"An Intelligent Electronic System for Offensive Meme Detection using BERT and VGG-19"

Sindhu Ravindran
School of computer science and engineering,
Vellore Institute of Technology, Chennai,India
sindhu.ravindran@vit.ac.in

Ashwini K
Department of Computer Science and Engineering,
Amrita School of Computing,
Amrita Vishwa Vidyapeetham, Chennai, India
k_ashwini@ch.amrita.edu

V Bharathi
Department of Computer Science and Engineering,
Vels Institute of science, Technology & Advanced Studies (VISTAS), Pallavaram, Chennai
vbharathi.se@velsuniv.ac.in

S Divya Bairavi
Department of Computer Science and Engineering,
Vels Institute of science, Technology & Advanced Studies (VISTAS), Pallavaram, Chennai
divyabairavi.se@vistas.ac.in

Vikneswaran Vijean,
Faculty of Electronic Engineering & Technology,
Universiti Malaysia Perlis, Kampus Alam UniMAP, Pauh Putra, 02600 Arau,Perlis, Malaysia
vikneswaran@unimap.edu.my

Praveen Balaji* (Corresponding author)
Department of Computer Science and Engineering,
Amrita School of Computing,
Amrita Vishwa Vidyapeetham, Chennai, India
ch.sc.u4cse23040@ch.students.amrita.edu*

***Abstract*—With user-generated content continuing to proliferate on social media, the application of offensive content in memes has created a critical problem that requires sophisticated means of Classification. This paper designs a novel framework that involves deep learning models for image and text classification. Specifically, the system utilizes Optical Character Recognition (OCR) technology to extract text from images, and afterward, applies the BERT (Bidirectional Encoder Representations from Transformers) model to analyze text and the VGG-19 convolutional neural network to analyze images. The framework can provide effective integration of various data types and thus enable the identification of offensive memes quickly and accurately. Experimental analysis on benchmarking datasets was used to measure the success of this approach for meme classification and its performance and stability. Multi-modal offensive content moderation was the identified issue, and an issue to the wider field of automated method to content study and enhanced user protection has been provided.**

Keywords—Electronic based Offensive Meme Detection, Optical Character Recognition (OCR), BERT, VGG-19, Automated Content Analysis.

I. INTRODUCTION

In today's world of social media, Memes have become increasingly prevalent on social media in today's world. Mostly, this is because of the ability to share user-generated content, which combines text and images. Although many memes are created for fun, a good number of them can be harmful or outright offensive. And with the combination of visual and textual components, memes pose a unique problem for automated content moderation systems. In general, these systems do not work well with user- generated content as they rely solely on text or image analysis. As a result, a more advanced multi-modal detection system is required. Memes are single frames, the challenge comes when incorporating so many visual and textual cues within it, some may range from harmless to extremely provocative. For example, an image that may be viewed as neutral can produce an offensive message when paired with a controversial caption. However, controversial images may have benign text overlaid on them to produce offensive memes. Due to these superimposed structures upon the various elements, a single photo may be free of offending context. To properly respond to such a subtlety, a more advanced analysis is needed in which all the elements of a photo are aggregated.

The advent of deep learning methods has substantially enhanced the evaluation of offensive content. Algorithms such as BERT (Bidirectional Encoder Representations from Transformers) and VGG19 Convolutional Neural Network revolutionized natural language comprehension and image classification, respectively. While BERT was able to capture context in subtle textual phrases, VGG19 achieved image categorization. Moreover, the combination of these models shows a very promising approach to addressing meme offensiveness detection. Optical Character Recognition provides a method of extracting the text within the images so that it can be included with the visual features of the meme for more comprehensive assessment. This paper proposes a new approach to identify offensive memes by putting into application the collaboration of several technologies: OCR for extracting texts, BERT for text classification, and VGG19 for image classification. This software identifies and analyses the text and images of the meme at the same time to identify its offensiveness and context more effectively. Thus, the system tries to address issues present in unimodal detection methods and the complexity involved in multimodal data detection.

The key contribution of this work bolsters the field of offensive meme detection. Lying at the heart of this work is a comprehensive multimodal analysis framework that can be applied in broader fields of automated content moderation and online safety. Moreover, this analysis highlights the imperative of an interdisciplinary strategy, showing how natural language processing, computer vision, and machine learning can work

synergistically to address today's challenges in the digital space. The application of the suggested model has been to the memes collection databases that have been developed to probe the ability of detecting some kinds of offensive images. The results also further show that the combination of the text and image modes of input provides higher performance for the classifier than a single mode.

II. Related Works

In recent years, the use of methods for categorizing offensive memes has become popular. Different researchers have worked on different deep learning methods to improve the detection of hate speech in text and images. For example, Mohammed Abulais et al. proposed a detection system for figurative language with an emphasis on the identification of metaphor, sarcasm, and irony. Supervised learning, Deep learning, and Ensemble learning approaches were taken and yielded a very high AUC of 0.99. Nevertheless, the model was able to solve low resource language problems, but it failed in the opposite case [1]. Kiela et al. have developed the Hateful Memes Challenge and the accompanying dataset which consists of more than 10,000 multimodal units. The results of the study also showed that the integration of the two modalities led to better classification results but pointed out the insufficient methods of feature fusion [2].

MultiOFF dataset for the multimodal offensive content detection system was designed by Suryawanshi et al. identified how both modalities synergistically contribute towards the enhancement of classification accuracy of offensive content. Additionally, this research brings out the importance of ensuring there is enough diversity in the datasets as a support to model robustness in real-world contexts [3]. HateXplain, a dataset that serves as a benchmark aimed at enhancing the performance of hate speech detection models through explainable AI features was released by Binny Mathew et al. Several models, including CNN-GRU, BiRNN, and BERT, were tested, achieving the best performance with HateXplain BERT which had an accuracy of 0.698. This research reported issues with mitigating bias that were unintentionally used in the dataset [4].

Shang et al. devised an analogy-knowledgeable offensive meme detection method. This method brought into focus the requirement of context comprehension using learning-with-analysis, obtaining benchmark-level accuracy using accessible datasets but incurring monstrous computational expenses [5]. Chhabra et al. reviewed the literature on the identification of multi-modal multilingual hate speech and provided a critique of the research, with respect to dataset diversity and model generalizability .On top of that, they call for greater attention to underrepresented languages and cultures to ensure equitable hate speech detection systems [6].

Briskilal et al. founded on CNN and LSTM architectures that created high accuracy but misclassifications involving ambiguous instances. This study yielded context-sensitive models as a requirement for good content understanding [7]. In 2024, Abdullakutty F. et al, an advanced multimodal zero-shot classification approach is introduced for offensive meme detection, leveraging large language models (LLMs) and BLIP-based image captioning [8]. In 2025, El-amrany, S et al have proposed GuardHarMem, a large-scale multimodal dataset for harmful meme analysis, is proposed together with HarMDetect, a classification architecture that exploits its rich annotation structure [9].

In 2025, an innovative multimodal framework, DecodEM-X, is introduced to strengthen harmful meme detection. The study indicates its capability to advance AI-powered moderation systems in terms of robustness and ethical adherence [10]. The study addresses offensive content detection related to women's harassment by incorporating sentiment and emotion intensity analysis. A large-scale WCSEoff dataset is created, and a multitask model is developed to detect offensiveness, its severity, and associated emotional context. The proposed framework improves performance by over 7% compared to conventional methods without sentiment–emotion integration [11]. The study proposes an AI-driven IoT-fog framework to detect hate speech and symbols in educational institutions using YOLOv8 for object detection, EasyOCR for text extraction, and a modified DistilBERT for contextual analysis. The system achieves high performance (accuracy 0.94, F1-score 0.87) and enables real-time identification of harmful content. This approach supports safer and more inclusive campus environments by improving automated moderation and alert mechanisms [12].

Although recent studies have advanced offensive meme detection using multimodal datasets, deep learning models, and even LLM-based zero-shot approaches, several limitations remain. Existing methods often rely on weak feature fusion, struggle with contextual or ambiguous content such as sarcasm and symbolism, and suffer from dataset bias and limited cultural diversity, leading to poor generalization in real-world scenarios. Some context-aware and large models improve performance but introduce high computational cost and low interpretability. Therefore, there is still a need for an efficient, scalable multimodal framework that enables deeper cross-modal understanding, reduces bias, and provides reliable offensive content detection across diverse environments. This paper introduces a BERT–VGG19 architecture by combining contextual textual understanding with robust visual feature extraction and provides a reliable and scalable solution for offensive meme classification across diverse contexts.

III. Methodology

This section details the techniques used in the selection of datasets, the processes of transforming data, and the creation of the model that classifies memes. The research incorporates the Multimodal Meme Dataset (MultiOFF) [3], along with extra memes gathered from sources including Kaggle, Twitter, Instagram, and Facebook. MultiOFF is a prominent dataset tailored for offensive meme detection, offering a rich array of examples that facilitate the effective training and evaluation of our model. The combined dataset encompassed a wide variety of memes, which increases the model's strength. The emphasis is on experiments performed using different modes of approaches on a common dataset for the careful evaluation of meme related aggressiveness.

A. Dataset Preprocessing

Data preprocessing is an important phase that impacts the quality and integrity of the data to be supplied to the model. The dataset distributed among three main folders: training, validation, and testing. Each folder has subfolders that were named according to the respective classes (for instance, “Offensive” and “Non-Offensive”). The images were adjusted to the VGG19 model standard by cropping them to pixel 224x224.

Data augmentation by rescaling the pixel value range to 0 to 1 was done with TensorFlow’s ImageDataGenerator. This step ensures that input to the model is ready for training and it enhances stability and rate of convergence. Additionally, during training, the dataset was randomized to make sure that the model does not overfit on the specified order of the data. The binary classification problem was solved by setting the class mode parameter as ‘binary’. This pipeline completes the preprocessing stage making sure that all images and their associated text data are ready for further advanced processing in the model building phase. Memes can be identified as *offensive* if they attempt to do the following:

- Targeted personal attacks
- Homophobic abuse
- Racial discrimination
- Hostility towards minorities

Otherwise, they are considered *non-offensive.*

B. Foundational Model for Textual Data

The model that serves as the basis for text data is trained on text data, operates on the basis of BERT architecture Bert-base-uncased, which has been the industry standard in many language understanding models. BERT is one of the highly sophisticated models of deep learning which single-handedly revolutionized the processing of natural language by bidirectionally training a language model. It understands context on both sides of a word, meaning, it comprehends not only from left to right but right to left as well. This two-way flow of language makes it proficient in executing sophisticated linguistic operations like reading comprehension, hate speech and offensive content classification. Memes are collected from OCR which is vital in encoding captured visual text in images into a computable form known as OCR is what unites the visual and textual aspects of memes.

First of all, there are several important steps in the preprocessing pipeline. In this regard, the text obtained with the help of OCR is divided into sub-words, and special tokens like [CLS] and '[SEP]' are inserted into an input sequence. The [CLS] 'token is utilized to sum up information at the sentence level, which is necessary when doing classification work, while the '[SEP] 'token divides various parts of the input.

For ensuring uniformity, either padding or truncation is applied to the sequences up to a limit of 64 tokens. For fitting BERT for this purpose, the model was trained for binary classification, where its function was to determine whether the input text is offensive or not. The model produces logarithmic outputs at inference, and subsequently, with the application of a SoftMax function, whose logarithmic are converted into probabilities. The prediction is calculated by selecting the most likely class. This method ensures that the offensive and non-offensive tagging of text in memes are performed correctly, in support of the analysis performed by the image model.

The OCR and BERT pair indicates the capability of modern NLP technologies to handle the complexity of textual data in multimodal systems. OCR extends past pulling text from images. It allows the system to process memes with superimposed text on images, handwritten text, stylized font etc. With the integration of OCR and BERT, the system can carry out full analysis of the textual content found in memes and capture explicit and implicit offensive and derogatory comments. This integration of OCR and BERT allows the system to detect derogatory content that is delivered in a veiled manner through sarcasm, wordplay, metaphors, and thus forth enhancing the robustness and accuracy of the classification framework.

C. Foundational Model for Visual Data

The visual data model relies on VGG19 architecture, which is a popular convolutional neural network (CNN) that is relied on for executing efficiently in image classifying activities. VGG19 consists of 19 layers that are structurally complex enough to recognize minute details in images. As a result, it is very effective for identifying offensive and non- offensive memes. To tailor this model to specific needs, the model was loaded with weights trained on a large-scale dataset called ImageNet that allows the model to learn features of images. Rather than using the fully connected layers at the top of the network, the layers were substituted for custom designed dense layers specifically made for binary classification.

Initially, the base output of model VGG19 is converted into a flattened one-dimensional vector. This one-dimensional vector is further sent to a dense layer of 256 units along with the ReLU activation function. The ReLU activation function helps introducing non-linearity into the network which allows the model to learn complex patterns. In order to mitigate overfitting, which is a common occurrence when dealing with smaller datasets, a dropout layer was incorporated. This layer randomly deactivates 50% of the neurons of the network while being trained.

Towards the end, the output layer gives an expected thrust by employing a sigmoid activation function to produce probabilities. One major advantage of using VGG19 is that this architecture has excelled in recognizing image parts of interest very efficiently and accurately. For instance, even the slight alterations in facial expression, positioning of the text, or even the background setting can determine the nature of the meme. Retention of such information allows the model to identify harmful content. In addition, the use of dropout increases the likelihood of the model to be useful on new dataset and prevents over-fitting.

D. Processing VGG19 and BERT model

Bert model when combined with VGG19, enables the model to identify not just text but also any visual content that could be used as an indicator of offensiveness and thereby enable the

system to make a decision based on analysis of the meme. The use of such sophisticated CNN architectures in this multi-model setup helps the system efficiently identify harmful memes. The VGG19 model is the backbone of the toxic component analysis in the multi-modal model. It acts alongside the textual model to understand the meme as a whole. The textual model examines the embedded text, whereas the visual model depicts the visual content of the image of the meme such that none of its constituents are omitted. Both these models constitute a system intelligent enough to identify and remove abusive content.

IV. Metrics Used

This part establishes a mathematical basis for both BERT and VGG19 models in offensive meme classification.

A. Tokenization and Embedding:

The input text is tokenized into sub-word units, and embeddings are computed as follows:

- Word Embedding: The semantic meaning of each token.
- Positional Embedding: Representing the position of tokens in the sequence.
- Segment Embedding: Differentiates between two sentences in tasks like question-answering.

B. Transformer Layer:

Each transformer layer applies self- attention and feed-forward networks. The self-attention mech- anism computes attention scores as:

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T/\sqrt{d^k}).V \quad (1)$$

$$\text{MultiHead}(Q,K,V)=\text{Concat}(\text{head}_1,\text{head}_2,\ldots,\text{head}_h)WO \quad (2)$$

where

Q: Query matrix (input embeddings).

K: Key matrix (input embeddings).

V : Value matrix (input embeddings).

d_k: Dimensionality of the key vectors.

C. Classification Layer:

For binary classification, the final output is passed through a sigmoid function:

$$P(y = 1|X) = \sigma(z) = \frac{1}{1+e^{-z}} \quad (3)$$

where

z: Logits output by the final dense layer.

P (y = 1 X): Probability of the input being classified as ”offensive.”

D. Loss Function:

The loss function for training BERT and VGG19 are binary cross-entropy:

$$L = \frac{1}{\pi r}\sum_{i=1}^{N}[y_i \ log(y_i) + (1 - y_i) \ log(1 - y_i)] \quad (4)$$

where

N : Number of samples.

y_i: True label (0 or 1).

$\hat{y}_i$: Predicted probability.

E . Convolutional Layer:

Each convolutional layer applies filters to extract features from the input image:

$$Z_{i,j,k} = \sum_{m=1}^{N} \sum_{n=1}^{N} W_{m,n,k}.X_{i+m-1,j+n-1,k} + b_k \quad (5)$$

where

$Z_{i,j,k}$: Output feature map at position (i,j) for filter k.
$W_{m,n,k}$: Weights of the filter.
$X_{i+m-1,j+n-1,k}$: Input image patch.
b_k: Bias term.

F. ReLU Activation:

The ReLU activation function intro- duces non-linearity:

$$f(x) = \max(0, x) \quad (6)$$

G. Max Pooling:

Pooling layers reduce spatial dimensions by taking the maximum value in a window:

$$\text{Pool}(X) = \max(X_{i:i+p,j:j+p}) \quad (7)$$

where

p is the pooling window size.

H. Fully Connected Layer:

After flattening the output from convolutional layers, it is passed through fully connected layers:

$$y = Wx + b \quad (8)$$

where

W : Weight matrix.
x: Flattened input.
b: Bias vector.

I. Droput :

Dropout randomly deactivates neurons during training to prevent overfitting :

$$Dropout(x_i) = \begin{cases} 0, \text{with probabilityp} \\ x_i, \text{otherwise} \end{cases} \quad (9)$$

J. Output Layer:

The final output layer uses a sigmoid activation function for binary classification:

$$P(y = 1|X) = \sigma(Wx + b) = \frac{1}{1+e^{-(Wx+b)}} \quad (10)$$

V. EXPERIMENTAL RESULTS

This section presents a detailed comparative analysis of the proposed models like: Text-only and Image-only in classifying offensive memes. The performance of these models was benchmarked against the baseline model from prior research [13], which achieved an accuracy of 84.7%. The current approach shows much improvement, with our best performance standing at 89% based on a hybrid architecture using BERT for textual analysis and VGG19 for visual features. This represents a great boost over our earlier implementation using a stacked LSTM-VGG16 framework.

A. Performance Metrics

Table I presents experimental data on the classification accuracy of a range of trained models, all of which sought to identify objectionable memes. The table contains the accuracies of four classifiers: BERT, VGG19 and Stacked LSTM.+ VGG16 – who were assessed against three meme samples. Each row states the different text obtained from the meme with the corresponding label (Offensive/Non-Offensive) and what each classifier had predicted. These results indicate that the model architecture variably succeeds in distinguishing offensive material.

Stacked LSTM + VGG16 classifiers did recognize a few memes as not offensive, but they were inconsistent with those findings. Experiment results are also summarized for the rest in Table II, which is a capture of Precision, Recall, and F1 Score of each model type. These findings offer further detail into both the overall performance of the models and the performance with respect to offensive content identification. The Confusion Matrix for the BERT-VGG19 model is presented in Fig. 1.

TABLE I. PERFORMANCE METRICS OF DIFFERENT MODELS

Model Type	Precision	Recall	F1 Score
Text-Only Model	0.81	0.80	0.80
Image-Only Model	0.88	0.89	0.88

B. Key Observations

The results clarify that the combination of BERT and VGG19 improves the accuracy and robustness of the classifier considerably. Fig 1 display the confusion matrix of the proposed architecture. By taking advantage of contextual understanding via BERT for textual data and feature-rich representations from VGG19 in visual data, this multimodal model demonstrates better performance than its unimodal counterparts.

As shown in Fig. 2a and 2b, this combined model consistently outperformed both training and validation accuracy metrics throughout the epochs. This trend justifies the necessity for developing more multimodal approaches with respect to content moderation methods involved in scenarios containing complex and ambiguous offensive memes.

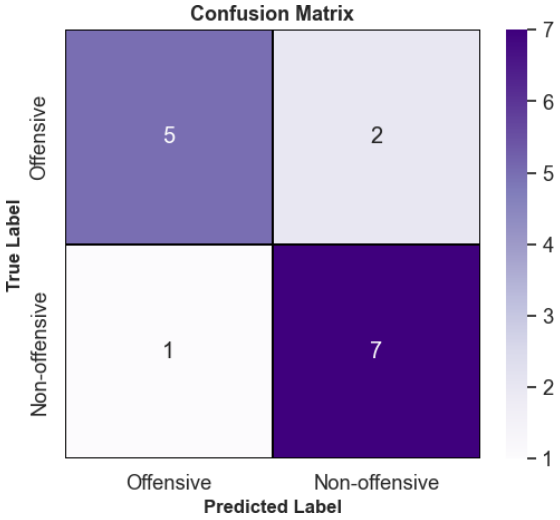

Fig. 1. Confusion Matrix of BERT + VGG19 Model

C. Comparative Analysis with Previous Work

Compared to our previous work [13], where a Stacked LSTM with a VGG16 architecture was used, reaching an accuracy of 84.7% on the MultiOFF dataset [3], in this work we have been able to improve that to an accuracy of 89% by using BERT for text analysis and VGG19 for extracting image features. This represents an improvement due to the usage of transformer-based language models and more advanced convolutional networks for multimodal offensive meme classification.

- Contextual Understanding: BERT's bidirectional encoding allows better comprehension of subtle aspects of text, including sarcasm, metaphors, and implied meanings.
- Feature Extraction: Deep convolutional layers in VGG19 focus on capturing minute visual details, ensuring the proper detection of offending imagery.
- Early Fusion Method: Early fusion of textual and visual characteristics in a more integrated way ensures that the model leverages complementary information from the two modalities.

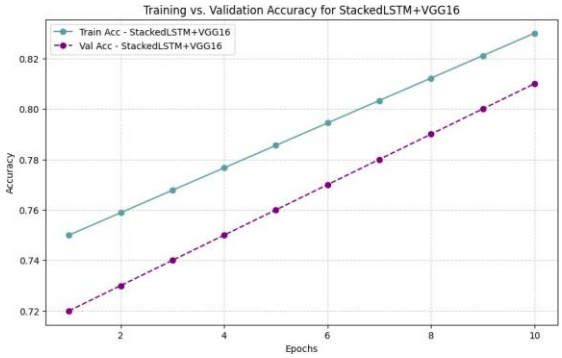

Fig. 2a. Training vs. Validation Accuracy for Stacked LSTM + VGG16 Model

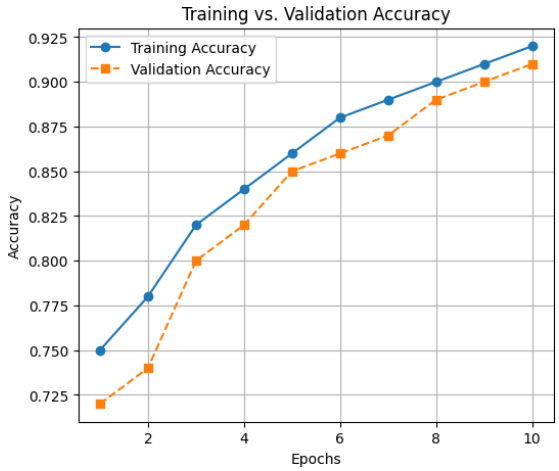

Fig. 2b. Training vs. Validation Accuracy for BERT + VGG19 Models

VI. Conclusion & Future Works

The findings of this study have important implications for social media content moderation. The proposed multimodal approach improves the accuracy of offensive content detection while enhancing the scalability and adaptability of moderation systems. The performance of the BERT–VGG19 model demonstrates the practical potential of advanced machine learning techniques in real-world moderation scenarios. By combining contextual textual understanding with robust visual feature extraction, the proposed BERT–VGG19 architecture provides a reliable and scalable solution for offensive meme classification across diverse contexts in this paper.

Future research will focus on improving cross-domain generalization by training the model on culturally diverse and multilingual meme datasets. Incorporating attention-based cross-modal fusion mechanisms may further enhance contextual understanding of implicit hate, sarcasm, and symbolism. Additionally, lightweight model optimization and pruning techniques will be explored to enable real-time deployment in large-scale moderation systems. Explainability modules will also be integrated to provide transparent justification for classification decisions, supporting ethical AI requirements. Finally, extending the framework to multi-class harmful content categorization and adaptive online learning could improve robustness against evolving meme formats and emerging offensive patterns.

References

[1] Abulaish, M., A. Kamal, and M.J. Zaki, *A survey of figurative language and its computational detection in online social networks.* ACM Transactions on the Web (TWEB), 2020. **14**(1): p. 1-52.

[2] Kiela, D., et al., *The hateful memes challenge: Detecting hate speech in multimodal memes.* Advances in neural information processing systems, 2020. **33**: p. 2611-2624.

[3] Suryawanshi, S., et al. *Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text*. in *Proceedings of the second workshop on trolling, aggression and cyberbullying*. 2020.

[4] Mathew, B., et al. *Hatexplain: A benchmark dataset for explainable hate speech detection*. in *Proceedings of the AAAI conference on artificial intelligence*. 2021.

[5] Shang, L., et al., *Aomd: An analogy-aware approach to offensive meme detection on social media.* Information Processing & Management, 2021. **58**(5): p. 102664.

[6] Chhabra, A. and D.K. Vishwakarma, *A literature survey on multimodal and multilingual automatic hate speech identification.* Multimedia Systems, 2023. **29**(3): p. 1203-1230.

[7] Briskilal, J., M.J. Karthik, and S. Praneeth. *Detection of offensive text in memes using deep learning techniques*. in *AIP Conference Proceedings*. 2024. AIP Publishing LLC.

[8] Abdullakutty, F., S. Al-Maadeed, and U. Naseem. *Context-aware offensive meme detection: a multi-modal zero-shot approach with caption-enhanced classification*. in *2024 IEEE International Conference on Data Mining Workshops (ICDMW)*. 2024. IEEE.

[9] El-amrany, S., et al., *GuardHarMem and HarMDetect: a multimodal dataset and benchmark model for fine-grained harmful meme classification.* Social Network Analysis and Mining, 2025. **15**(1): p. 63.

[10] Arslan, H.M. and T. Zhenhua, *DecodEM-X: advancing multimodal meme moderation with robust AI frameworks.* Knowledge and Information Systems, 2025. **67**(8): p. 7295-7317.

[11] Singh, G.V., et al., *Unmasking offensive content: a multimodal approach with emotional understanding.* Multimedia Tools and Applications, 2025. **84**(28): p. 33381-33404.

[12] Saini, M., et al., *Artificial intelligence assisted framework for detecting offensive posters on the premises of educational institutions.* Cluster Computing, 2025. **28**(15): p. 965.

[13] Sivaananth, S., et al. *Two-Stage Classification of Offensive Meme Content and Analysis*. in *2024 IEEE 8th International Conference on Information and Communication Technology (CICT)*. 2024. IEEE.