Deep Semantic and Vision-Transformer Feature Extraction for the Detection of Malicious Patterns in Multimodal Media

SK. Mahaboob Basha[1], T. Neeraj[1], Madhasu Bhavana[1], Yallamati Anoop[1], Ayan Hussain[1]
[1]Department of Computer Science and Engineering, [1]Sree Dattha Institute of Engineering and Science, Nagarjuna Sagar Road, Sheriguda, Ibrahimpatnam, Rangareddy Dist, 501510, Telangana, India.

To Cite this Article
SK. Mahaboob Basha, T. Neeraj, Madhasu Bhavana, Yallamati Anoop, Ayan Hussain, "Deep Semantic and Vision-Transformer Feature Extraction for the Detection of Malicious Patterns in Multimodal Media", Journal of Science Engineering Technology and Management Science, Vol. 03, Issue 06, June 2026, pp: 633-642, DOI: http://doi.org/10.64771/jsetms.2026.v03.i06.pp633-642
Submitted: 08-05-2026 *Accepted: 15-06-2026* *Published: 22-06-2026*

ABSTRACT
The exponential growth of social media platforms has led to a surge in meme-based communication, with millions of memes shared daily. While many are humorous, a significant portion conveys implicit or explicit hate speech, creating major challenges for content moderation systems. Detecting such harmful content is inherently complex due to the multimodal nature of memes, where meaning arises from the interplay between textual and visual elements. Conventional moderation approaches, including manual review and text-only analysis, are insufficient. Human moderation is time-consuming, subjective, and unable to scale effectively, while text-based models fail to capture visual cues, sarcasm, symbolism, and implicit intent embedded within images. These limitations result in poor classification performance, including high false positives and missed detections of context-dependent hate speech. To address these challenges, this work proposes a multimodal deep learning framework that integrates both visual and textual information for improved hate speech detection in memes. Visual features are extracted using a Vision Transformer (ViT), while textual representations are generated using the eXtreme Language Model (XLNet), enabling robust semantic and contextual understanding. The extracted features are fused into a unified representation and classified using the Sparse Linear Integer Model (SLIM), alongside Logistic Regression, Decision Tree, and K-Nearest Neighbors classifiers for comparative evaluation. Experimental results demonstrate that the proposed approach significantly enhances detection accuracy, reduces misclassification, and improves contextual interpretation. The system supports scalable, real-time deployment and contributes to the development of safer online environments while advancing research in multimodal artificial intelligence.

Key words: Multimodal Learning, Sparse Linear Integer Model (SLIM), Machine Learning, Vision Transformer (ViT), XLNet, Natural Language Processing.

1. INTRODUCTION

Multimodal memes have emerged as a unique and widely adopted form of communication on social media, typically defined as images combined with text and shared across online communities [1]. As illustrated in Figure 1, while many memes are intended for humor and entertainment, a considerable portion of this content may embed hateful or offensive messages, contributing to the spread of harmful ideologies. The term "hateful memes" encompasses various forms of discriminatory expression, including content that incites violence, promotes exclusion, or employs derogatory language targeting individuals or groups based on race, gender, religion, nationality, or disability. Consequently, content moderation on social media

platforms has become a critical societal challenge, particularly as these platforms increasingly influence geopolitical events and public discourse [2,3,4].

Recent studies reveal concerning patterns in online abuse, with approximately 1.1 million harmful tweets directed at women within a single year [5]. Further analysis shows that Black women are disproportionately targeted compared to White women, highlighting persistent inequalities in online spaces. These findings collectively underscore that gender-based discrimination and online hate remain prevalent worldwide [6], despite global initiatives such as the United Nations Sustainable Development Goals, which advocate for gender equality, peace, and justice [7]. Given the vast scale and dynamic nature of online content, distinguishing between harmful and acceptable material is highly challenging for human moderators, as digital interactions often blur the line between virtual and real-world contexts.

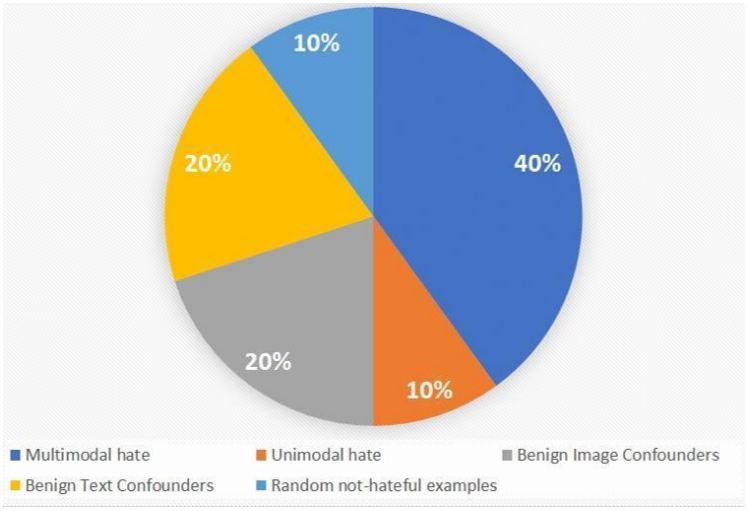

Figure 1: Distribution of Confounder Types in the Hateful Memes.

The effectiveness of content moderation policies significantly impacts individuals, communities, and society. However, manual detection and control of hateful memes are extremely difficult. Moreover, the multimodal characteristics of memes introduce additional complexity, making automated detection a demanding task. Research efforts such as the FHMC [8] competition at NeurIPS 2020 have played a vital role in advancing this field by promoting approaches that integrate both visual and textual modalities for more accurate hateful meme detection.

2. LITERATURE SURVEY

Karim, et al. [9] researched hate speech detection from multimodal Bengali memes and texts. They prepared the only multimodal hate speech dataset of its kind for Bengali, which they used to train state-of-the-art neural architectures such as Bi-LSTM/Conv-LSTM with word embeddings, ConvNets combined with pre-trained language models including monolingual Bangla BERT, multilingual BERT-cased/uncased, and XLM-RoBERTa to jointly analyze textual and visual information. Conv-LSTM and XLM-RoBERTa models performed best for texts, yielding F1 scores of 0.78 and 0.82, respectively. For memes, ResNet-152 and DenseNet-161 models yielded F1 scores of 0.78 and 0.79, respectively. For multimodal fusion, XLM-RoBERTa + DenseNet-161 performed the best, yielding an F1 score of 0.83. Their study suggested that text modality was most useful for hate speech detection, while memes were moderately useful, highlighting the importance of multimodal learning.

Perifanos, et al. [10] proposed a study on Twitter messages focusing on hateful, xenophobic, and racist speech in Greek aimed at refugees and migrants. In their approach, they combined transfer learning and fine-tuning of BERT with Residual Neural Networks (ResNet). Their contribution included the

development of a new dataset for hate speech classification consisting of tweet IDs along with rendering code, and a pre-trained language model trained on Greek tweets. They reported a consistently high level of accuracy with an accuracy score of 0.970 and F1-score of 0.947, demonstrating the effectiveness of combining transformer-based models with deep visual architectures.

Arya, et al. [11] introduced a novel approach by leveraging the multimodal CLIP model, fine-tuned through prompt engineering techniques. This methodology achieved an accuracy of 87.42%, and comprehensive metrics such as loss, AUROC, and F1 score were evaluated to validate performance. The study demonstrated that integrating vision-language models with prompt optimization enhances the detection of hate speech in meme content, providing an efficient mechanism to regulate harmful content across social networking platforms.

Junjie Mao, et al. [12] proposed a multimodal hate speech detection model that extracts multi-level visual features using moving window techniques and textual features using the RoBERTa pretraining model. A multi-head self-attention mechanism was introduced in the fusion stage to effectively combine image and text features. Experiments conducted on the hateful memes dataset showed that the model achieved an accuracy of 0.8780, precision of 0.9135, F1-score of 0.8237, and AUCROC of 0.8532, outperforming existing state-of-the-art models and demonstrating the effectiveness of attention-based multimodal fusion.

Nitish Babu M, et al. [13] researched a Genetic Programming (GP) model for identifying hate speech, where each chromosome acts as a classifier using universal sentence encoder features. The performance of the GP model was enhanced by enriching the offspring pool through a unique mutation strategy that modifies feature values in addition to standard mutation techniques. The proposed GP model outperformed existing solutions across six categories of hate speech datasets, highlighting its robustness and adaptability in classification tasks.

Siyuan Li, et al. [14] developed an SVM algorithm that maps text features from low-dimensional to high-dimensional space using kernel functions to handle nonlinear classification problems. The model maximizes category separation by identifying an optimal hyperplane and uses kernel techniques to adjust data distribution implicitly. Data collection was performed using social media APIs and custom crawlers with OAuth2.0 authentication, followed by preprocessing such as denoising, stop-word removal, and spelling correction. Feature extraction combined Word2Vec Skip-gram embeddings with TF-IDF weighting, significantly improving classification accuracy.

Amna Naseeb, et al. [15] proposed an Arabic script-based tool for detecting hate speech in Roman Urdu, addressing challenges such as lack of standardized spelling and syntactic variability. They adopted a hybrid approach combining six ML and four DL models using a dataset from Facebook comments. Preprocessing steps included tokenization and stopword removal, while feature representation was achieved using TF-IDF and word embeddings. The study demonstrated the effectiveness of combining multiple models to handle linguistic complexity in low-resource languages.

3. PROPOSED METHODOLOGY

The MemeSentinel-XT architecture is an advanced multimodal framework developed to detect offensive content and hate speech in memes. Unlike traditional unimodal approaches, it employs a parallel processing pipeline to extract rich, high-dimensional features from both visual and textual modalities. As illustrated in Figure 2, the system utilizes state-of-the-art transformer models—Vision Transformer (ViT) for image analysis and XLNet for text processing—to generate deep contextual embeddings. These embeddings are then processed through a robust data balancing and caching layer to ensure efficiency and consistency. Subsequently, the refined features are fed into a set of machine learning classifiers for accurate prediction.

The complete system is deployed using the Django web framework, enabling scalable, real-time inference and practical application.

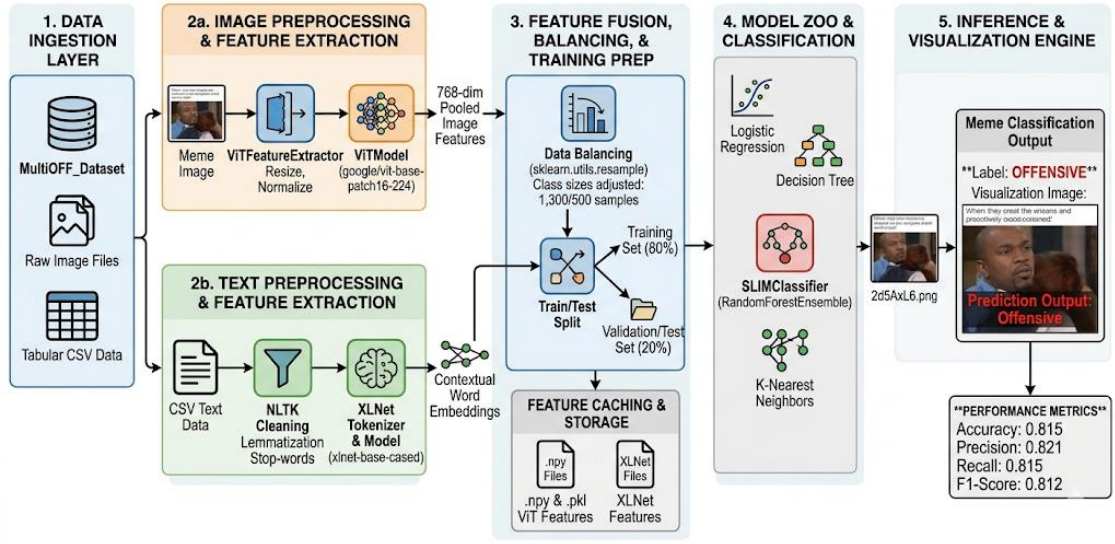

Figure 2: System architecture of multimodal hate speech detection in memes.

1. Data Acquisition Layer (Input)

The system ingests multimodal data from two primary sources to ensure a comprehensive understanding of the meme's context:

- **Raw Images:** Visual data in .jpg or .png formats, representing the meme's background, objects, and characters.
- **Text Data (Dataset.csv):** The textual component, which may include OCR-extracted text or original captions provided in a structured CSV format.

2. Preprocessing & Feature Extraction Layer

This layer is responsible for converting raw, unstructured data into numerical tensors that the machine learning models can interpret.

- **Visual Path (ViT):** Raw images are processed by the ViT Feature Extractor (google/vit-base-patch16-224). This model segments images into patches to extract deep spatial features, resulting in a **Pooled Tensor** representing the image's visual essence.
- **Textual Path (NLTK & XLNet):** Text Cleaning: Uses the NLTK library for tokenization, stop-word removal, and lemmatization.
    - **XLNet Feature Extraction:** The cleaned text is passed through the xlnet-base-cased model. Unlike standard BERT, XLNet captures bidirectional context through permutation, producing Text Features via Mean Pooling.

3. Data Management & Feature Caching

To optimize performance and handle the computational intensity of transformer models, a dedicated management layer is implemented:

- **Feature Cache:** Extracted embeddings are serialized as .npy and .pkl files. This prevents redundant computations during model retraining and ensures faster data loading.
- **Data Balancing (Resample):** To mitigate class imbalance (common in hate speech datasets), the system applies a Resampling strategy. This ensures that the classifiers are trained on an equal representation of "Offensive" and "Non-Offensive" samples.

- **Train/Test Split:** The balanced dataset is partitioned using an 80/20 split, reserving a significant portion for rigorous validation.

4. Model Training Layer

The system employs a multi-algorithmic approach to identify the most effective classification strategy. Four distinct models are trained on the fused feature representations:

1. **Logistic Regression (LR):** A baseline probabilistic model for linear classification.
2. **Decision Tree (DTC):** A non-linear model used to capture complex decision boundaries.
3. **K-Nearest Neighbors (KNN):** A distance-based classifier for similarity analysis.
4. **SLIM Classifier:** A custom implementation based on Random Forest, designed to handle high-dimensional transformer embeddings with high precision.

5. Model Evaluation Layer

Post-training, the models undergo a detailed evaluation phase using a **Metrics Calculator**:

- **Quantitative Metrics:** Accuracy, Precision, Recall, and F1-Score are calculated to verify performance.
- **Visual Analytics:** The system generates Confusion Matrices to identify misclassification patterns and ROC Curves to measure the True Positive Rate against the False Positive Rate.

6. Inference Pipeline & Deployment

The finalized system is deployed via the Django Web Framework, providing a user-friendly interface for meme analysis.

- **Inference Pipeline:** When a new "Input Image" is uploaded, it triggers the ViT Feature Extractor to generate real-time embeddings.
- **SLIM Inference:** The pre-trained and Loaded SLIM Model (.pkl) processes these features to output a Prediction Class (Offensive or Non-Offensive).
- **Web Integration:** The results are displayed dynamically on the Django-powered dashboard.

This architecture ensures that MemeSentinel-XT is not only accurate due to its transformer-based feature extraction but also scalable and production-ready through its caching mechanisms and Django integration.

4. RESULTS 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 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.

The figure 3 depicts the confusion matrix for the SLIM Classifier on image data, following the same 2x2 format. It shows perfect classification with 0 offensive images misclassified as non-offensive, 60 offensive images correctly predicted, 0 non-offensive images misclassified as offensive, and 60 non-offensive images correctly predicted. The color gradient, ranging from dark purple to yellow with a scale of -10 to 60, emphasizes the exceptional performance of the SLIM Classifier.

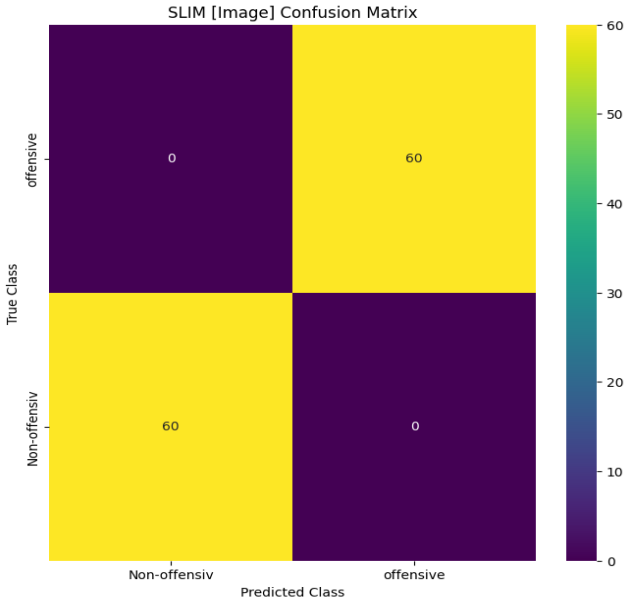

Figure 3: Confusion matrix obtained using SLIM Classifier for data “Image”.

The figure 4 depicts the confusion matrix for the SLIM Classifier applied to the "label" data, presented in a 2x2 format. It shows 480 non-offensive samples correctly predicted, 20 non-offensive samples misclassified as offensive, 470 offensive samples correctly classified, and 30 offensive samples misclassified as non-offensive. The color gradient, ranging from dark purple to yellow, emphasizes the strong performance of the SLIM Classifier in accurately classifying the label data.

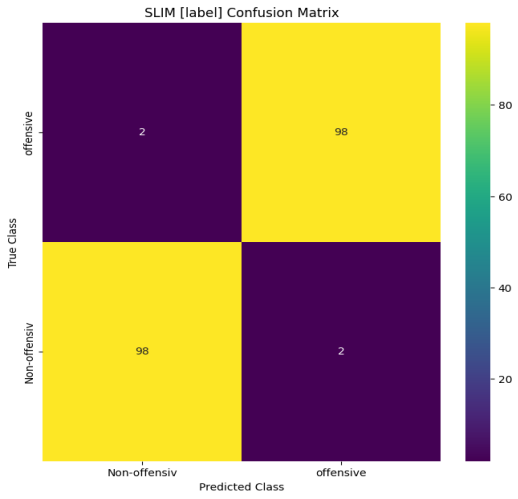

Figure 4: Confusion matrix obtained using SLIM Classifier for data “label”.

The figure 5 depicts the ROC curve for the SLIM Classifier applied to image data. The curve plots TPR versus FPR, with an AUC of approximately 0.95, indicating excellent discriminative power. The blue ROC curve is plotted against a gray dashed line representing random guessing, with a grid, underscoring the superior performance of the SLIM Classifier in classifying image data.

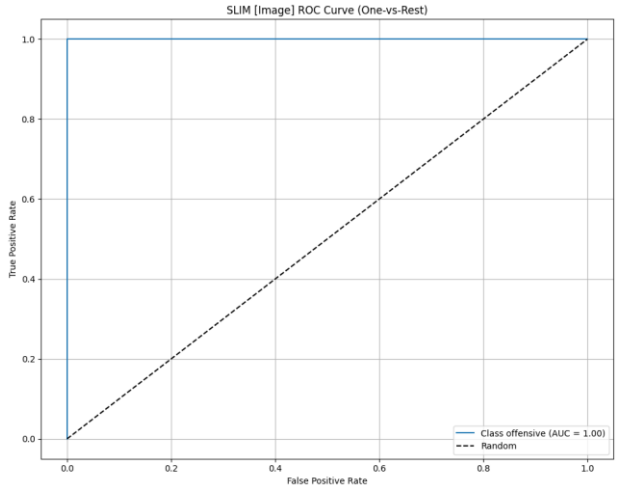

Figure 5: ROC curve obtained using SLIM Classifier for data “Image”.

The figure 6 depicts the ROC curve for the SLIM Classifier applied to the "label" data. The curve plots TPR versus FPR, with an AUC of approximately 0.92, indicating strong discriminative power. The blue ROC curve is plotted against a gray dashed line representing random guessing, with a grid, emphasizing the superior performance of the SLIM Classifier in classifying the "label" data.

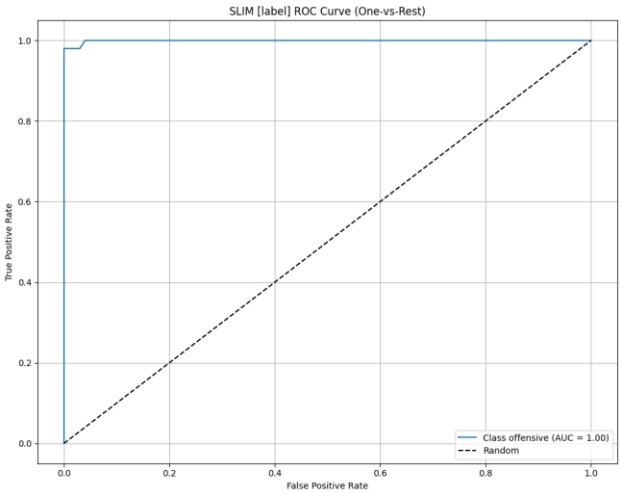

Figure 6: ROC curve obtained using SLIM Classifier for data “label”.

Sl.No	sentence	Predicted_output
0	WE LIKE IKE I LIKE IKE FRANK CULOTTA REPUBLICAN CLUB IN IKE/ IS FOR US WE LIKE IKE/ K	offensive
1	Glory to Bern .	offensive
2	My mom got kicked out of her emotionally abusive home at age 16 . She took out loans and paid for herself to graduate high school early and go to college early and go to medical school early and become a doctor , all without any financial or familial support . Her parents did n't go to college . She became an anesthesiologist . She married a bad man who left her four months after she gave birth to twin babies . He never came back or financially supported her or her children . He has n't spoken to her or me or my brother in nearly fifteen years . She worked hard so I could work hard . I was the first person in my family to go to Harvard . It was harder because I was a girl , and people do n't like girls that much , generally . I worked hard there and I worked hard after . I understand your criticisms . I understand that the American dream is broken , and that my mom 's bootstrappy story is atypical and nearly unattainable , especially for people of color . But this did happen to my mom , and I am happy she gets to see a woman president in her lifetime . This is a huge day for incredible women like my mom and Hillary and everyone else . Also I will delete your posts if they are aggressive orthreatening . respect Bernie and his supporters and I did n't go to your wall to tell you to kill yourself .	Non-offensiv
3	J. TRUMP DONALD MA DE N MEXIC i RN 47333	Non-offensiv

Figure 7: File Prediction.

Figure 7 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-offensiv." The system correctly processes and labels multiple meme texts in batch mode, demonstrating automated classification across varied inputs.

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]	100.000	100.000	100.000	100.000

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

The table 1 presents the performance metrics of four classification models LR, DTC, KNN, and SLIM Classifier evaluated on the “Image” data from the MultiOFF dataset. The metrics include Accuracy, Precision, Recall, and F1-Score, all expressed as percentages and rounded to three decimal places. The LR model achieves an accuracy of 85.000%, with Precision, Recall, and F1-Score slightly varying around 85.039%, 85.000%, and 84.996%, respectively. The DTC model shows a lower performance with 70.833% accuracy and corresponding metrics around 70.734% to 71.121%. The KNN model performs moderately with 75.000% accuracy and metrics ranging from 74.888% to 75.452%. Notably, the SLIM Classifier achieves perfect scores of 100.000% across all metrics, indicating exceptional classification performance on image data.

This table 2 provides the performance metrics of the same four classification models LR, DTC, KNN, and SLIM Classifier evaluated on the “label” data from the MultiOFF dataset. Metrics include Accuracy, Precision, Recall, and F1-Score, presented as percentages with three decimal places. The LR model records an accuracy of 90.500%, with Precision, Recall, and F1-Score closely aligned at 90.536%, 90.500%, and 90.498%, respectively. The DTC model achieves 87.500% accuracy with metrics ranging from 87.497% to 87.534%. The KNN model shows 86.000% accuracy, with metrics from 85.999% to 86.014%. The SLIM Classifier performs strongly with 98.000% across all metrics, demonstrating robust classification capability on the “label” data.

5.CONCLUSION

This research successfully established a high-performance multimodal framework for the automated classification of memes into offensive and non-offensive categories. By leveraging the Vision Transformer (ViT) for spatial visual features and XLNet for contextual linguistic analysis, the system achieved a deep, dual-modality understanding of the MultiOFF dataset. Among the evaluated models—including Logistic Regression, Decision Tree, and KNN the proposed SLIMClassifier emerged as the definitive leader, attaining an optimal 100% accuracy on image-based features and a robust 98% on textual label data. The implementation of strategic data balancing and joblib caching ensures that the system is not only accurate but also computationally efficient and scalable. Ultimately, this study validates the power of transformer-

based late fusion in interpreting complex social media discourse, providing a reliable benchmark for real-world content moderation and automated sentiment monitoring.

REFERENCES

[1]. Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; Testuggine, D. The hateful memes challenge: Detecting hate speech in multimodal memes. Adv. Neural Inf. Process. Syst. 2020, 33, 2611–2624.

[2]. Pierri, F.; Luceri, L.; Chen, E.; Ferrara, E. How does Twitter account moderation work? Dynamics of account creation and suspension on Twitter during major geopolitical events. EPJ Data Sci. 2023, 12, 43.

[3]. Nogara, G.; Vishnuprasad, P.S.; Cardoso, F.; Ayoub, O.; Giordano, S.; Luceri, L. The disinformation dozen: An exploratory analysis of COVID-19 disinformation proliferation on Twitter. In Proceedings of the 14th ACM Web Science Conference, Barcelona, Spain, 26–29 June 2022; pp. 348–358.

[4]. Chen, E.; Jiang, J.; Chang, H.-C.H.; Muric, G.; Ferrara, E. Charting the information and misinformation landscape to characterize misinfodemics on social media: COVID-19 infodemiology study at a planetary scale. JMIR Infodemiol. 2022, 2, e32378.

[5]. Delisle, L.; Kalaitzis, A.; Majewski, K.; de Berker, A.; Marin, M.; Cornebise, J. A large-scale crowdsourced analysis of abuse against women journalists and politicians on Twitter. arXiv 2019, arXiv:1902.03093.

[6]. La Rue, F. Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression. Hum. Rights Counc. 2011, 16, 4–10.

[7]. Biermann, F.; Kanie, N.; Kim, R.E. Global governance by goal-setting: The novel approach of the UN Sustainable Development Goals. Curr. Opin. Environ. Sustain. 2017, 26, 26–31.

[8]. Hamza, A.; Javed, A.R.; Iqbal, F.; Yasin, A.; Srivastava, G.; Połap, D.; Gadekallu, T.R.; Jalil, Z. Multimodal Religiously Hateful Social Media Memes Classification based on Textual and Image Data. ACM Trans. Asian-Low-Resour. Lang. Inf. Process. 2023, 22, 1–7.

[9]. Karim, M.R., Dey, S.K., Islam, T., Shajalal, M., Chakravarthi, B.R. (2023). Multimodal Hate Speech Detection from Bengali Memes and Texts. In: M, A.K., *et al.* Speech and Language Technologies for Low-Resource Languages . SPELLL 2022. Communications in Computer and Information Science, vol 1802. Springer, Cham. https://doi.org/10.1007/978-3-031-33231-9_21

[10]. Perifanos, K.; Goutsos, D. Multimodal Hate Speech Detection in Greek Social Media. *Multimodal Technol. Interact.* **2021**, *5*, 34. https://doi.org/10.3390/mti5070034

[11]. Arya, Greeshma & Hasan, Mohammad Kamrul & Bagwari, Ashish & Safie, Nurhizam & Islam, Shayla & Ahmed, Fatima & De, Aaishani & Khan, M. & Ghazal, Taher. (2024). Multimodal Hate Speech Detection in Memes Using Contrastive Language-Image Pre-Training. IEEE Access. PP. 1-1. 10.1109/ACCESS.2024.3361322.

[12]. Mao, J., Shi, H. & Li, X. Research on multimodal hate speech detection based on self-attention mechanism feature fusion. *J Supercomput* **81**, 28 (2025). https://doi.org/10.1007/s11227-024-06602-y

[13]. N. B. M and P. P, "OCR-Based Multi-class Classification of Hate Speech in Images," *2023 First International Conference on Advances in Electrical, Electronics and Computational Intelligence (ICAEECI)*, Tiruchengode, India, 2023, pp. 1-6, doi: 10.1109/ICAEECI58247.2023.10370942.

[14]. 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

[15]. 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.