International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(14s) 2026 pp. 644-658
https://doi.org/10.70917/ijcisim-2026-4272
Article

SEMTEX++: Semantic & Threat-aware Explainable Framework for Malicious Content

N. Deepika[1], Kempanna M.[2]

[1]Bangalore Institute of Technology, Bengaluru, Affiliated to Visvesvaraya Technological University (VTU), Belagavi, Karnataka, India.
Email: deepikajvijay@gmail.com
ORCID: 0000-0001-8956-1397
[2]Department of Artificial Intelligence and Machine Learning, Bangalore Institute of Technology, Bengaluru, Karnataka, India.
Email: kempsindia@gmail.com
ORCID: 0000-0001-6999-2130

Abstract: The information that is circulating in social networks is no longer only text but mostly text along with images and videos (35,36,37). And this content is not always genuine. When there is content, people try to make it fraudulent in various media formats such as internet memes, posts combining text + images and messages combined with visual context. The amount of information spreading is countless, with this much bulk information finding the messages that are genuine is a major challenging task for the owners who are maintaining the social networking sites. For identifying misinformation, if we completely rely on traditional filtering tools then it won't work as this malicious intent not only comes from the images or texts but also through the semantic relationship between these channels (4,5,19,20). This research introduces SEMTEX++, which is an ensemble framework with hybrid clustering, semantics, threat-aware, and explainable generative AI to increase the chance of identifying and analyzing toxic content within social media. Our unified framework produced good performance with an overall accuracy of 91%, F1-score of 0.90, and AUC of 0.95 when compared to text-only, vision-language, and LLM baselines. Results demonstrate that this multimodal fusion approach improves the detection of indirect harmful content, sarcasm, meme-based hate speech and misinformation. The explanations module achieved the highest faithfulness and human evaluation scores among all evaluated methods.
Keywords: Malicious detection, Machine Learning, Multi modal threat index, Classification, Clustering, LLMs, Misinformation spread, Social media content, Visual language.

1. Introduction

Social media has become the main bridge to people in many ways in terms of spreading information, news and sharing opinions. At the same time, it has become the main victim of spreading harmful content like hate speech, cyberbullying, misinformation and harassment. Through these networks, toxic people send toxic posts so accurately that no one could imagine that the information is completely wrong.

During the initial stages, the research is mostly focused on textual analysis. Hats off to Machine learning and natural language processing, which helped us to succeed in identifying offensive language, abusive messages and toxic conversations (25,38). The next era (26,27) is with transformer-based models such as BERT (7,24), RoBERTa(8) and DeBERTa (9), which improved the detection through contextual understanding and linguistic meaning. Despite these improvements, text-based approaches faced various limitations as the posts are now no longer only text but have become multimodal.

Memes provide a clear example of this challenge. In many cases, the text and the image embedded within a meme seem to be harmless when seen independently. The harmful meaning comes out when both components are seen together. A sarcastic message combined with an image might communicate discrimination, mockery or hatred, without containing any direct offensive words. Similar patterns can be seen in misinformation campaigns where

misleading stories or messages are often amplified through carefully created visual content. As a result, approaches that analyse only text or only images normally fail to capture the actual intent of the message if the messages are indirect. Recent multi modal researches have addressed this problem (11,12,14).

Vision-supported models like Visual BERT(4), BLIP-2(5), CLIP(20) and LLaVA(19), which could identify relationships between text and visual data, have shown good results in multimedia analysis and meme understanding. Applying these techniques to harmful content detection remains still ON, with several unresolved challenges. One such challenge is explainability. In practical situations, simply predicting whether content is harmful is often not enough. Users who are checking the results need to understand why a particular piece of content was flagged as malicious. Many existing approaches miss these explanations in the model decision(23,24,27). Another challenge is the dynamic nature of online communication. Content writers usually find strategies to avoid harmful content detection and try to include new slang, coded expressions, cultural references and visual symbols.

Models trained on static datasets may struggle to identify malicious patterns in posts with different language and cultural contexts. Instead of solely depending on predictive performance, it is very important to understand the severity of harmful content with broader semantic patterns. These aspects remain relatively under-explored in current multimodal moderation research(35,37).

1.1.Research Problem

Keeping all these challenges in mind, a hybrid Malicious Content Detection framework is developed that performs multilingual text understanding, visual representation learning, semantic clustering, ensemble classification, explainable AI and generative reasoning sequentially. The main principle of our research is that toxic intent in social media content is often contextual and multi modal; hence, it is difficult to explain using conventional approaches. But it is feasible with an ensemble approach, where we integrate textual semantics, visual understanding, clustering-based discovered patterns along with human-readable explanations that lead to better accuracy and transparency of malicious content detection.

To evaluate its effectiveness, experiments were conducted using the Facebook Hateful Memes(1), MUStARD(2) and MemeCap(3) datasets. These datasets capture different appearances of multi modal messages including hate speech, sarcasm, meme interpretation and unspoken harmful intent. Through comprehensive evaluation, we have investigated whether multi modal reasoning and explainable AI can really provide a reliable basis in complex online environments.

1.2. Research Objectives

The objectives of this research are:

RO1: Develop a framework that can analyse both textual and visual language for malicious content detection.

RO2: Improve the detection of misinformation, harmful memes, multi modal hate speech and image - grounded misinformation.

RO3: Evaluate cross-modal and multi modal strength with different benchmark datasets and content types.

RO4: Generation of clear and understandable explanations along with model decisions.

1.3. Contributions

This research addresses several limitations of existing malicious content detection approaches.

The key contributions are:

- A multi modal malicious content detection framework SEMTEX++, is developed to analyse textual and visual language within a unified architecture. Conventional moderation systems predominantly depend on textual embeddings, but our framework supports posts coming from cross modals to capture harmful intent.
- A semantic structure identification approach is adopted to identify harmful communication that may not be evident from class labels alone.
- The framework combines machine learning models with vision-language reasoning models through a hybrid ensemble strategy, which enables both classification and contextual understanding of content such as memes, sarcasm and implicit hate speech.

- To help end users to have transparent, readable explanations of model decisions, an explainable layer is introduced that integrates feature attribution methods with generative rationale generation.
- A new metric called Multimodal Threat Index (MTI) is created to measure the severity of possibly harmful content. Along with binary classification, risk assessment information is also provided so that the analyzing team can prioritize which messages need immediate attention.
- Three benchmark datasets Facebook Hateful Memes(1), MUStARD(2), and MemeCap(3) are experimented considering multiple multi modal moderation scenarios.

2. Literature Review

2.1 Text-Based Malicious Content Detection

In the past, researchers used traditional machine learning methods like SVM, Logistic Regression, Naïve Bayes and Random Forest mostly to detect malicious content(12,13,25,36). These techniques mostly work based on n-grams, lexical analysis, and syntactic patterns. Although these approaches were good at identifying explicit hate speech and offensive language, they had some major drawbacks like heavy training, rich data set requirements.

Over the period, there has been an improvement from ML to deep learning models like CNNs, RNNs and LSTMs (15,26,27,28,29) which are good at contextual understanding to understand semantic representations from text. More recently, LLM-based techniques were leveraged to identify harmful content by capturing contextual and multilingual patterns(14).

Our analysis is that text-based methods perform well for plain harmful content but struggle with context-dependent and visually grounded communication.

2.2 Multimodal Harmful Content Detection

The growth of memes, screenshots, and multimedia posts has increased interest in multimodal harmful-content detection(12,21,22,23). Harmful meaning in memes comes from relationships between visual and textual elements.

Benchmark datasets such as Facebook Hateful Memes(1), MUStARD(2), and MemeCap(3) would be leveraged as datasets to conduct investigations. Harmful multimodal content is difficult to detect because the interactions through hatefulness, sarcasm or misinformation inclusion are not plain in messages. Many models experience reduced performance when content comes from different platforms, languages and cultures.

Our analysis shows that multimodal architectures outperform text-alone approaches but mostly work on classification metrics with less explainability.

2.3 Explainable AI for Content Moderation

Stakeholders not only require results but also often require clear justifications or explanations for why content has been classified as harmful.

To achieve explainability, methods such as LIME(16) and SHAP(17) have been utilized, which provide influential textual or visual features. Recently, large language models have been used to generate natural-language rationales that are easier to understand(23, 24).

What is missing with these feature-attribution methods is contextual reasoning, and generated explanations miss evaluation mechanisms.

2.4 Research Gap and Emerging Challenges

The literature study shows progress in in-text-based moderation, learning that combines different types of data, explainable AI, and understanding how vision and language work together. Despite this progress, there are still some gaps that need to be filled.

- Existing models mostly prioritize accuracy of classification over semantic structures.
- Harmful memes, sarcasm and implicit hate speech remain a challenge in current multimodal systems.
- Explainability is frequently considered as a post-step after classification instead of an internal part of decision-making.
- Validation of rationale through faithfulness or completeness is not mandatory as part of risk-sensitive assessment.
- Cross-platform messages and multilingual posts are not completely covered in the detection process.

These challenges motivated us to think and work to get a unified framework that combines multimodal understanding, semantic discovery, explainable reasoning and threat-aware assessment.

2.5. Comparative Analysis of Existing Approaches

The table summarizes the tools and techniques along with the feasibility to identify malicious detection in textual, image and multilingual messages.

Approach	Text	Image	Multilingual	Explainability	Rationale Generation
SVM / Logistic Regression	✓	✗	Limited	partial	✗
CNN-Based Models	✓	Partial	Limited	Low	✗
BERT / RoBERTa / DeBERTa	✓	✗	Moderate	Limited	✗
CLIP / VisualBERT	✓	✓	Moderate	Limited	✗
BLIP / LLaVA	✓	✓	Moderate	Partial	✓
SEMTEX++ (Proposed)	✓	✓	High	High	✓

Table 1.Comparision analysis of existing approaches to detect malicious content

Based on the literature review, existing approaches address only subsets of the challenges associated with multimodal malicious content detection. Based on this observation, SEMTEX++ was developed, which integrates multiple techniques.

3. Research Methodology

SEMTEX++ is designed to simultaneously execute multimodal harmful content classification, semantic community finding, threat severity and rationale generation.

3.1 Problem Formulation

Malicious content detection in social networks has grown into a multimodal learning problem in which harmful intent is not only present in text but also in textual and visual content information. Unlike regular text content, the meaning of a social media post now comes from combinational text + image content. We can represent this combined multimodal dataset as

$$D = \{(T_i, V_i, Y_i)\}_{(i=1)\hat{}N}$$

Where, T_i denotes textual content, V_i denotes visual content, Y_i denotes the corresponding class label, and N represents the total number of samples.

The objective is to learn a multimodal mapping function: $F: (T, V) \rightarrow Y$, Where, $Y \in$ {Benign, Hate, Offensive, Threat, Misinformation}

3.2 Framework Overview

The Algorithm of the proposed SEMTEX++ framework covers five interconnected layers/ steps.

Step1: Collecting all sorts of content from social media posts in multiple forms. This includes collecting text, images and even text that's been extracted from images using OCR extraction along with metadata that might be useful.

Step2: Converting the textual and visual information into semantic embeddings using text encoders like Sentence BERT (32) and vision-language encoders CLIP(20).

Step3: Constructing joint semantic embeddings using projection layers and cross-modal interaction algorithms CLIP(20).

Step4: Discovering semantic structures using UMAP(39) and HDBSCAN(40) dimensionality reduction before doing classification of latent semantic structures through clustering and ensemble prediction.

Step5: Natural language explanations along with predictions are listed with SHAP(17) features, with MTI scores.

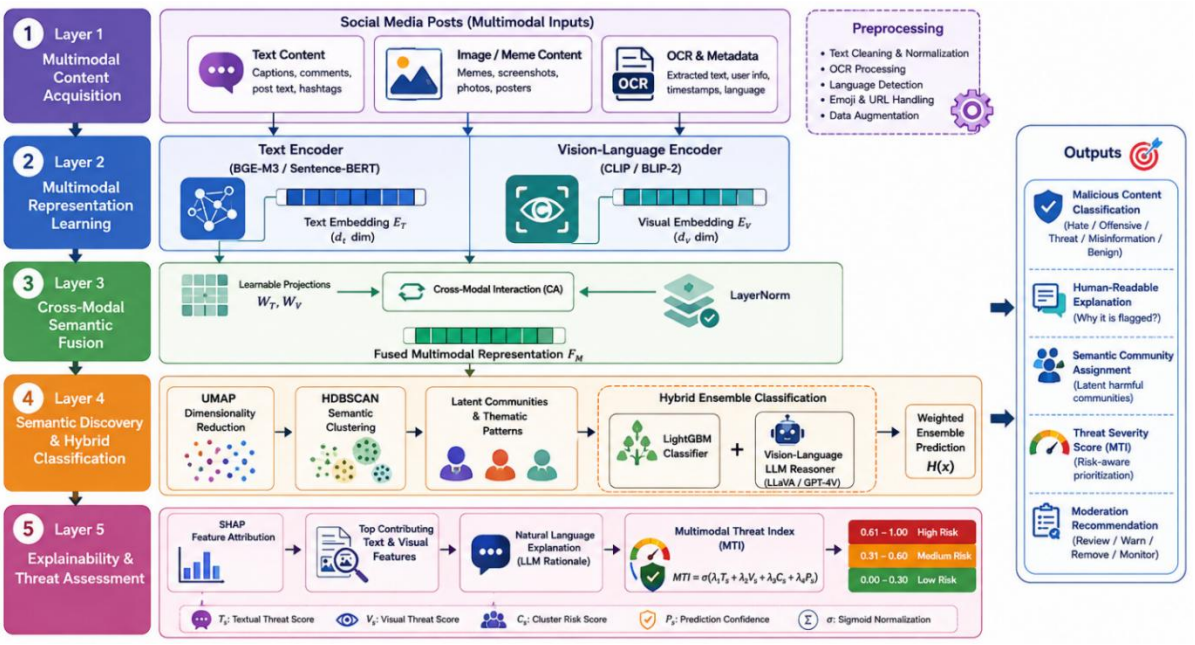

Figure 1: SEMTEX++ Architecture diagram

3.3 Multimodal Dataset Collection

For experimentation, we have gathered three publicly available benchmark datasets: Facebook Hateful Memes(1), Multimodal Sarcasm Detection Dataset - MUStARD (2) and MemeCap (3, 18), which represent different forms of multimodal malicious and complex social media content. These datasets were selected because they capture hate speech, multimodal sarcasm, meme communications, visual content and indirect harmful intent.

Data Set	Content Type	Samples	Labels
Facebook Hateful Memes	Meme images +text	5000	Hateful/ Non Hateful
MUStARD	Video frames + subtitles +context	690	Sarcastic/ Non Sarcastic
MemeCap	Memes + captions + explanations	6300	Multiple semantic annotations

Table 2: Total of 11990 samples considered for the research

Synthesized Visual language example

Due to copy rights of data sets, we are not showing any original visual language pictures in the paper, but to just give the readers a view of what it looks like below examples are shown. As an example1, the visual language with

sarcasm, and another example2 as a visual meme with “cheerful caption over a visibly miserable scene” is shown below.

3.3.1 Example Dataset Instances

Table 3. consists of a samples from each data set mentioned in Table 2, with four columns: data set, text, task and key challenge mentioned. Before feature extraction, pre-processing tasks were performed for both texts and images. Tesseract(41) Optical Character Recognition(OCR) was applied to extract embedded text content from images. This is merged with the original text content.

Data Set	Text	Task	Key challenges
Facebook Hateful Memes	ID 82403 – "everybody loves chocolate chip cookies, even hitler" (Label: Non-hateful)	Hate speech detection	joint image–text reasoning; text alone might be misleading
MUStARD	ID 1_60 – "It's just a privilege to watch your mind at work."	Sarcasm detection	The right meaning also depends on the conversation and who is speaking.
MemeCap	Post ID f49x1x – "Ayy sexy lady"	Meme understanding	interpreting visual metaphors and implicit meaning .

Table 3. Multimodal Dataset Tasks and Challenges Table

3.4 Semantic Embedding & Representation

The processed text is transformed into semantic vector representations with a multilingual foundation model called Sentence BERT (32).

For a textual input T_i, the embedding representation is defined as, $E_T = (f_T)T_i$

Where, $E_T \in R\,\hat{}(d_t)$ and f_T represents the textual encoder.

These embeddings capture semantic meanings, contextual relationships, linguistic patterns and sentiments, which might not be direct; for example, sarcastic messages look positive and the actual hidden meaning is negative.

3.5 Visual Representation Learning

Visual content usually convey meaning through symbols, gestures, contextual objects or the way picture described. These elements can add hidden context that makes harmful intent clean. Each image undergoes steps like resizing, normalization, OCR extraction and augmentation. Visual features are extracted using a vision-language encoder called as CLIP (20).

The visual representation is defined as:

$E_V = (f_V)V_i$

where: $E_V \in R\hat{}(d_V)$

The resulting representations capture details about objects, the overall scene, hidden meanings in visuals and unique patterns that usually found in memes.

3.6 Cross-Modal Fusion Strategy

The framework combines features from both text and images with meanings matched, so they fit together in one shared representation. This fused representation is then calculated as below,

$$F_M = LayerNorm(W_T E_T + W_V E_V \ + CA(E_T, E_V))$$

where: W_T is the textual projection matrix, W_V is the visual projection matrix and CA(42) is the cross-modal interaction function. CA aligns text and image representations through attention-based semantic matching to capture conflicting relationships between modalities. The purpose of including LayerNorm is to improve representation stability.

This setup helps the model to pick up on how text and images work together, how their meanings match up and how they depend on one another. With this it becomes easier to spot harmful intent that comes from how words and visuals interact.

3.7 Semantic Structure Discovery

Traditional supervised classifiers depend mostly on predefined labels and might fail to recognize unseen and new malicious patterns. To address this problem, SEMTEX++ incorporates semantic structure discovery. We are performing dimensionality reduction and density-based clustering here.

Dimensionality Reduction: Unified embeddings are projected into a lower-dimensional matrix using the function

$$\text{UMAP: } Z = UMAP(F_M) \text{ where: } Z \in R\hat{}(d_Z)$$

Density-Based Clustering: The reduced embeddings obtained from reducing dimensions are clustered using the HDBSCAN model,

$$C = HDBSCAN(Z)$$

Through these clusters, we can find hidden hate communities, misinformation patterns in messages, coordinated messaging behaviour and meme themes. Based on the results class labels along with additional meanings are assigned. This step helps later in identifying harmful content accurately.

3.8 Hybrid Ensemble Classification

Rather than depending on a single predictive model, SEMTEX++ combines discriminative and generative reasoning.

Classifier 1: Machine Learning Predictor

A LightBGM(33) classifier learns structured decision boundaries from multimodal representations.

$$P_1 = h_1(F_M)$$

Classifier 2: Vision-Language Reasoner

A multimodal LLM CLIP performs context-aware reasoning:

$$P_2 = h_2(T, V)$$

Ensemble Prediction: Final prediction is computed through weighted aggregation of classifier 1 and classifier 2.

$$H(x) = \lambda_1 P_1 + \lambda_2 P_2 \quad , \qquad \text{subject to} \quad \lambda_1 + \lambda_2 = 1$$

The reason for combining LightBGM (33) and CLIP is that the former efficiently detects numerical features and the latter technique provides high-level semantic understanding, especially when the data is ambiguous and context-dependent.

3.9 Explainable Decision Layer

This layer provides interpretations both locally(factors influencing individual predictions) and globally(feature importance across the entire data set). We generate explanations using SHAP(17), which shows which

features influenced the decision, and a generative AI model that can put those findings into plain natural language. This way, the system offers both numbers for transparency and easy-to-read explanations for people.

Feature Attribution : SHAP values estimate the contribution of each feature as,

$$\emptyset_i = SHAP(x_i)$$

Where, $\emptyset_i$ represents the importance of feature i.

Rationale Generation: The multimodal LLM generates a natural-language explanation based on the features obtained.

$$R = g(T, V, H(x))$$

Where, R denotes the generated rationale and g(·)represents the explanation model.

3.10 Multimodal Threat Index (MTI)

MTI is a combined risk score that calculates proof from text threats, image threats, risk patterns found in clusters and model confidence. This score helps moderation systems decide which content to review first by ranking posts according to how severe the threat is estimated to be.

$$MTI = \sigma(\lambda_1 T_s + \lambda_2 V_s + \lambda_3 C_s + \lambda_4 P_s)$$

where,T_s= textual threat score, V_s= visual threat score, C_s= cluster risk score, P_s= prediction confidence, σ is sigmoid normalization.

The resulting value satisfies $0 \leq MTI \leq 1$; risk interpretation is identified with the ranges below:

MTI Score	Threat Level
0.00-0.30	Low Risk
0.31-0.60	Medium Risk
0.61-1.00	High Risk

4. Experimental Design

4.1 Experimental Setup

The SEMTEX++ framework was implemented in Python 3.12, PyTorch 2.2 and the Hugging Face transformers library. The system configuration used is an NVIDIA GPU, Intel Core i9 processor and 64 GB RAM.

Visual representations were extracted using CLIP(20) vision-language model and textual features were generated using Sentence BERT model (32).

A stratified 5-fold cross-validation strategy (30) was adopted to ensure good evaluation.

Hyper parameter tuning was performed using grid search optimization on the training folds. We have repeated the experiment five times with different random seeds. The final result is the mean of all runs. To maintain robustness, multimodal content samples were included from datasets mentioned in Section 3.3.

The table indicates hyper parameter details used for experimentation,

Parameter	Text Encoder	Vision Encoder	Classifier	Batch size	Learning Rate	Epochs	Random seeds	Cross validation
Value	Sentence-BERT	CLIP ViT-B/32	LightBGM	32	2e-5	10	5	5-fold stratified

4.2 Statistical Validation Protocol

To figure out if the differences in performance were real, we used tests to see if they were statistically significant with estimated confidence intervals. We ran each experiment five times independently using different random seeds and then looked at the average performance.

Model stability was assessed through 95% confidence intervals CI:

$$CI = \overline{X} \pm 1.96 \times \left({}^{s}/_{\sqrt{n}}\right)$$

where $\overline{X}$ is the sample mean, s is the standard deviation, and n is the number of experimental runs.

To determine whether differences between models, paired t-tests were conducted between SEMTEX++ framework and baseline models. The significance threshold was set to α=0.05. Effect sizes were measured using Cohen's d metric to measure the size of the performance gains beyond statistical significance.

$d = \frac{M_1 - M_2}{SD}$ where d is performance gain and M_1 is one message1 and M_2 is message2.

All the experimental results are discussed in Section 5 of the paper.

5. Evaluation and Results

5.1 Overall Performance Evaluation

SEMTEX++ was evaluated using accuracy, precision, recall, and F1-score and ROC-AUC (31) metrics on all bench mark datasets. Results had proved that out approach is better than uni modal and multi modal baseline approaches in all runs with minimal variance. This shows that combining different types of data, finding semantic structure and using a mix of reasoning methods not only makes predictions better but also makes the model more stable. All experiments were conducted using a 5-fold cross-validation strategy (30). During each iteration, four folds were used for training and one fold for testing.

Figure 5. presents the average accuracy, precision, recall, and F1-score with mean ± standard deviation.

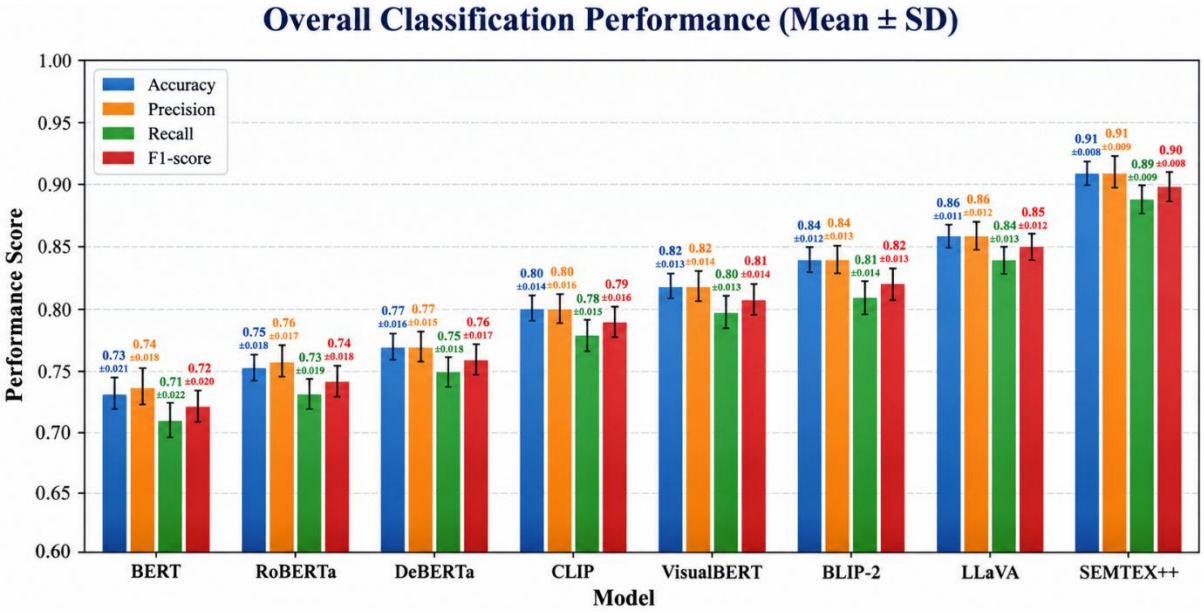

Figure 5. Overall performance comparison of SEMTEX++

5.2 Confidence Interval Analysis

To evaluate the reliability of the reported performance, 95% confidence intervals were calculated for F1-score (31) obtained across repeated experiments. The narrow confidence interval for SEMTEX++ shows that its performance is consistent across different runs.

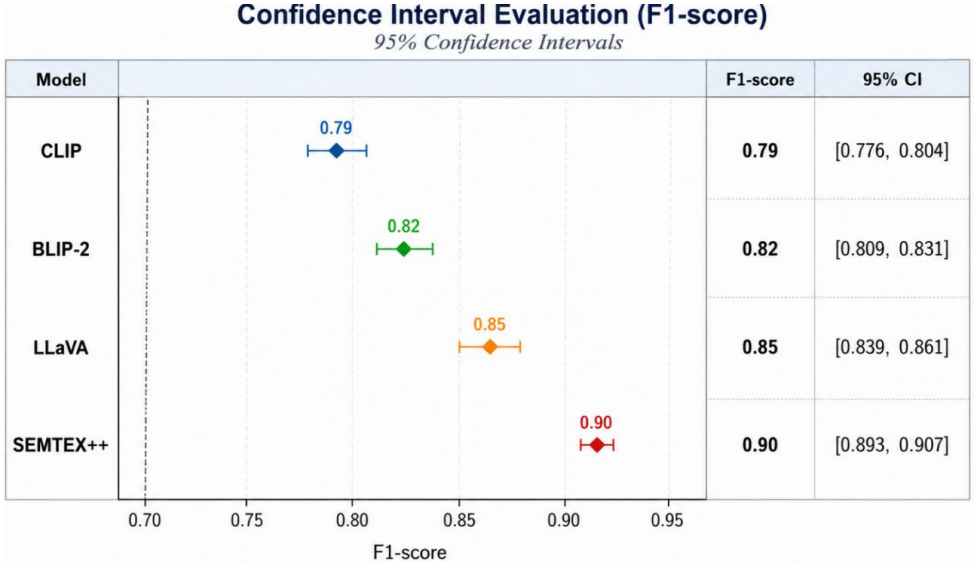

Figure 6. Mean F1-score and corresponsing 95% Confidence Intervals for all evaluated models

5.3 Cross Dataset Generalisation

To examine robustness across different multimodal contexts, we tested it on three datasets: Facebook Hateful Memes(1), MUStARD(2), and MemeCap(3). We found that our framework maintained strong performance across all three datasets despite of differences in characteristics of content and annotations. This shows that the framework was able to capture semantic patterns well instead of remembering the patterns. This suggests that our approach is learning things that can be applied to many different situations, not just specific characteristics of one datasets.

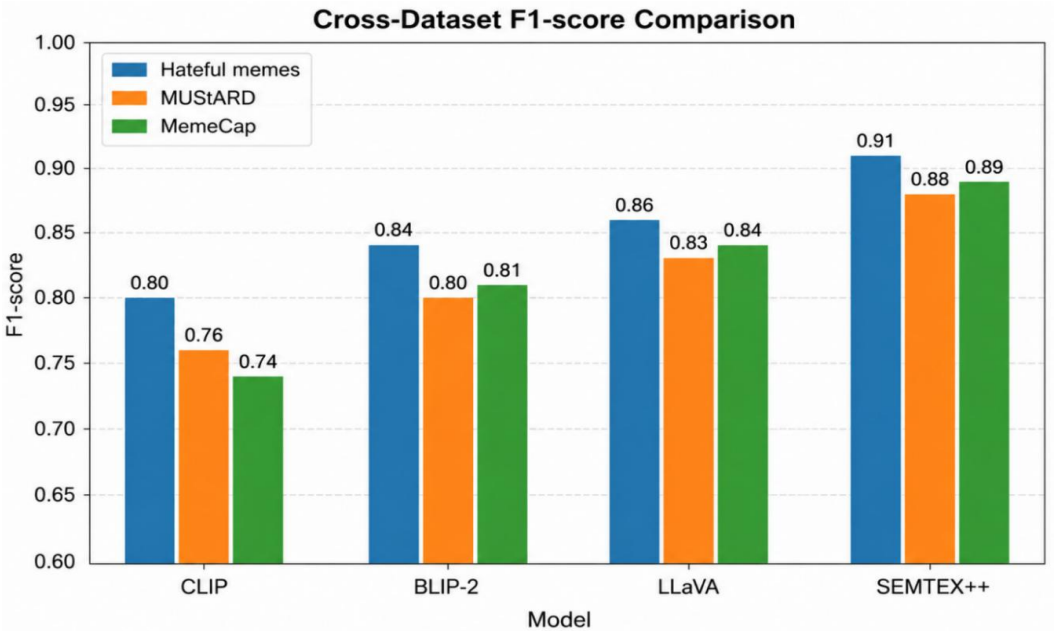

Figure 7. Grouped bar chart showing F1-score obtained across three benchmark datasets

5.4 ROC-AUC Evaluation

We have conducted Receiver Operating Characteristic (ROC) (31) analysis to check how well the model can tell harmful content apart at different decision thresholds. The Area Under the Curve (AUC) (31) was used as a threshold-independent performance indicator.

As shown in the figure, SEMTEX++ consistently achieved higher TP rates while keeping lower FP rates compared to the baseline methods.

Model	AUC	95% CI
CLIP	0.85	[0.84, 0.87]
BLIP-2	0.88	[0.87, 0.89]
LLaVA	0.91	[0.90, 0.92]
SEMTEX++	0.95	[0.94, 0.96]

Table 5. ROC-AUC Performance

5.5 Ablation Analysis

To measure the contribution of individual components, an ablation study was performed by removing major modules one by one from SEMTEX++ architecture.

As shown in the table below there is a performance downgrade when core component was excluded. The removal of generative reasoning decreased F1-score from 0.90 to 0.85, and excluding semantic clustering reduced the score to 0.86, which highlights the importance of semantic structure discovery.

Configuration	Accuracy	F1-score
Full SEMTEX++	0.91	0.90
Without OCR	0.88	0.87
Without Semantic Clustering	0.87	0.86
Without Explainability Layer	0.89	0.88
Without Generative Reasoning	0.86	0.85
Text Only	0.75	0.74

Table 6. Ablation Analysis

OCR module also contributed positively to overall performance, showing how important textual information extraction is when it is embedded within images.

5.6 Explainability Evaluation

The quality of generated explanations was measured using faithfulness, consistency and human evaluation scores. Faithfulness measures how much the resultant explanations reflect model behaviour, consistency measures explanation stability across similar inputs. Human evaluation scores range from 1 to 5, measuring clarity and correctness.

Method	Faithfulness	Consistency	Human Evaluation
LIME	0.72	0.70	3.5
SHAP	0.81	0.79	4.0
GPT-Based Rationale	0.86	0.84	4.3
Hybrid Explanation	0.91	0.89	4.7

Table 7. Explainability Assessment

5.7 Multi modal Threat Index (MTI) Evaluation

To validate the effectiveness of the proposed threat-ranking method, MTI scores were compared against experts' indicated labels. The table captures the distribution of MTI Scores across risk categories.

Risk Level	MTI Range	Human Agreement
Low	0.00-0.30	89%
Medium	0.31-0.60	87%
High	0.61-1.00	93%

Table 8. MTI Validation Results

The strong agreement between MTI predictions and expert assessments indicates that the proposed index can be used as a practical tool for moderation prioritization. The highest human agreement is seen in the high-risk category, then low risk and finally medium risk.

5.8 Statistical Significance and Effect Size Analysis

To determine whether performance improvements were statistically meaningful, we have conducted paired t-test using F1-scores from our experiments. The analysis indicated statistically significant improvements, explaining not only significance but also meaningfulness.

With 5-fold-cross-validation (30), confidence intervals are computed with 11990 samples; the CI interval is obtained as $0.11 \pm 1.96 \times 0.004 = 0.11 \pm 0.0078 = [0.102, 0.118]$, $\Delta F1 = 0.11$ and $SD = 0.004$. where $\triangle$ F1 indicates the difference in F1-score, and SD is the standard deviation.

The improvements we have seen are statistically reliable at the 95% confidence level. The effect sizes show that these gains represent practical improvements compared to existing multimodal baseline models.

5.9 Error Analysis

To better understand the limitations of the framework, we have manually checked misclassified samples. Most errors happened when the data needed outside cultural knowledge or having new online slang or has tricky sarcasm.

Error Category	Proportion (%)
Cultural References	31
Emerging Slang	24
Ambiguous Sarcasm	19
Poor Image Quality	15
OCR Errors	11

Table X. Major Sources of Misclassification

These results show that future research should focus on models that can keep learning over time.

6. Discussion

The experimental results demonstrate that SEMTEX++ outperforms text-only, vision-language and LLM baselines across classification, cross dataset generalisation, discrimination strength and quality of explanations. The experiments indicate that malicious content is not only conveyed through interactions between modalities but also through text or images alone.

The ablation study shows that both semantic clustering and generative reasoning plays an important role in getting best results. In addition, the explanability layer makes interpretable predictions which helps in improving transparency and trust.

The framework achieved an accuracy of 0.91, an F1-score of 0.90, and an AUC of 0.95. These findings suggest that combining visual-textual fusion with explainable and generative reasoning mechanisms would well detect multimodal malicious content in social platforms.

6.1. Limitations

Despite promising results, several limitations should be acknowledged.

1. The framework focuses on image-text content and does not deal with content from videos or live streams.
2. The inclusion of vision-language and generative AI models increases computational cost and may affect real-time deployment efficiency.
3. Content or inputs that mix multiple languages are still not well represented in the evaluation datasets. There is a need to include more code-mixed samples as a future step.
4. Benchmark datasets might not totally capture the latest memes, slang or how people communicate online, which is always changing and we can not control. This means we have a chance to make our framework better and see how well it really works.

6.2. Future work

As part of future work, we aim to research short-form videos, live streaming interactions, along with more visual - language patterns. Latest methodologies, including Agentic AI (34) workflows under Human in the loop (HITL), are interesting work to consider.

Future studies should also investigate continuous learning mechanisms so that the framework learns evolving online language and newly emerging memes.

7. Conclusion

Our framework SEMTEX++ is a multimodal framework that integrates multiple techniques to detect malicious content in text, visual-language, and multilingual social media datasets, which are borrowed from Facebook Hateful Memes (1), MUStARD (2), and MemeCap (3) datasets. The techniques that are integrated for finding results are semantic structure discovery, hybrid ensemble classification, explainable AI and threat-aware assessment in different layers of execution. Further checking performed with statistical tests, confidence intervals and ablation studies confirm that this framework is solid and dependable.

When we put together text and images, it really helps to spot nasty memes, sarcasm, false information and hate speech that's hidden between the lines and images. Also, using semantic clustering and explainable reasoning makes the decisions better and easier to understand, which is a big plus. This way, we can see how the system arrived at its conclusions, making it more transparent and trustworthy. By combining these approaches, a more powerful tool for detecting and dealing with harmful content can be used.

References

1. https://www.kaggle.com/datasets/parthplc/facebook-hateful-meme-dataset (web style)
2. https://www.kaggle.com/datasets/prashantpathak244/mustard-multimodal-sarcasm-detection-dataset (web style)
3. https://github.com/eujhwang/meme-cap/blob/main/data/memes-trainval.json (web style)
4. https://www.kaggle.com/datasets/prashantpathak244/mustard-multimodal-sarcasm-detection-dataset (web style)
5. Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, Kai-Wei Chang, “VisualBERT: A Simple and Performant Baseline for Vision and Language”, https://doi.org/10.48550/arXiv.1908.03557, 2019.(journal style)
6. Junnan Li, Dongxu Li, Silvio Savarese, Steven Hoi, “BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models”, Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.(journal style)
7. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger,Ilya Sutskever,”Learning Transferable Visual Models From Natural Language Supervision”, Proceedings of the 38 th International Conference on Machine Learning, PMLR 139, 2021.(conference style)
8. 7. Devlin J, Chang MW, Lee K, Toutanova K. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” Proc NAACL-HLT.4171-86. doi:10.18653/v1/N19-1423. 2019.(conference style)
9. 8. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. “RoBERTa: A Robustly Optimized BERT Pretraining Approach”. arXiv. 1907.11692.2019.(journal style)
10. 9. He P, Liu X, Gao J, Chen W. “DeBERTa: Decoding-enhanced BERT with Disentangled Attention”. arXiv:2006.03654v6 [cs.CL] 6 Oct 2021.(conference style)

11. 10. Piyush Verma, Manoj Kumar S, Harsh Deep, Kumar Chinmay, “Building Trust in AI: A Hybrid Approach to Combating Fake News and Misinformation”,IEEE Explorer, Dec 2025.(journal style)
12. 11.Yue Xue, “Using machine learning for toxic text detection on Android”,CNSSE '25: Proceedings of the 2025 5th International Conference on Computer Network Security and Software Engineering, Jun 2025.(conference style)
13. 12. Jiadi Liu, Zhuodong Liu, Qiaoqi Li, Weihao Kong, Xiangyu Li, “Multi-Domain Controversial Text Detection Based on a Machine Learning and Deep Learning Stacked Ensemble”,IEEE International Conference on Control & Automation, Electronics, Robotics, Internet of Things, and Artificial Intelligence (CERIA).Vol.13,Issue 9, 2024.(conference style)
14. 13. Jesslyn Tanuwijaya; Ivan Sebastian Edbert; Hans Christian Arinardi; Derwin Suhartono, “Offensive Text Detection on Social Media Using Machine Learning”,IEEE Explorer, Mar 2025.(journal style)
15. 14. Elyas Meguellati, Assaad Zeghina, Shazia Sadiq, Gianluca Demartini, “LLM-based Semantic Augmentation for Harmful Content Detection”, arXiv:2504.15548. 2025.(journal style)
16. 15. A Suresh; R M Mallika; S Gireesh; G Hari Kiran Singh; E S Jeevanandham; A Hemalatha, Empowering Fake News Detection Through Innovative Hybrid Deep Learning-Based Approach, “2025 International Conference on Computing for Sustainability and Intelligent Future (COMP-SIF)”, 2025.(conference style)
17. 16. Yanyun Wang, Xumei Fang, Zan Xu, Jianye Li, Luping Wang, “Exploring the Impact of Explainability in Large Language Model (LLM) Applications on User Experience”, CHI EA '25: Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, Article No.266, Pg.1 - 8,2025.(conference style)
18. 17. Rohini Jadhav, Vishal Meshram, Amol Bhosle, Kailas Patil, Sital Dash, Shrikant Jadhav, “Explainable multilingual and multimodal fake-news detection: toward robust and trustworthy AI for combating misinformation”, PMC, 2025.(journal style)
19. 18. EunJeong Hwang, Vered Shwartz, “MemeCap: A Dataset for Captioning and Interpreting Memes”, ACL Anthology, pg.1433–1445,2023.(journal style)
20. 19. Evaluating the Efficacy of Prompt-Engineered Large Multimodal Models versus Fine-Tuned Vision Transformers in Image-Based Security Applications “ACM Transactions on Intelligent Systems and Technology", Volume 16, Issue 4,Article No.: 89, Pages 1 - 22.2025.(journal style)
21. 20. Alec Radford, Ilya Sutskever, Jong Wook Kim, Gretchen Krueger, Sandhini Agarwal, “CLIP: Connecting text and images”, https://openai.com/index/clip/.2021.(web style)
22. 21. Hyewon Choi, Yejun Yoon, Seunghyun Yoon, Kunwoo Park, “How does fake news use a thumbnail? CLIP-based Multimodal Detection on the Unrepresentative News Image ”, Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations, Pg. 86 - 94. 2022.(journal style)
23. 22. Xinnong Zhang, Haoyu Kuang, Xinyi Mou,, Hanjia Lyu, , Kun Wu, Siming Chen,, Jiebo Luo, , Xuanjing Huang, , Zhongyu Wei1, SoMeLVLM: A Large Vision Language Model for Social Media Processing 2024", pages 2366–2389.2025.(journal style)
24. 23. Noushad Rahim M, Mohamed Basheer K.P,A hybrid explainability framework for recommender systems using SHAP with natural language interpretations, “International Journal of Innovative Research and Scientific Studies “, 8(6) , Pg.687-703,2025.(journal style)
25. 24. Manish Shukla,Interpreting BERT Using LIME and SHAP, DOI:10.21203/rs.3.rs-7376225/v1,2025.(journal style)
26. 25. Koushik Chowdhury, Spam Identification on Facebook, Twitter and Email using Machine Learning, “Central European Researchers Journal”, Vol.6 Issue 1,2020.(journal style)
27. 26. Dun Li, Haimei Guo , Zhenfei Wang, And Zhiyun Zheng, Unsupervised Fake News Detection Based on Autoencoder, “IEEE Access” ,Vol. 9, 2021.(journal style)
28. 27. E. Hashmi, S. Y. Yayilgan, M. M. Yamin, S. Ali and M. Abomhara, "Advancing Fake News Detection: Hybrid Deep Learning With FastText and Explainable AI," IEEE Access”, vol. 12, pp. 44462-44480, 2024(journal style)
29. 28. A. J. Keya, S. Afridi, A. S. Maria, S. S. Pinki, J. Ghosh and M. F. Mridha, "Fake News Detection Based on Deep Learning," 2021 International Conference on Science & Contemporary Technologies (ICSCT)”, , pp. 1-6, 2021(Conference style)
30. 29. O. Bashaddadh, N. Omar, M. Mohd and M. Nor Akmal Khalid, "Machine Learning and Deep Learning Approaches for Fake News Detection: A Systematic Review of Techniques, Challenges, and Advancements," IEEE Access”, vol. 13, pp. 90433-90466, 2025(journal style)
31. 30. Verma, V.K., Saxena, K., Banodha, U. (2024). Analysis Effect of K Values Used in K Fold Cross Validation for Enhancing the Performance of Machine Learning Model with Decision Tree. In: Garg, D., Rodrigues, J.J.P.C., Gupta, S.K., Cheng, X., Sarao, P., Patel, G.S. (eds) Advanced Computing. IACC 2023. Communications in Computer and Information Science, vol 2053. Springer, Cham. https://doi.org/10.1007/978-3-031-56700-1_30 (conference style)
32. 31. Oona Rainio, Jarmo Teuho, Riku Klén , Evaluation metrics and statistical tests for machine “www.nature.com/scientificreports”,2024. (journal style)
33. 32. Nils Reimers, Iryna Gurevych, " Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,"arXiv:1908.10084”, 2019. (conference style)
34. 33. Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree, “Advances in Neural Information Processing Systems 30 (NIPS 2017)”, 2017. (journal style)
35. 34. Juan Ren, Mark Dras, Usman Naseem. Agentic Moderation: Multi-Agent Design for Safer Vision-Language Models, “arxiv.org/pdf/2510.25179v1 [cs.AI] 29 oct 2025”. (journal style)

36. 35. Esma Aïmeur, Sabrine Amri, Gilles Brassard. Fake news, disinformation and misinformation in social media: a review, “PMC, Soc Netw Anal Min.” 2023 Feb 9;13(1):30. doi: 10.1007/s13278-023-01028-5 (journal style)
37. 36. Yazdi KM, Yazdi AM, Khodayi S, Hou J, Zhou W, Saedy S. Improving fake news detection using k-means and support vector machine approaches. “Int J Electron Commun Eng. 2020”14(2):38–42. doi: 10.5281/zenodo.3669287. (conference style)
38. 37. Priyanka Meel, Dinesh Kumar Vishwakarma. Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities, “Expert Systems with Applications”, Vol.153,,112986,ISSN 0957-4174,2020.(conference style)
39. 38. N.Deepika, N.Guruprasad. Malicious Content detection in social networks using hybrid machine learning model, “International Journal of Computer Science & Engineering”, e-ISSN:0976-5166, Volume 14 No.3, June 2023.(journal style)
40. 39. McInnes L, Healy J, Melville J. “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction”. arXiv;1802.03426.2020.(journal style)
41. 40. McInnes L, Healy J, Astels S. “hdbscan: Hierarchical density based clustering.” The Journal of Open Source Software .2(11):205. doi:10.21105/joss.00205.2017.(journal style)
42. 41. Santosh Kumar, Nilesh Kumar Sharma, Mridul Sharma, Nikita Agrawal. Text Extraction from Images Using Tesseract. “in Deep Learning Techniques for Automation and Industrial Applications“, Wiley, 2024, pp.1-18, doi: 10.1002/9781394234271.ch1.(journal style)
43. 42. Laura Wenderoth, Konstantin Hemker, Nikola Simidjievski, Mateja Jamnik1.Nikita Agrawal. Measuring Cross-Modal Interactions in Multimodal Models. arXiv:2412.15828v2 [cs.LG] 28 Jan 2025.(journal style)