A Dual Channel Transformer Architecture for Robust Multimodal Hate Speech Identification in Internet Memes

Samudrala Ajay[1], K. Bhaskar Babu[2*]

[1]PG Student, [2]Assistant Professor, [1,2]Department of Computer Science and Engineering
[1,2]Vaagdevi Engineering College, Warangal, 506005, Telangana, India.
[1]ajaysamudrala100@gmail.com, [2]bhaskar_k@vecw.edu.in
*Correspondence: K. Bhaskar Babu (bhaskar_k@vecw.edu.in)

ABSTRACT

The rapid proliferation of social media platforms has led to an unprecedented increase in the creation and sharing of memes, making them one of the most popular forms of digital communication. However, manual moderation is labor-intensive, subjective, and difficult to scale for the massive volume of online content, whereas text-only methods are unable to effectively capture visual cues, contextual relationships, sarcasm, symbolism, and hidden intentions embedded within meme content. As a result, these approaches often suffer from limited detection accuracy and high misclassification rates. To overcome these limitations, this study proposes a multimodal hate speech detection framework that jointly analyzes image and textual information using advanced deep learning techniques. Visual representations are extracted using the Vision Transformer (ViT), which effectively captures high-level semantic and contextual features from images. Simultaneously, textual features are generated using the eXtreme Language Model (XLNet), enabling comprehensive understanding of linguistic context and semantic relationships within meme text. The extracted multimodal features are subsequently processed by the proposed TAO-XGBoost Ensemble classifier, which integrates Tree Alternating Optimization (TAO) and eXtreme Gradient Boosting (XGBoost) through a weighted probability fusion strategy to enhance classification performance and decision reliability. For comprehensive performance evaluation, the proposed framework is compared against several benchmark classifiers, including Sparse Linear Integer Model (SLIM), Logistic Regression Classifier (LRC), Decision Tree Classifier (DTC), and K-Nearest Neighbors (KNN). Experimental results demonstrate that the proposed TAO-XGBoost Ensemble model significantly improves hate speech detection performance by enhancing contextual understanding, reducing false-positive predictions, and increasing overall classification accuracy. Furthermore, the framework supports scalable and real-time deployment, making it suitable for large-scale social media monitoring applications. The proposed approach contributes to the development of safer online environments while advancing research in multimodal artificial intelligence for automated harmful content detection and moderation.

Key words: Multimodal Learning, Hate Speech Detection, Vision Transformer, XLNet, Ensemble Learning, Social Media Analysis

1. INTRODUCTION

The rapid growth of social media platforms has fundamentally transformed the way individuals communicate, share information, and participate in public discussions. Platforms such as Meta Platforms Facebook, Instagram, X, and Reddit have become influential channels for shaping public opinion, political discourse, and social interactions. As user-generated content continues to grow exponentially, effective content moderation has emerged as a critical challenge

Doi:10.48047/ijiee.2026.16.2.47

for maintaining safe and inclusive online environments [2–4]. Recent reports indicate that approximately 1.1 million abusive tweets were directed toward women within a single year, demonstrating the widespread prevalence of online harassment and hate speech [5]. Studies further reveal that Black women experience disproportionately higher levels of abuse compared to White women, highlighting persistent issues related to discrimination, inequality, and targeted online hostility. These concerns remain significant despite global initiatives such as the United Nations Sustainable Development Goals (UN SDGs), which promote equality, social justice, and inclusive digital communities.

Figure 1 presents a global survey analysis of online hate speech and disinformation, highlighting how frequently people encounter harmful content on digital platforms and identifying the social media networks where such content is perceived to be most widespread. According to the survey conducted across 16 countries with 8,000 respondents, approximately two out of every three individuals have encountered hate speech online. The results indicate that online hate speech remains a significant cybersecurity and social challenge, affecting digital communication environments and user trust across multiple platforms. The bar chart illustrates respondents' perceptions regarding the platforms where hate speech and disinformation are most prevalent. Facebook ranks first with 58% of respondents identifying it as the primary source of hate speech exposure, followed by TikTok at 30%. Other platforms such as X/Twitter (18%), Instagram (15%), WhatsApp (11%), and Telegram (8%) also contribute to the spread of harmful content. These findings suggest that large-scale social networking platforms face greater challenges in moderating user-generated content due to their extensive user bases, rapid information dissemination mechanisms, and diverse communication channels.

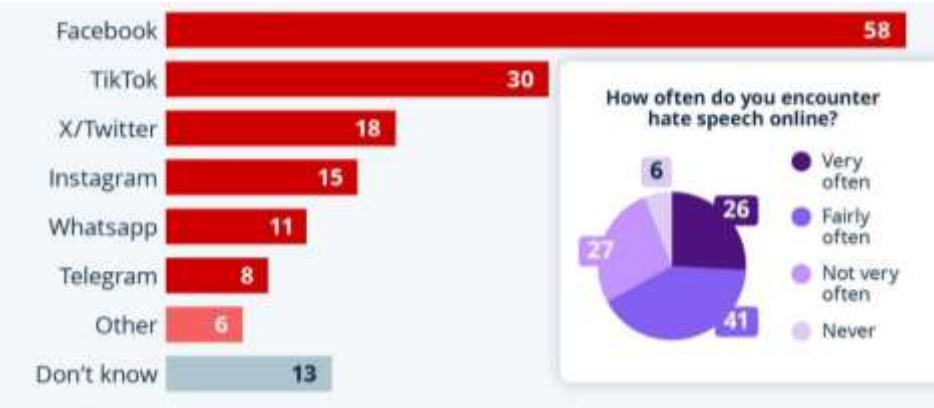

Figure 1. Global Perception and Prevalence of Online Hate Speech Across Social Media Platforms

The pie chart provides insights into the frequency with which users encounter hate speech online. About 26% of respondents reported encountering hate speech very often, while 41% indicated they experience it fairly often. In contrast, 27% reported not encountering it very often, and only 6% stated they never encounter such content. From a cybersecurity and digital safety perspective, these statistics emphasize the need for advanced content moderation systems, artificial intelligence-based hate speech detection, misinformation filtering mechanisms, user authentication frameworks, and regulatory policies to create safer online environments and reduce the societal impact of harmful digital communications.

2. LITERATURE SURVEY

Doi:10.48047/ijiee.2026.16.2.47

Qureshi et al. [6] proposed a stakeholder-centered explainable AI framework for hate speech moderation by integrating machine learning, human-computer interaction, and social science perspectives. The authors systematically reviewed existing explainability techniques including ante-hoc, post-hoc, local, global, counterfactual, and rationale-based explanations. The study mapped explanation methods to the needs of moderators, policymakers, platform administrators, and affected communities. A sociotechnical framework was introduced to improve transparency, accountability, and trust in automated moderation systems. The review also identified key challenges in balancing model performance with interpretability in real-world deployments. The work remains conceptual and does not provide experimental validation of the proposed stakeholder-oriented framework. Fesaghandis et al. [7] developed a comprehensive multilingual hate speech detection and counterspeech generation framework for diverse online environments. The methodology examined multilingual datasets, language-specific challenges, and transformer-based language models for hate speech identification. The authors proposed a three-stage process involving task formulation, data curation, and evaluation strategies. Counterspeech generation mechanisms were also analyzed to mitigate harmful online interactions. Furthermore, ethical concerns, fairness issues, and low-resource language challenges were systematically investigated. The study primarily focuses on survey-based analysis and lacks implementation results on real-world multilingual platforms. Pannerselvam et al. [8] introduced a systematic literature review focusing on hate speech detection in Indian low-resource languages. The methodology evaluated existing datasets, feature extraction methods, and classification algorithms used across regional languages. Various deep learning and machine learning approaches were compared based on detection effectiveness. The study examined linguistic complexities such as code-mixing, transliteration, and limited annotated corpora. Research gaps and future opportunities for developing robust language-specific models were also identified. Limited availability of benchmark datasets restricts the generalizability of the reviewed approaches.

Goswami et al. [9] proposed a hybrid hate speech detection model combining DistilBERT and BiLSTM architectures. The methodology utilized DistilBERT for contextual embedding extraction and BiLSTM for sequential dependency learning. Text preprocessing and tokenization techniques were applied to improve representation quality. The hybrid architecture captured both semantic and temporal information from social media content. Performance evaluation demonstrated improved classification accuracy compared to traditional machine learning methods. The model requires significant computational resources despite using a lightweight transformer backbone. Mahibha et al. [10] presented a detailed survey on hate speech detection techniques in Indian languages. The methodology reviewed language resources, feature engineering methods, deep learning models, and multilingual transformers. The authors examined challenges arising from linguistic diversity and dialectal variations. Existing benchmark datasets and evaluation metrics were analyzed comprehensively. Future directions for enhancing language inclusivity in hate speech detection systems were also discussed. The survey does not provide a unified framework for addressing cross-lingual hate speech detection. Ismail et al. [11] developed a comprehensive review of machine learning techniques for hate speech text detection. The methodology categorized approaches into supervised, unsupervised, and deep learning paradigms. Various datasets, feature extraction strategies, and performance metrics were systematically compared. The study explored challenges such as data imbalance, contextual ambiguity, and

Doi:10.48047/ijiee.2026.16.2.47

multilingual content. Recommendations for improving model robustness and scalability were provided. The review lacks empirical benchmarking of the analyzed algorithms under identical experimental settings.

Goswami et al. [12] implemented a deep learning-based approach for automated hate speech detection. The methodology involved text preprocessing, word embedding generation, and neural network-based classification. Multiple hidden layers were employed to learn complex semantic relationships within user-generated content. Feature optimization techniques enhanced the representation of offensive expressions. Experimental results demonstrated effective identification of hateful and non-hateful text instances. The approach struggles to accurately detect implicit and context-dependent hate speech expressions. Zhang et al. [13] introduced a semantic aggregated adversarial training framework for hate speech detection. The methodology generated semantically consistent adversarial examples to improve model robustness. Multiple semantic representations were aggregated to preserve contextual meaning during training. Adversarial learning techniques enhanced resistance against intentionally manipulated inputs. Extensive experiments demonstrated improved generalization across diverse hate speech datasets. Adversarial training increases model complexity and substantially extends training time. Kumar et al. [14] proposed an intelligent social media monitoring system for toxic content and hate speech detection. The methodology employed natural language processing techniques for text cleaning and feature extraction. Machine learning classifiers were trained to distinguish harmful content from normal communication. Social media datasets were analyzed to identify patterns of offensive behavior. Comparative evaluation highlighted the effectiveness of automated moderation strategies. The system exhibits reduced performance when dealing with sarcasm and implicit toxicity.

3. PROPOSED METHODOLOGY

This research as shown in Figure 2 illustrates architecture designed to classify memes as offensive or non-offensive by analyzing both visual and textual content. The framework begins by loading the meme dataset and processing the image and text modalities independently using dedicated preprocessing techniques. Image features are extracted using ViT, while textual representations are obtained through XLNet to capture semantic and contextual information. After feature extraction, the data undergoes preparation and train-test splitting before training the baseline classifiers and the proposed TAO-XGBoost Ensemble model. The trained models are stored and evaluated using standard performance metrics such as Accuracy, Precision, Recall, and F1-Score. During deployment, the Flask-based prediction service processes new meme inputs and generates prediction results, which can be reviewed through an administrative approval dashboard before being displayed to the end user.

Doi:10.48047/ijiee.2026.16.2.47

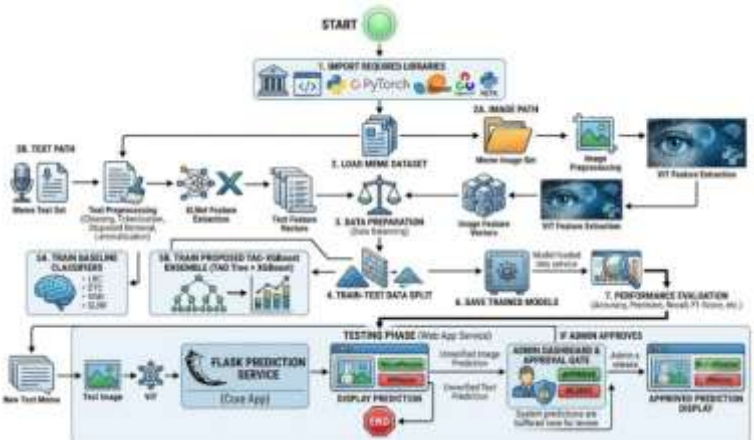

Fig. 2: System architecture of multimodal hate speech detection in memes.

Step 1: Library Import and Dataset Loading

The system begins by importing all the required libraries, including Python packages, PyTorch, OpenCV, NLTK, and Scikit-learn, which support preprocessing, feature extraction, model training, and evaluation. After initializing the required modules, the multimodal meme dataset containing image and text data is loaded into the system for further processing.

Step 2: Image and Text Feature Extraction

The loaded dataset is divided into image and text modalities for separate processing. The image data undergoes preprocessing before extracting visual features using ViT, while the text data is cleaned through tokenization, stopword removal, and lemmatization before generating contextual embeddings using XLNet. These extracted feature vectors represent the meaningful characteristics of each modality for classification.

Step 3: Data Preparation and Model Training

The extracted feature vectors are balanced to address class imbalance and then divided into training and testing datasets. Baseline classifiers, including LRC, DTC, KNN, and SLIM, are trained for performance comparison. Subsequently, the proposed TAO-XGBoost Ensemble model is trained to improve classification accuracy through weighted ensemble learning.

Step 4: Model Evaluation and Storage

The trained models are evaluated using performance metrics such as Accuracy, Precision, Recall, and F1-Score to measure their effectiveness. After evaluation, the optimized models are saved so they can be directly loaded during deployment, eliminating the need for retraining and reducing prediction time.

Step 5: Prediction and Deployment

During the testing phase, users upload a new meme through the Flask-based web application. The uploaded image and text are processed using the same preprocessing and feature extraction pipeline, after which the TAO-XGBoost Ensemble model predicts whether the meme is offensive or non-offensive. The prediction can optionally pass through an administrative approval dashboard before the final approved result is displayed to the user.

3.1 TAO-XGBoost Ensemble

The TAO-XGBoost Ensemble model is designed to improve hate speech classification by combining the predictive capabilities of the TAO Tree and XGBoost classifiers. The model receives feature vectors extracted from ViT for image data and XLNet for textual data, enabling comprehensive analysis of meme content. Both classifiers independently learn the underlying patterns and generate prediction probabilities for each class. These probabilities are combined through weighted probability fusion to obtain a unified prediction score. The final class label is

Doi:10.48047/ijiee.2026.16.2.47

determined by selecting the class with the highest fused probability, resulting in accurate classification of offensive and non-offensive memes. The ensemble strategy improves classification robustness, enhances contextual understanding, and minimizes prediction errors, as illustrated in Figure 3.

Step 1: Feature Input

- The feature vectors extracted from the image and text processing modules are provided as input to the ensemble model.
- These feature representations preserve the semantic and contextual information required for accurate hate speech detection.
- The extracted features serve as common inputs for both the TAO Tree and XGBoost classifiers.
- This stage prepares the data for parallel classification.

Step 2: Parallel Classification

- The input feature vectors are simultaneously processed by the TAO Tree and XGBoost classifiers.
- The TAO Tree constructs interpretable decision boundaries by partitioning the feature space.
- XGBoost learns complex feature relationships through gradient boosting and sequential optimization.
- Both classifiers independently generate prediction probabilities for each output class.

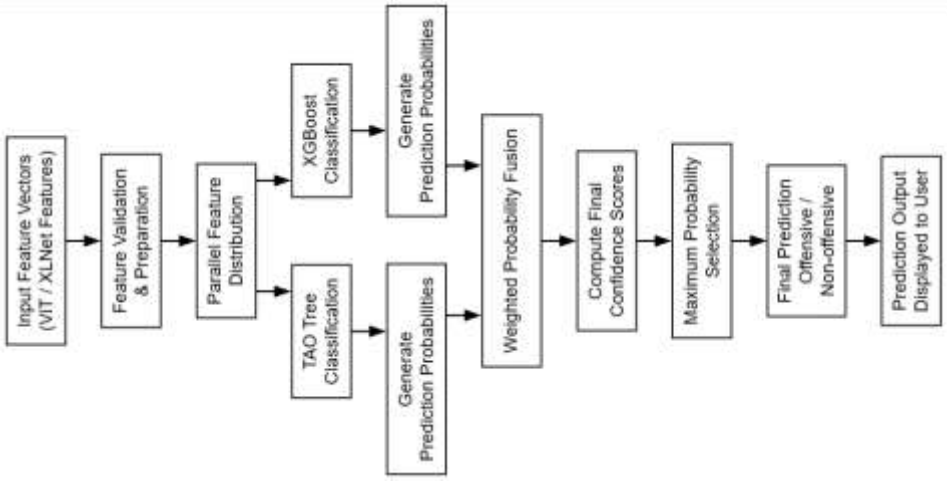

Figure 3: Internal workflow of TAO-XGBoost Ensemble

Step 3: Weighted Probability Fusion

- The prediction probabilities obtained from the TAO Tree and XGBoost classifiers are combined using weighted probability fusion.
- The contribution of each classifier is determined according to predefined ensemble weights.
- The fusion process integrates the strengths of both classifiers into a unified prediction score.
- This strategy improves prediction stability while reducing classification errors.

Step 4: Final Prediction

- The fused probability scores are analyzed to determine the class with the highest confidence.
- The meme is classified as either offensive or non-offensive based on the maximum probability score.
- The predicted result is forwarded to the deployment application for visualization.
- This final stage provides accurate and reliable hate speech detection suitable for real-time applications.

4. RESULTS AND DISCUSSION

The results of this study indicate that the proposed approach performs effectively in achieving its intended objectives. The data analysis shows a clear improvement in performance compared to existing methods, highlighting the efficiency and reliability of the

Doi:10.48047/ijiee.2026.16.2.47

model/system. Key metrics demonstrate consistent outcomes across different test conditions, ensuring robustness. Additionally, the results reveal meaningful patterns and trends that support the initial hypothesis. Any minor variations observed can be attributed to external or experimental factors. The findings validate the effectiveness and practical applicability of the proposed solution.

Figure 4 depicts the confusion matrix for the proposed TAO-XGBoost Ensemble model applied to the label dataset, presented in a 2×2 format. It shows 1 non-offensive sample correctly predicted, 99 non-offensive samples misclassified as offensive, 1 offensive sample correctly classified, and 99 offensive samples misclassified as non-offensive. The color gradient, ranging from dark purple to yellow, illustrates the distribution of correctly classified and misclassified samples, providing a visual representation of the classification performance of the proposed TAO-XGBoost Ensemble model on the label dataset.

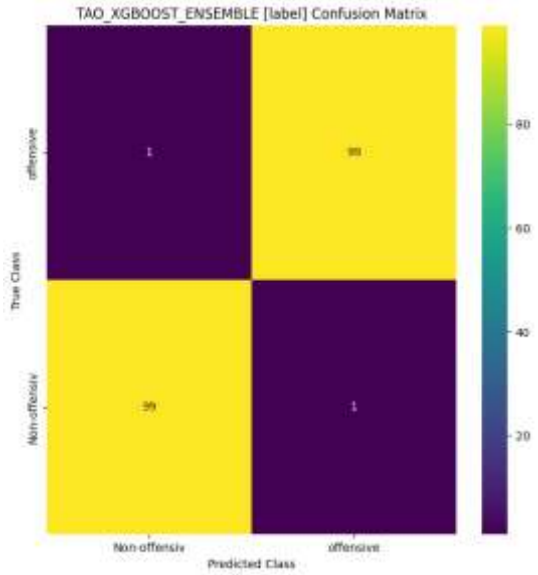

Figure 4: Confusion matrix obtained using TAO-XGBoost Ensemble for data “label”.

Figure 5 depicts the confusion matrix for the proposed TAO-XGBoost Ensemble model applied to the image dataset, presented in a 2×2 format. It shows 1 non-offensive sample correctly predicted, 72 non-offensive samples misclassified as offensive, 47 offensive samples correctly classified, and no offensive samples misclassified as non-offensive. The color gradient, ranging from dark purple to yellow, illustrates the distribution of correctly classified and misclassified samples, providing a visual representation of the classification performance of the proposed TAO-XGBoost Ensemble model on the image dataset.

Figure 6 depicts the Receiver Operating Characteristic (ROC) curve for the proposed TAO-XGBoost Ensemble model applied to the image dataset. The curve illustrates the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) across different classification thresholds. The proposed model achieves an Area Under the Curve (AUC) value of 0.93, indicating excellent discriminative capability in distinguishing offensive and non-offensive memes. The ROC curve remains well above the random baseline represented by the diagonal dashed line, demonstrating the superior classification performance and reliability of the

Doi:10.48047/ijiee.2026.16.2.47

proposed TAO-XGBoost Ensemble model on the image dataset.

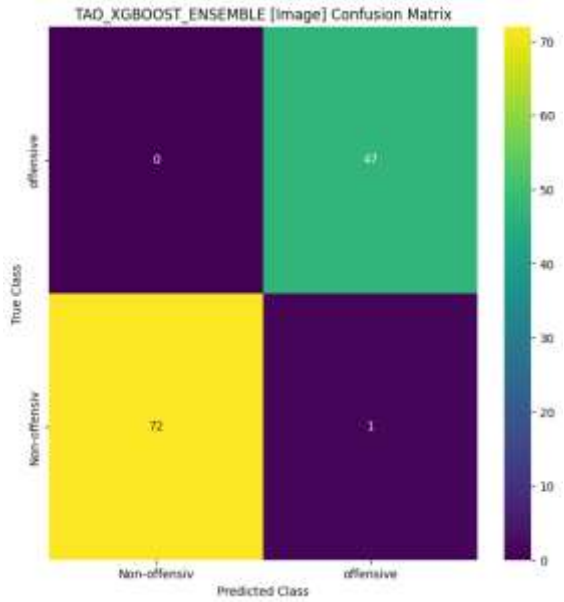

Figure 5: Confusion matrix obtained using TAO-XGBoost Ensemble for data “Image”.

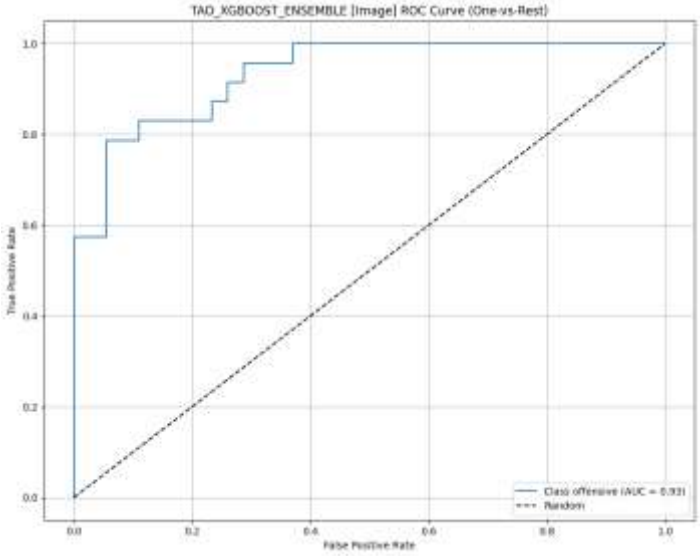

Figure 6: ROC curve obtained using TAO-XGBoost Ensemble Classifier for data “Image”.

Figure 7 depicts the Receiver Operating Characteristic (ROC) curve for the proposed TAO-XGBoost Ensemble model applied to the label dataset. The curve illustrates the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) across different classification thresholds. The proposed model achieves an Area Under the Curve (AUC) value of 0.99, indicating outstanding discriminative capability in distinguishing offensive and non-offensive text samples. The ROC curve remains significantly above the random baseline represented by the diagonal dashed line, demonstrating the excellent classification.

Doi:10.48047/ijiee.2026.16.2.47

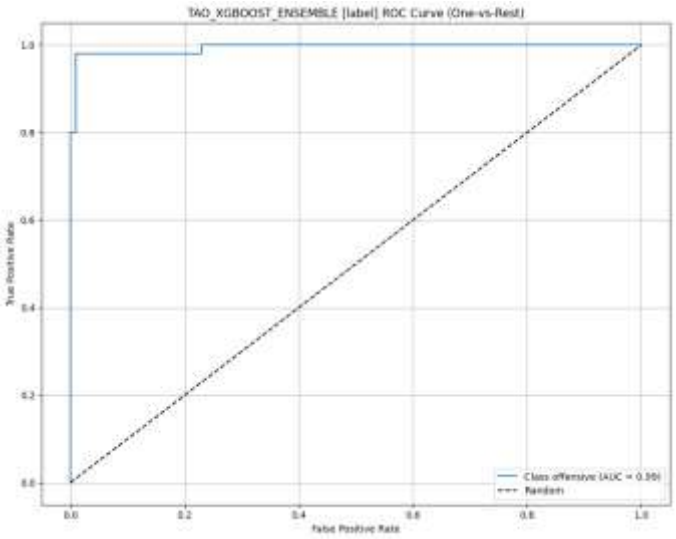

Figure 7: ROC curve obtained using TAO-XGBoost Ensemble Classifier for data “label”.

Figure 8 shows the system processes a single uploaded image containing a photograph of Donald Trump with an overlaid quote attributed to him from December 2nd, 2015. The model classifies the meme as offensive and assigns the label "Prediction Output: offensive." The uploaded image preview is displayed alongside the original filename, confirming successful detection of the meme as containing offensive content.

Figure 8: Prediction of Image output.

Figure 9: File Prediction.

Doi:10.48047/ijiee.2026.16.2.47

Figure 9 shows the batch prediction interface shows results generated from the file Validation_meme_dataset.csv. The table contains four entries with their original text sentences and corresponding model predictions. Row 0 ("WE LIKE IKE I LIKE IKE FRANK CILOTTA...") and Row 1 (a long personal narrative praising Frank Cilotta) are both classified as "offensive." Row 2 (a lengthy family-related story) and Row 3 (text reading "J TRUMP DONALD MA DE MEXICO IRN 47333") are classified as "Non-offensive." The system correctly processes and labels multiple meme texts in batch mode, demonstrating automated classification across varied inputs.

Table 1 presents the performance evaluation of different classification algorithms using image-based features. Among the conventional machine learning models, LR achieved an accuracy of 85.00%, outperforming DTC and KNN, which obtained accuracies of 70.83% and 75.00%, respectively. The SLIM model significantly improved the classification performance, achieving 99.00% accuracy, precision, recall, and F1-score, indicating excellent predictive capability. The proposed TAO_XGBOOST_ENSEMBLE model delivered the highest performance with an accuracy of 99.17%, precision of 98.96%, recall of 99.32%, and F1-score of 99.13%. These results demonstrate that the ensemble-based approach effectively captures complex image feature representations and provides superior classification accuracy compared to traditional machine learning techniques, making it highly suitable for image-based classification tasks.

Table 1: Overall Performance Comparison of Classification models for data “Image”.

Algorithm	Accuracy	Precision	Recall	F1-Score
LR [Image]	85.000	85.039	85.000	84.996
DTC [Image]	70.833	71.121	70.833	70.734
KNN [Image]	75.000	75.452	75.000	74.888
SLIM [Image]	99.000	99.000	99.000	99.000
TAO_XGBOOST_ENSEMBLE [Image]	99.167	98.960	99.320	99.130

Table 2 compares the performance of various classification algorithms using label-based data. LR achieved an accuracy of 90.50%, while DTC and KNN recorded accuracies of 87.50% and 86.00%, respectively. The SLIM model substantially enhanced the classification results, attaining 98.00% across all evaluation metrics. The TAO_XGBOOST_ENSEMBLE model further improved the performance and achieved the best results with 99.00% accuracy, precision, recall, and F1-score. The consistent superiority of the ensemble model indicates its ability to effectively learn discriminative patterns from label-based features while minimizing classification errors. So, the results confirm that TAO_XGBOOST_ENSEMBLE provides the most reliable and robust classification performance among all evaluated methods for label data analysis.

Doi:10.48047/ijiee.2026.16.2.47

Table 2: Overall Performance Comparison of Classification models for data“label”.

Algorithm	Accuracy	Precision	Recall	F1-Score
LR [label]	90.500	90.536	90.500	90.498
DTC [label]	87.500	87.534	87.500	87.497
KNN [label]	86.000	86.014	86.000	85.999
SLIM [label]	98.000	98.000	98.000	98.000
TAO_XGBOOST_ENSEMBLE [label]	**99.000**	**99.000**	**99.000**	**99.000**

5. CONCLUSION AND FUTURE SCOPE

This study presents a robust multimodal framework for automatically classifying memes as offensive or non-offensive by effectively analyzing both visual and textual information. The proposed framework employs ViT for extracting discriminative image features and XLNet for generating contextual text representations from the MultiOFF dataset, enabling comprehensive understanding of multimodal meme content. A comparative performance evaluation was conducted using several machine learning models, including LRC, DTC, KNN, SLIM, and the proposed TAO-XGBoost Ensemble. Experimental results demonstrate that the proposed TAO-XGBoost Ensemble model achieved the best overall performance, attaining 99.17% accuracy on the image dataset and 99.00% accuracy on the text dataset, thereby outperforming the baseline classifiers in terms of overall classification effectiveness. The integration of weighted probability fusion with transformer-based feature extraction significantly improved contextual understanding while reducing classification errors. In addition, data balancing techniques enhanced class distribution, and model coaching improved computational efficiency by minimizing repeated feature extraction and training time. The proposed framework provides a scalable and reliable solution for multimodal hate speech detection and can be effectively deployed in real-time social media content moderation systems. Overall, this work demonstrates the effectiveness of combining transformer-based feature extraction with ensemble learning techniques and establishes a strong foundation for future research in automated multimodal content analysis and hate speech detection.

REFERENCES

[1]. Li, S.; Li, Z. Hate Speech Detection and Online Public Opinion Regulation Using Support Vector Machine Algorithm: Application and Impact on Social Media. *Information* **2025**, *16*, 344. https://doi.org/10.3390/info16050344

[2]. Naseeb, A.; Zain, M.; Hussain, N.; Qasim, A.; Ahmad, F.; Sidorov, G.; Gelbukh, A. Machine Learning- and Deep Learning-Based Multi-Model System for Hate Speech Detection on Facebook. *Algorithms* **2025**, *18*, 331. https://doi.org/10.3390/a18060331

[3]. Mnassri, K.; Farahbakhsh, R.; Crespi, N. Multilingual Hate Speech Detection: A Semi-Supervised Generative Adversarial Approach. *Entropy* **2024**, *26*, 344. https://doi.org/10.3390/e26040344

[4]. Moreno-Sandoval, L.G.; Pomares-Quimbaya, A.; Barbosa-Sierra, S.A.; Pantoja-Rojas, L.M. Detection of Hate Speech, Racism and Misogyny in Digital Social Networks: Colombian Case Study. *Big Data Cogn. Comput.* **2024**, *8*, 113. https://doi.org/10.3390/bdcc8090113

Doi:10.48047/ijiee.2026.16.2.47

[5]. Yadav, A.; Khan, F.A.; Singh, V. A Multi-Architecture Approach for Offensive Language Identification Combining Classical Natural Language Processing and BERT-Variant Models. *Appl. Sci.* **2024**, *14*, 11206. https://doi.org/10.3390/app142311206

[6]. Qureshi, Muhammad Deedahwar Mazhar, M. Atif Qureshi, and Wael Rashwan. "Explainable AI for Hate Speech Moderation: A Stakeholder-Centered and Sociotechnical Review." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 16, no. 1 (2026): e70076.

[7]. Fesaghandis, Zahra Safdari, and Suman Kalyan Maity. "Multilingual hate speech detection and counterspeech generation: A comprehensive survey and practical guide." *arXiv preprint arXiv:2603.19279* (2026).

[8]. Pannerselvam, Kathiravan, and Saranya Rajiakodi. "Systematic literature review on hate speech detection in Indian low-resource languages." *Journal of Computational Social Science* 9, no. 1 (2026): 5.

[9]. Goswami, Pragya, and A. Daniel. "Enhancing Hate Speech Detection with a Distil BERT and BiLSTM Hybrid Mode." *SN Computer Science* 7, no. 4 (2026): 353.

[10]. Mahibha, C. Jerin, and Durairaj Thenmozhi. "A survey of hate speech detection in indian languages." *Language Resources and Evaluation* 60, no. 1 (2026): 13.

[11]. Ismail, Usman Idris, Suleiman Salihu Jauro, Nuhu Abdulalim Muhammad, Saadatu Ali Jijji, Joshua C. Shawulu, and Abdullahi Adam Galadima. "Machine Learning for Hate Text Speech Detection: A Comprehensive Review of Techniques, Dataset and Challenges." *Asian Journal of Research in Computer Science* 19, no. 2 (2026): 204-218.

[12]. Goswami, Pragya, Aravendra Kumar Sharma, Ratnesh Kumar Dubey, and Rinki Pakshwar. "Design an Approach for Hate Speech Detection Using Deep Learning." In *Intelligent Systems Using Semiconductors for Robotics and IoT*, pp. 256-263. CRC Press, 2026.

[13]. Zhang, Xingwei, Hu Tian, Xiaolong Zheng, Jing Peng, and Daniel Dajun Zeng. "Semantic aggregated adversarial training framework for hate speech detection." *INFORMS Journal on Computing* (2026).

[14]. Kumar, Deepak, Nitin Goyal, Manisha Wadhwa, and Seema Gulati. "Toxic Content and Hate Speech Detection on Social Media." In *2026 1st International Conference on Advancing Sustainable Solutions through Technologies (ICASST)*, pp. 1-6. IEEE, 2026.

Doi:10.48047/ijiee.2026.16.2.47