0.89	267
1.0	0.67	0.63	0.65	89
Accuracy		0.83		356
Macro Avg	0.77	0.76	**0.77**	356
Weighted Avg	0.83	0.83	0.83	356

Table 2: Test Classification Report for ChebyshevKAN on Tamil dataset

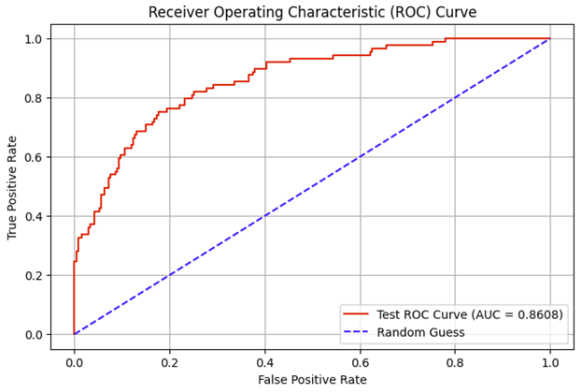

Figure 2: ROC curve for ChebysevKAN on Tamil dataset

PyKAN

Class	Precision	Recall	F1-Score	Support
0.0	0.88	0.87	0.88	267
1.0	0.63	0.65	0.64	89
Accuracy		0.82		356
Macro Avg	0.76	0.76	**0.76**	356
Weighted Avg	0.82	0.82	0.82	356

Table 3: Test Classification Report for PyKAN on Tamil dataset

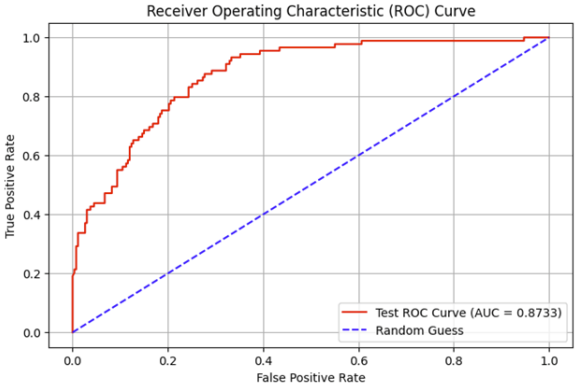

Figure 3: ROC cureve for PyKAN on Tamil dataset

FastKAN

Class	Precision	Recall	F1-Score	Support
0.0	0.84	0.94	0.89	267
1.0	0.74	0.47	0.58	89
Accuracy		0.83		356
Macro Avg	0.79	0.71	**0.73**	356
Weighted Avg	0.82	0.83	0.81	356

Table 4: Test Classification Report for FastKAN on Tamil dataset

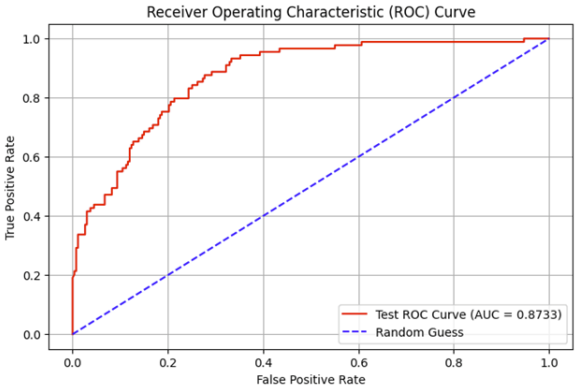

Figure 4: ROC curve for FastKAN on Tamil dataset

All three models achieve similar accuracy (around 0.83) and perform well on class 0, with high precision and recall (above 0.84). Each model demonstrates strengths and weaknesses depending on the classification task.

As observed from Figure (2,3,4) Chebyshev KAN offers the best balance, PyKAN is a close alternative, and Fast KAN performs well but needs recall improvements for class 1. KAN models misclassifies a higher number of class 1 samples, likely due to feature overlaps or data imbalance. Further tuning and data adjustments can enhance overall performance, particularly for class 1 detection.

6.2 Results for Malayalam

The same model architecture was applied on Malayalam dataset. The results for each model (FastKAN, ChebysevKAN, and PyKAN) are presented below, along with their respective classification reports and ROC curves.

All three models (FastKAN, ChebysevKAN, and PyKAN) from Table 5,6,7, The models demonstrated strong performance on the Malayalam dataset, with FastKAN achieving the highest accuracy (0.88) and F1-score (0.87). ChebysevKAN and PyKAN followed closely, with accuracies of 0.87 and 0.86, respectively. The ROC curves for all models as observed from Figure (5,6,7) indicate excellent discrimination capabilities, with high AUC values.

Additionally, the classification reports reveal that FastKAN consistently achieved higher recall for class 1, making it particularly effective in identifying positive instances. ChebyshevKAN demonstrated a more balanced performance across both classes, while PyKAN exhibited slightly lower recall but maintained competitive precision. The ROC curves further validate the model's classification capabilities, with all AUC values exceeding 0.85.

These findings suggest that the proposed architecture not only generalizes well across different languages but also adapts effectively to the nuances of the Malayalam dataset, with FastKAN being the most effective among the three.

FastKAN

Class	Precision	Recall	F1-Score	Support
0.0	0.88	0.93	0.90	122
1.0	0.87	0.79	0.83	78
Accuracy		0.88		200
Macro Avg	0.87	0.86	**0.87**	200
Weighted Avg	0.87	0.88	0.87	200

Table 5: Test Classification Report for FastKAN on Malayalam dataset

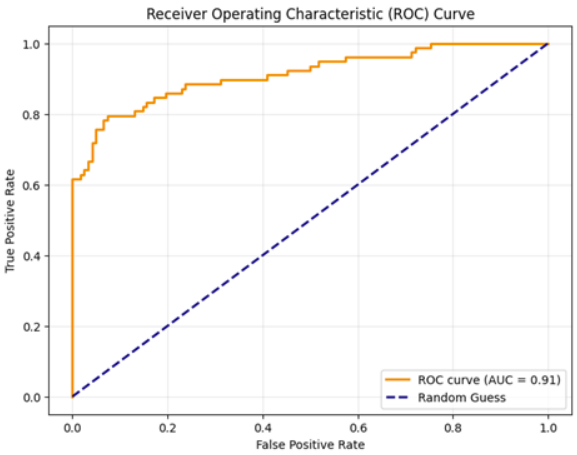

Figure 5: ROC curve for FastKAN on Malayalam data

ChebysevKAN

Class	Precision	Recall	F1-Score	Support
0.0	0.87	0.93	0.90	122
1.0	0.87	0.78	0.82	78
Accuracy		0.87		200
Macro Avg	0.87	0.85	**0.86**	200
Weighted Avg	0.87	0.87	0.87	200

Table 6: Test Classification Report for ChebysevKAN on Malayalam dataset

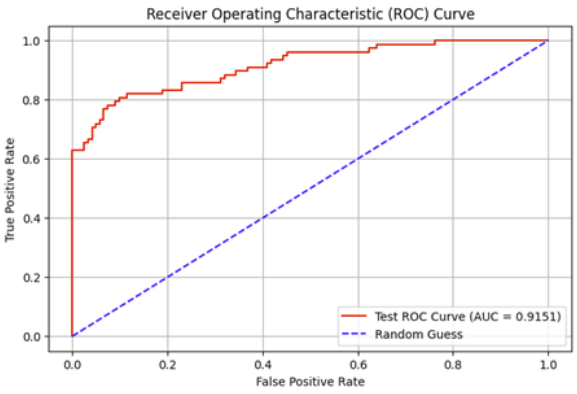

Figure 6: ROC curve for ChebysevKAN on Malayalam dataset

PyKAN

Class	Precision	Recall	F1-Score	Support
0.0	0.88	0.89	0.89	122
1.0	0.83	0.81	0.82	78
Accuracy		0.86		200
Macro Avg	0.85	0.85	**0.85**	200
Weighted Avg	0.86	0.86	0.86	200

Table 7: Test Classification Report for PyKAN on Malayalam dataset

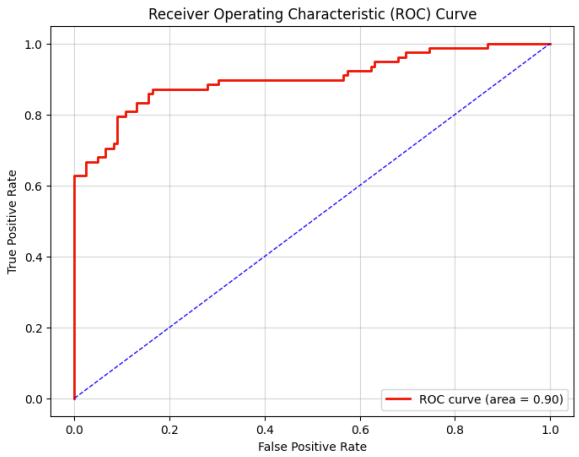

Figure 7: ROC curve for PyKAN on Malayalam dataset

To mitigate the class imbalance in our dataset, where misogynistic instances were significantly underrepresented compared to non-misogynistic ones, we used various oversampling techniques to generate a more balanced distribution. Specifically, we applied SMOTE (Synthetic Minority Oversampling Technique), KMeansSMOTE, ADASYN (Adaptive Synthetic Sampling), and BorderlineSMOTE, each of which synthesizes new samples for the minority class using different interpolation strategies. These methods allowed us to expand the misogynistic class while preserving the overall data distribution, ensuring that our KAN models had sufficient representative samples to learn nuanced patterns associated with misogynistic content.

After incorporating the oversampled data into our training pipeline, we retrained our KAN models and observed that the overall performance remained comparable to our initial results. However, the class balance improved significantly, leading to an increase in key metrics for the misogynistic class (Class 1). This indicates that the model's ability to correctly identify misogynistic content improved without introducing substantial biases toward the majority class. The results highlight the effectiveness of oversampling in handling class imbalances and ensuring better model performance across both classes.

7 Inference

From the classification reports, we observe that all models achieve high precision and recall for class 0 (non-misogynistic text) but exhibit lower recall for class 1 (misogynistic text), indicating that they struggle to correctly identify some misogynistic instances.

Among the models, ChebyshevKAN performed best relative to F1 score on Tamil dataset and FastKAN on Malayalam dataset.

We could enhance recall through advanced feature engineering, leveraging larger and more diverse datasets, or incorporating multimodal approaches to improve robustness and fairness in misogyny detection.

8 Conclusion

This study shows that combining image and text features using Vision Transformer Encoder (ViTE) and ModernBERT with Kolmogorov Arnold Networks (KAN) is effective for detecting misogynistic memes in Tamil and Malayalam. Among the models tested, FastKAN performed best for Malayalam (F1 score: 87), while Chebyshev KAN was the most effective for Tamil (F1 score: 77). Despite the class imbalance—where misogynistic content was underrepresented in the dataset—the models achieved reasonable scores, demonstrating their robustness. However, further fine-tuning is needed to address the challenges posed by this imbalance.

Future work can focus on expanding the dataset to more Indian languages, improving fusion techniques for better accuracy, and optimizing models for real-time deployment in social media moderation tools. Additionally, incorporating techniques to handle class imbalance and integrating bias detection and explainability methods can make the system more transparent and fair. These advancements will help in automating content moderation, reducing harmful content, and promoting inclusive online communities.

References

Aanisha Bhattacharyya. 2022. Aanisha@ tamilnlp-acl2022: abusive detection in tamil. In *Proceedings*

of the Second Workshop on Speech and Language Technologies for Dravidian Languages, pages 214–220.

Ankita Duraphe, Ratnavel Rajalakshmi, and Antonette Shibani. 2022. Dlrg@ dravidianlangtech-acl2022: Abusive comment detection in tamil using multilingual transformer models. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics (ACL).

Nitesh Jindal, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Sajeetha Thavareesan, Saranya Rajiakodi, and Bharathi Raja Chakravarthi. 2024. Mistra: Misogyny detection through text–image fusion and representation analysis. *Natural Language Processing Journal*, 7:100073.

Sidharth Mahesh, D Sonith, Gauthamraj Gauthamraj, G Kavya, Asha Hegde, and H Shashirekha. 2024. Mucs@ lt-edi-2024: Exploring joint representation for memes classification. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 282–287.

Rahul Ponnusamy, Kathiravan Pannerselvam, R Saranya, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, S Bhuvaneswari, Anshid Ka, Susminu S Kumar, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. From laughter to inequality: Annotated dataset for misogyny detection in tamil and malayalam memes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7480–7488.

S Remya, T Anjali, S Abhishek, Somula Ramasubbareddy, and Yongyun Cho. 2024. The power of vision transformers and acoustic sensors for cotton pest detection. *IEEE Open Journal of the Computer Society*.

Sunil Saumya, Abhinav Kumar, and Jyoti Prakash Singh. 2024. Filtering offensive language from multilingual social media contents: A deep learning approach. *Engineering Applications of Artificial Intelligence*, 133:108159.

Deepawali Sharma, Vedika Gupta, and Vivek Kumar Singh. 2024. Abusive comment detection in tamil using deep learning. In *Computational Intelligence Methods for Sentiment Analysis in Natural Language Processing Applications*, pages 207–226. Elsevier.

H Shaun, Samyuktaa Sivakumar, R Rohan, Nikilesh Jayaguptha, and Durairaj Thenmozhi. 2024. Quartet@ lt-edi 2024: A svm-resnet50 approach for multitask meme classification-unraveling misogynistic and trolls in online memes. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 221–226.

Aakash Singh, Deepawali Sharma, and Vivek Kumar Singh. 2024. Mimic: Misogyny identification in multimodal internet content in hindi-english code-mixed language. *ACM Transactions on Asian and Low-Resource Language Information Processing*.