TOXISCOPE: Toxic Meme Classification Using Machine Learning

Noell Biju Michael[1], Fabeela Ali Rawther[1], Dr Geevarghese Titus[2]
[1]Computer Science & Engineering, [2]Electronics & Communications Engineering
Amal Jyothi College of Engineering (Autonomous)
Kanjirappally, Kerala, India
noellbijumichaelmtech2027@cs.ajce.in, fabeelaalirawther@amaljyothi.ac.in
geevarghesetitus@amaljyothi.ac.in

Abstract—Social media platforms have significantly increased the use of memes as a medium for communication and online interaction. Unlike plain text, memes combine visual and textual elements to convey meaning, making toxicity detection more difficult due to sarcasm, implicit hate, and contextual dependence. Conventional toxicity detection methods that rely on a single modality are often unable to accurately interpret such content [1], [3].

This work presents Toxiscope, a multi-modal meme toxicity classification framework developed using machine learning and deep learning approaches. Multiple models including a rule-based classifier, TF-IDF with Logistic Regression, a BERT-based text classification model [2], and an image-based NSFW classification model were implemented and comparatively analyzed. The study evaluates how different approaches perform in identifying toxic meme content and examines the limitations of uni-modal systems in meme understanding tasks.

To improve prediction capability, textual and visual toxicity scores were combined using a multi-modal scoring approach. Experimental results show that the multi-modal framework performs more consistently than individual text-only or image-only models, particularly in memes where harmful meaning emerges from the interaction between text and image. The study also identifies practical challenges such as contextual ambiguity, symbolic hate, dataset imbalance, and limited visual semantic understanding. The findings demonstrate the importance of multi-modal learning in content moderation systems and provide a baseline for future improvements using advanced vision-language models and fusion techniques.

I. INTRODUCTION

The rapid growth of social media platforms in this generation has resulted in a large increase in content such as memes, which combine text and images to communicate ideas, opinions, humor, and social commentary. Although memes are widely used for entertainment, they are also frequently misused to spread hate speech, offensive language, racism, and misinformation. Detecting toxicity in memes is more challenging than traditional text classification because the intended meaning often depends on the relationship between textual and visual elements rather than a single modality alone [1].

Initial toxicity detection systems mainly relied on keyword matching and classical machine learning approaches for textual analysis. While these methods are effective in identifying explicit offensive language, they perform poorly when handling sarcasm, implicit hate, contextual meaning, or ambiguous expressions commonly found in memes. Recent advancements in natural language processing, in architectures such as BERT, have improved contextual understanding in text classification tasks by analyzing semantic relationships between words and sentences [2]. However, text-only models are still limited when the visual component of a meme significantly alters or reinforces its meaning.

Visual analysis models provide additional information by identifying unsafe or inappropriate image content. Here, an NSFW-based image classifier was used to analyze meme images and generate image toxicity scores. While such models can detect explicit visual content, they are less effective in identifying symbolic hate, stereotypes, or context-dependent offensiveness conveyed through imagery [3]. These limitations highlight the need for multi-modal approaches that combine textual and visual analysis to better interpret meme semantics.

To tackle the problem of meme toxicity detection, this study explores and compares several approaches: a rule-based model, TF-IDF with Logistic Regression, a BERT-based text classifier, and an image-based toxicity detector. Building on these, a multi-modal framework is introduced that combines text and image scores to enhance prediction accuracy. The models are assessed using standard metrics such as accuracy, precision, recall, and F1-score. In addition, the study highlights the strengths and weaknesses of each method, offering insight into the practical challenges of developing effective multi-modal systems for meme toxicity detection.

II. RELATED WORK

Memes have become one of the most common ways people communicate online, blending images and text to share humor, opinions, or social commentary. This mix of modalities makes toxicity detection particularly challenging, since the meaning of a meme often depends on how the caption and image interact. Past work, such as the Hateful Memes Challenge, has shown that text-only or image-only models often miss harmful intent when it is implied indirectly. Researchers have also noted that sarcasm, cultural references, and contextual ambiguity add further complexity, making meme classification harder than traditional text-based toxicity detection.

Early attempts to address toxicity relied on rule-based systems and keyword matching. While these methods were simple and efficient, they struggled to capture context or subtle meaning. Classical machine learning approaches, like TF-IDF with Logistic Regression, improved performance by learning patterns from data rather than relying solely on predefined keywords. Yet, they still treated words as isolated

features, limiting their ability to understand semantics. Transformer models such as BERT marked a significant step forward by analyzing relationships between words in context, enabling better detection of indirect or nuanced toxicity. Even so, text-only models remain limited when the image alters or reinforces the meaning of the meme.

To overcome these limitations, researchers have explored image-based and multi-modal approaches. Visual classifiers can detect explicit content but often fail to recognize symbolic hate or stereotypes. More recent work has focused on combining text and image features, with tasks like Memotion Analysis showing improved results when both modalities are considered together. Models such as CLIP and multi-modal ensembles demonstrate stronger performance by jointly learning relationships between captions and visuals, though they demand complex architectures and large datasets. Against this backdrop, the present study compares a range of methods—from simple rule-based systems to deep learning and multi-modal frameworks—examining how each performs individually and how fusion strategies enhance toxicity detection. It also reflects on the practical challenges these approaches face, especially in dealing with implicit hate and contextual ambiguity.

III. METHODOLOGY

A. Dataset

For this study, two datasets were used: Hateful Memes dataset and a custom dataset. The Hateful Memes dataset provides a collection of meme images paired with text captions, each annotated with binary labels indicating whether the meme is toxic or non-toxic. Because many memes derive their meaning from both text and visuals, this dataset is particularly well-suited for multi-modal analysis, offering examples where harmful intent is often implied through subtle relationships between image and caption.

A custom dataset of 100 memes was assembled to capture a broader range of content commonly encountered on social media. This collection includes examples of harmless humor alongside offensive, racist, sexist, Islamophobic, and dark humor memes. Each entry in the dataset contains three key components: the meme image, its associated text, and a binary toxicity label. By using these categories, the custom dataset adds diversity reflecting the kinds of challenges toxicity detection systems face in online environments.

Before training and evaluation, the dataset was pre-processed by converting text to lowercase, removing missing entries, and generating valid image paths. The combined dataset was then divided into training and testing sets using an 80:20 ratio. Textual data was processed using TF-IDF vectorization and transformer-based tokenization, while image data was processed separately for visual toxicity analysis.

TABLE 1
DATASET STATISTICS

Dataset Component	**Count**
Hateful Memes Dataset Samples	8400
Custom Meme Samples	100
Toxic Samples	3019
Non-Toxic Samples	5481
Total Samples	8500

```
data[['img', 'text', 'toxicity_score']].head(10) #For getting 10 entries
```

[6]:

	img	text	toxicity_score
0	img/42953.png	its their character not their color that matters	0.1
1	img/23058.png	don't be afraid to love again everyone is not ...	0.1
2	img/13894.png	putting bows on your pet	0.1
3	img/37408.png	i love everything and everybody! except for sq...	0.3
4	img/82403.png	everybody loves chocolate chip cookies, even h...	0.1
5	img/16952.png	go sports! do the thing! win the points!	0.1
6	img/76932.png	fine you're right. now can we fucking drop it?	0.3
7	img/70914.png	tattoos are bad for your health i know 5 milli...	0.6
8	img/02973.png	how long can i run? till the chain tightens	0.1
9	img/58306.png	what is he hiding? we need to see his tax retu...	0.1

Fig. 1. Sample entries from the dataset showing image paths, text, and computed toxicity scores

B. Data Pre-processing

Before model implementation, the dataset was pre-processed to ensure consistency between text and image data. All textual content was converted to lowercase, and missing or invalid entries were removed. Image paths were verified to ensure that each meme image was correctly linked to its corresponding text and label.

The custom dataset was merged with the benchmark dataset, and the indices were reset after concatenation. During experimentation, random sampling was also used in some stages to reduce computational load during evaluation. These pre-processing steps helped maintain proper synchronization between textual and visual data throughout the pipeline.

C. Model 1: Rule-Based Approach

A simple rule-based model was implemented as a baseline approach for toxicity detection. The model uses a manually defined list of offensive and toxic keywords commonly associated with hate speech or abusive language. If any keyword is detected in the input text, a toxicity score is assigned accordingly.

This approach is computationally simple and easy to implement, but it cannot understand contextual meaning, sarcasm, or implicit toxicity. The model also fails in cases where toxicity depends on the interaction between text and image.

D. Model 2: Image-Based NSFW Classification

To analyze the visual component of memes, a pre-trained NSFW image classification model was used. The model generates a probability score representing the likelihood of unsafe or explicit visual content. In this work, the generated probability was used as the image toxicity score.

Although the model can identify explicit visual content, it cannot detect contextual hate, stereotypes, or symbolic toxicity within images. This limits its effectiveness for meme understanding tasks where meaning depends on deeper semantic interpretation.

E. Model 3: TF-IDF + Logistic Regression

A classical machine learning approach using TF-IDF vectorization and Logistic Regression was implemented for text classification. TF-IDF converts text into numerical feature vectors based on word importance within the dataset.

In this implementation, the TF-IDF vectorizer was configured with a maximum feature size of 5000.

The dataset was divided into training and testing sets using an 80:20 ratio. Logistic Regression was then trained on the generated TF-IDF vectors for binary toxicity classification. Compared to rule-based methods, this approach provides better generalization, but it still struggles with sarcasm, contextual meaning, and implicit toxicity.

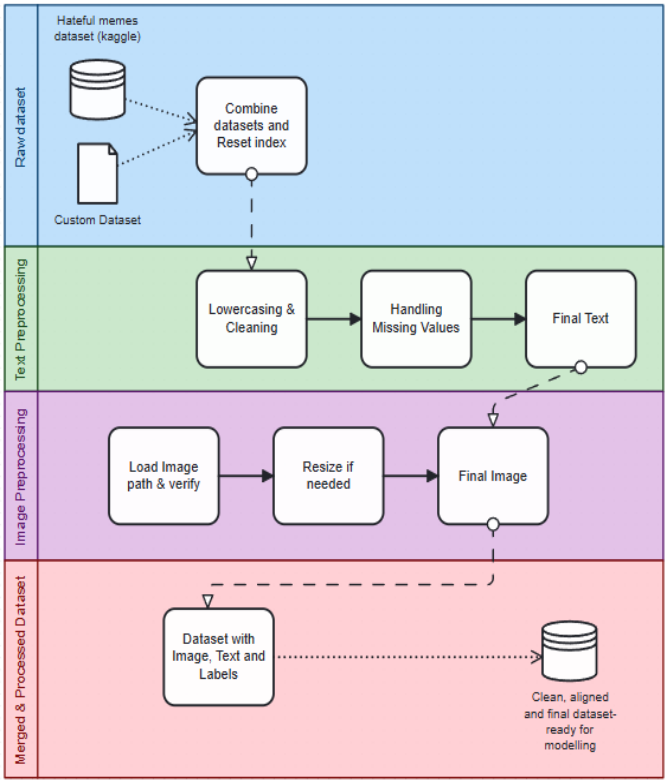

Fig. 2. Data preprocessing and multi-modal pipeline for meme toxicity detection

F. Model 4: BERT-Based Text Classification

A transformer-based BERT model was used to improve text toxicity classification by understanding contextual meaning within meme captions. The pre-trained model unitary/toxic-Bert was used to generate toxicity predictions and confidence scores for input text. Unlike traditional machine learning models that mainly rely on word frequency, BERT processes text bidirectionally and analyzes how words relate to one another within a sentence. This helps the model better interpret semantic meaning and contextual intent.

The model generates a predicted label along with a confidence score for each meme caption. Toxic predictions with higher confidence produce higher toxicity scores, which are later used in the multi-modal pipeline. During evaluation, the BERT model performed better than rule-based and TF-IDF approaches in identifying indirect toxicity, sarcasm, and context-dependent offensive language. Even though the model provides strong textual understanding, it only analyzes text and does not consider the associated meme image during prediction.

G. Model 5: Multi-modal Approach

A multi-modal approach was implemented by combining the text toxicity score generated by BERT and the image toxicity score generated by the NSFW classifier. The final prediction score was calculated using:

$$Final\ Score = Text\ Score + Image\ Score$$

The combined score was then filtered using a value of 0.5 to classify memes as toxic or non-toxic. By combining textual and visual information, the model performs better in cases where toxicity depends on the interaction between image and text. However, the overall performance is still limited by the capability of the image classification model and the difficulty of understanding implicit meme semantics.

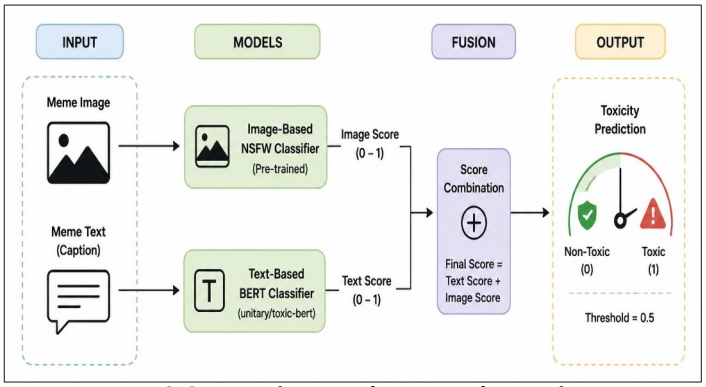

Fig. 3. System architecture of Toxiscope framework

IV. RESULTS AND DISCUSSION

A. Evaluation Metrics

The performance of the models was evaluated using standard classification metrics including accuracy, precision, recall, and F1-score. These metrics help in analyzing how effectively the models classify toxic and non-toxic memes. Since the dataset contains both toxic and non-toxic samples, relying on a single metric alone is not sufficient for proper evaluation. Multiple metrics were used to obtain a more balanced understanding of model performance.

- Accuracy measures the overall correctness of the classification model by calculating the ratio of correctly predicted samples to the total number of samples.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precision measures how many of the memes predicted as toxic were actually toxic. Higher precision indicates fewer false positive predictions.

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

- Recall measures how effectively the model identifies actual toxic memes from the dataset. A higher recall value indicates better detection of toxic content.

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

- The F1-score provides a balanced measure between precision and recall and is especially useful for imbalanced datasets.

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Among these metrics, the macro F1-score was considered particularly important in this study because it evaluates performance across both toxic and non-toxic classes equally, providing a more balanced assessment of model effectiveness.

B. *Comparative performance of Models*

The performance of the implemented models was evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score. Among these, the macro F1-score is particularly important as it ensures balanced evaluation across both toxic and non-toxic classes, making it suitable for this task.

The rule-based model, used as a baseline, exhibited the weakest performance. Its reliance on predefined keywords limits its ability to capture contextual meaning, leading to both false positives and false negatives. For instance, the presence of certain keywords may not necessarily indicate toxicity, while implicit or sarcastic toxic expressions often go undetected.

The image-based NSFW classification model also demonstrated limited effectiveness in this context, even though it is capable of identifying explicit visual content. It does not capture semantic or contextual toxicity present in memes. Which is why most images in the dataset received negligible image scores, indicating that explicit content detection alone is insufficient for meme toxicity classification.

The TF-IDF with Logistic Regression model showed moderate performance and improved upon the rule-based approach by learning patterns from data. But its inability to model contextual relationships between words restricts its effectiveness in handling complex or nuanced textual content.

In contrast, the BERT-based model achieved significantly better results due to its bidirectional contextual understanding. By analyzing the relationship between words within a sentence, BERT effectively captures implicit toxicity and subtle linguistic cues. This leads to improved precision and recall, establishing it as the most effective uni-modal model in this study.

TABLE 2
MODEL COMPARISON

Model	Type	Input	Strength
Rule Based	Heuristic	Text	Simple, Fast
TF-IDF+LR	Machine Learning	Text	Learns Patterns
BERT	Deep Learning	Text	Strong Contextual Analysis
NSFW	Image Model	Image	Detects Explicit Content
Multi-modal	Combined	Text+Image	Best Overall Performance

C. *Effectiveness and limitations of the multi-modal approach*

To reduce the limitations of text-only and image-only models, a multi-modal approach was implemented by combining the toxicity scores generated from the BERT model and the NSFW image classifier. The idea behind this approach was to use contextual information from text together with visual information from meme images to provide more balanced prediction by combining textual and visual information.

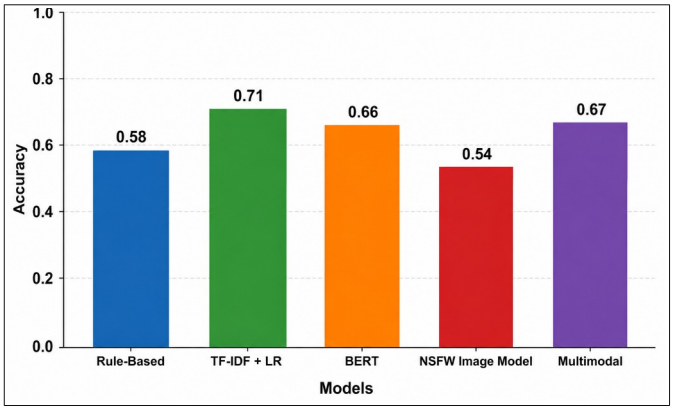

Fig. 4. Comparison of model accuracy chart

The multi-modal model achieved an accuracy of around 67% with a macro F1-score of 0.63. Although the performance was slightly better than some of the uni-modal approaches, the improvement was not very significant. During evaluation it was observed that the text-based model contributed more strongly to the final prediction compared to the image-based model. One major reason for this is the limitation of the NSFW classifier used in the project. The model is mainly designed to detect explicit or unsafe visual content and is not capable of understanding contextual or symbolic meaning within memes. Many toxic memes do not contain explicit visuals, but instead rely on sarcasm, stereotypes, or the relationship between the image and text. In such cases, the image model contributes very little useful information.

The fusion method used in this work was also relatively simple, where the text score and image score were directly combined to generate the final prediction. While this helped in implementing a basic multi-modal pipeline, it does not fully capture the complex interaction between textual and visual features.

Even with these limitations, the multi-modal approach demonstrated that combining both modalities can improve meme toxicity detection compared to relying only on text or image information individually. The results also indicate that better image understanding models and more advanced fusion techniques could further improve performance in future work.

TABLE 3
MULTI-MODAL MODEL RESULTS

Class	Precision	Recall	F1-score	Support
Non-Toxic (0)	0.71	0.76	0.73	123
Toxic (1)	0.57	0.51	0.53	77

D. *Future work*

The proposed system demonstrates that combining textual and visual information can improve meme toxicity detection compared to relying on a single modality alone. However, several limitations were observed during implementation and evaluation. One of the major limitations comes from the image classification model used in this work. The NSFW classifier is mainly designed to detect explicit visual content and is not capable of understanding contextual or symbolic toxicity present in many memes. As a result, the

contribution of the image modality to the final prediction remained limited in several cases. The multi-modal pipeline also uses a simple score combination method for generating predictions, where the text and image scores are directly combined. While this approach is easy to implement, it does not fully capture the complex interaction between textual and visual features. More advanced multi-modal models and fusion techniques, such as attention-based fusion or joint representation learning, could help improve performance by learning stronger relationships between image and text content.

Another important area for improvement is the dataset and overall model efficiency. Although a custom meme dataset was created and combined with the benchmark dataset, the collected samples still do not completely represent the diversity of toxic meme content commonly found on social media. Expanding the dataset with more varied meme styles, cultural references, and implicit forms of toxicity can improve model generalization and reduce bias. In terms of implementation, transformer-based models such as BERT provide strong contextual understanding but also increase computational cost and inference time. Future work can therefore explore lightweight transformer architectures, model compression techniques, and additional machine learning models such as ensemble methods or advanced multi-modal transformers. These improvements can help build a more accurate, scalable, and robust multi-modal toxicity detection system for practical content moderation applications.

V. Conclusion

In this project, different approaches for detecting toxicity in memes were explored, starting from simple rule-based methods to more advanced machine learning and deep learning models. The initial models showed that relying only on keywords or basic text features is not enough, especially when the meaning of a meme depends heavily on context. The BERT model performed much better compared to the earlier approaches, mainly because it understands the context of sentences instead of just individual words. This made it more reliable for identifying subtle or indirect forms of toxicity.

Since memes are not just text but also images, a combined approach was implemented using both text and image scores. While this multi-modal model did show some improvement, the difference was not very large. This was mainly because the image model used was limited to detecting explicit content and could not understand the actual meaning behind the image.

From this it is clear that text plays a major role in toxicity detection, but image understanding is still an area that needs improvement. Better image models and smarter ways of combining text and image information could lead to stronger results. Overall, this work helped in understanding how different models behave on the same problem and where they fall short. It also gives a clear direction on what can be improved in the next phase of the project.

References

[1] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine, “The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes,” arXiv:2005.04790, 2021.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” arXiv:1810.04805, 2019.

[3] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal Machine Learning: A Survey and Taxonomy,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.

[4] G. Arya, M. K. Hasan, A. Bagwari, N. Safie, S. Islam, and M. Ahmed, “Multimodal Hate Speech Detection in Memes Using Contrastive Language-Image Pre-Training,” IEEE Access, vol. 12, pp. 1–14, 2024.

[5] D. S. M. Pandiani, E. Tjong Kim Sang, and D. Ceolin, “Toxic Memes: A Survey of Computational Perspectives on the Detection and Explanation of Meme Toxicities,” arXiv:2406.07353, 2024.

[6] C. Sharma et al., “SemEval-2020 Task 8: Memotion Analysis – The Visuo-Lingual Metaphor,” arXiv:2008.03781, 2020.

[7] B. Chen et al., “Visual Perspective Taking for Opponent Behavior Modeling,” arXiv:2105.05145, 2021.

[8] A. Radford et al., “Learning Transferable Visual Models From Natural Language Supervision,” arXiv:2103.00020, 2021.

[9] Y. Zhou and Z. Chen, “Multimodal Learning for Hateful Memes Detection,” arXiv:2011.12870, 2020.

[10] A. El-Sayed et al., “Multimodal Hate Speech Detection Using CLIP and BERT-Based Ensemble Models,” Proceedings of CASE 2024 (ACL Workshop), 2024.

[11] “A Survey of Multimodal Hate Meme Detection,” Expert Systems with Applications, Elsevier, 2026.

[12] R. Xing et al., “Is AI Ready for Multimodal Hate Speech Detection? A Comprehensive Dataset and Benchmark Evaluation,” arXiv:2603.21686, 2026.