International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Semantic Fusion and Cross-Modal Intelligence for Discovering Covert Harmful Patterns in Multimedia Content

P. Vijay Goud[1*], Pitla Abhilash[2], Somani Hemanth Kumar[2], Dala Anirud[2]
[1]Assistant Professor, [2]UG Student, [1,2]Department of Computer Science and Engineering
[1,2]Kommuri Pratap Reddy Institute of Technology, Ghanpur, Ghatkesar, 501301, Telangana, India.
*Correspondence: P. Vijay Goud (panjala.vijay123@gmail.com)

ABSTRACT

The widespread expansion of social media has significantly increased the use of memes as a common form of communication, with millions being shared every day. Although many memes are intended for entertainment, a considerable portion includes implicit or explicit hate speech, creating major challenges for effective content moderation. Identifying such content is complex due to the multimodal nature of memes, where meaning emerges from the combination of visual and textual elements rather than from a single modality. Conventional methods primarily depend on human moderation or text-based analysis. However, human moderation is time-consuming, subjective, and lacks scalability, while text-only approaches fail to interpret visual context, sarcasm, symbolism, and hidden intent, resulting in poor accuracy and higher misclassification rates. To overcome these challenges, this study introduces a multimodal framework that combines both visual and textual features for enhanced hate speech detection. Visual representations are extracted using Vision Transformer (ViT), while textual features are derived using eXtreme Language Model (XLNet), allowing for deeper semantic and contextual comprehension. These features are then integrated into a single representation and classified using various machine learning models, including Sparse Linear Integer Model (SLIM), Logistic Regression Classifier (LRC), Decision Tree Classifier (DTC), and K-Nearest Neighbors (KNN) for comparative evaluation. The proposed approach enhances detection performance, minimizes false positives, and improves contextual interpretation. Furthermore, it enables scalable and real-time implementation, contributing to safer digital environments and promoting advancements in multimodal artificial intelligence research.

Key words: Vision Transformer (ViT), XLNet, Natural language processing (NLP), Computer vision, Sparse Linear Integer Model (SLIM), K-Nearest Neighbors (KNN).

1. INTRODUCTION

Multimodal memes have emerged as a widely used mode of communication across social media platforms, generally described as images paired with text that are extensively circulated among users [1]. As shown in Fig. 1.1, while many memes are designed for humor and entertainment, a considerable number contain offensive or hateful material, contributing to the dissemination of harmful narratives online. The term “hateful memes” refers to different forms of discriminatory content, including material that encourages violence, supports exclusion, or employs derogatory language aimed at individuals or groups based on attributes such as race, gender, religion, nationality, or disability.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

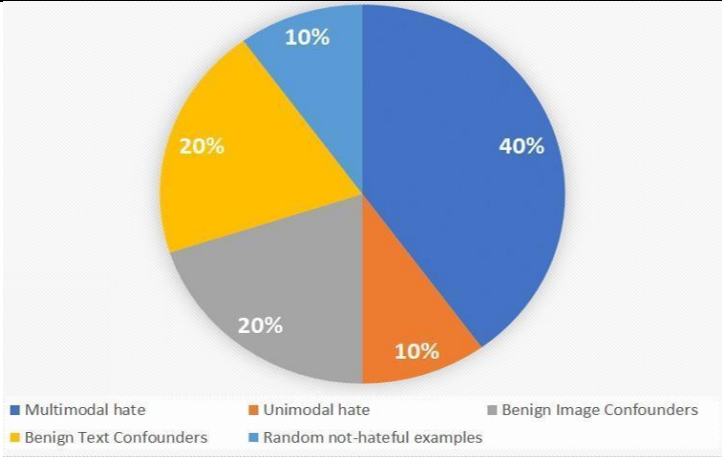

Figure 1: Distribution of confounder types in the hateful memes.

The large-scale and dynamic nature of online content makes it difficult for human moderators to accurately identify harmful material, as digital interactions often blur the boundaries between virtual and real-world contexts. Manual moderation is therefore time-consuming, subjective, and not scalable. Additionally, the multimodal nature of memes introduces further complexity, as their meaning depends on the interaction between visual and textual elements. Initiatives like the Facebook Hateful Memes Challenge (FHMC) at NeurIPS 2020 [8] highlight the need for integrated multimodal approaches to improve the detection of hateful content.

2. LITERATURE SURVEY

Karim et al. [9] explored hate speech detection in multimodal Bengali memes and textual data. They constructed a unique multimodal dataset for Bengali and employed advanced neural architectures such as Bi-LSTM/Conv-LSTM with word embeddings, along with ConvNets integrated with pre-trained language models including monolingual Bangla BERT, multilingual BERT (cased/uncased), and XLM-RoBERTa to jointly analyze visual and textual modalities. For text classification, Conv-LSTM and XLM-RoBERTa achieved F1-scores of 0.78 and 0.82, respectively. For meme analysis, ResNet-152 and DenseNet-161 achieved F1-scores of 0.78 and 0.79. In multimodal fusion, the combination of XLM-RoBERTa and DenseNet-161 yielded the highest F1-score of 0.83. The study concluded that textual features contribute most significantly, while visual features provide moderate support, highlighting the importance of multimodal learning.

Perifanos et al. [10] analyzed hateful, xenophobic, and racist speech in Greek Twitter data targeting refugees and migrants. Their approach combined transfer learning and fine-tuning of BERT with Residual Neural Networks (ResNet). They introduced a new dataset containing tweet IDs and rendering code, along with a pre-trained language model trained on Greek tweets. The model achieved an accuracy of 0.970 and an F1-score of 0.947, demonstrating the effectiveness of combining transformer-based models with deep visual architectures.

Arya et al. [11] proposed a novel approach using the multimodal CLIP model, fine-tuned through prompt engineering techniques. The method achieved an accuracy of 87.42%, and performance was evaluated using loss, AUROC, and F1-score. Their findings showed that integrating vision-language models with prompt optimization enhances hate speech detection in memes and provides an efficient mechanism for regulating harmful content.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Junjie Mao et al. [12] introduced 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 applied during the fusion stage to effectively combine image and text features. Experimental results on the hateful memes dataset showed an accuracy of 0.8780, precision of 0.9135, F1-score of 0.8237, and AUCROC of 0.8532, outperforming existing models and demonstrating the effectiveness of attention-based multimodal fusion.

Nitish Babu M et al. [13] proposed a Genetic Programming (GP) model for hate speech detection, where each chromosome acts as a classifier using universal sentence encoder features. The model's performance was improved through a novel mutation strategy that modifies feature values along with standard mutation operations. The GP model outperformed existing methods across six categories of hate speech datasets, demonstrating its robustness and adaptability.

Siyuan Li et al. [14] developed a Support Vector Machine (SVM) model that maps text features from low-dimensional to high-dimensional space using kernel functions to handle nonlinear classification. The model identifies an optimal hyperplane to maximize class separation while kernel techniques implicitly adjust data distribution. Data collection was performed using social media APIs and custom crawlers with OAuth2.0 authentication. Preprocessing included denoising, stop-word removal, and spelling correction, while feature extraction combined Word2Vec Skip-gram embeddings with TF-IDF weighting, improving classification accuracy.

Amna Naseeb et al. [15] proposed an Arabic script-based tool for detecting hate speech in Roman Urdu, addressing issues such as inconsistent spelling and syntactic variability. They adopted a hybrid approach combining six machine learning models and four deep learning models using Facebook comment datasets. Preprocessing involved tokenization and stop-word removal, while feature representation used 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 proposed study introduces a well-organized analytical framework for identifying hate speech in multimodal meme content using artificial intelligence methods. The process begins with the collection and structuring of the dataset, followed by preprocessing steps applied to both textual and visual information. The system utilizes advanced deep learning techniques to derive meaningful features from images and text. Visual features are extracted through a transformer-based vision model, which captures elements such as objects, symbols, and contextual patterns within images. At the same time, textual content is analyzed using a transformer-based language model to understand semantic meaning, contextual relationships, and linguistic structures. These extracted features are then integrated into a single unified representation and evaluated using multiple machine learning classifiers to carry out content classification.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

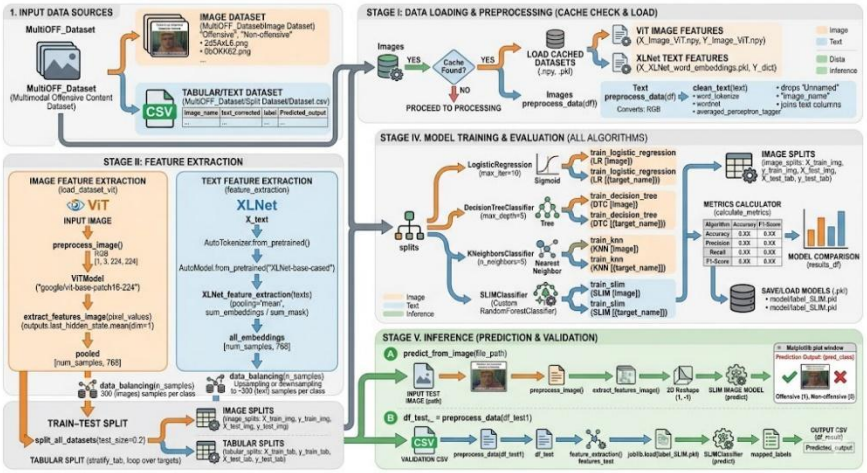

Fig. 4.1: System architecture of multimodal hate speech detection in memes.

The framework also incorporates evaluation and visualization components to assess performance and ensure reliability. A user interface supports interaction for data handling, model training, evaluation, and prediction tasks, as illustrated in Fig. 4.1. A lightweight storage mechanism is used to manage trained models and intermediate data, while the system supports prediction workflows for both text and image inputs. Continuous evaluation and retraining further enhance accuracy and adaptability to evolving data patterns.

1. User Interface (Client Application)

The system features a centralized graphical interface designed for cross-modality interaction and high accessibility.

- **Operational Control:** Enables users to upload meme images, enter associated text, or provide combined multimodal inputs for analysis.
- **Functionalities:** Provides dedicated modules for dataset management, triggering the preprocessing pipeline, and viewing real-time classification results.
- **Backend Integration:** All user-initiated requests are captured and routed to the computational backend for processing and inference.

2. Application Processing Layer

This layer serves as the core computational engine, orchestrating the complex flow of multimodal data.

- **Pipeline Coordination:** Manages the transitions between raw data input, feature extraction, and final model classification.
- **Data Management:** Ensures that textual and visual components are synchronized throughout the analytical lifecycle.
- **Efficiency:** Optimizes the execution of deep learning models to ensure rapid result generation.

3. Dataset (Multimodal Meme Collection)

The framework is powered by a diverse collection of labeled memes representing real-world internet culture.

- **Content:** Contains memes categorized into Offensive and Non-Offensive classes, including diverse linguistic expressions and visual cues.
- **Visual Semantics:** Captures critical contextual elements such as symbols, gestures, and text embedded within the images.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

- **Utility:** Serves as the ground truth for training both the XLNet and ViT components, ensuring the models understand the nuance of internet-based hate speech.

4. Text Preprocessing and Feature Extraction (XLNet)

Textual components are analyzed using a transformer-based language model to capture deep semantic dependencies.

- **Linguistic Cleaning:** Performs tokenization, stop-word removal, and lemmatization to prepare the text for embedding.
- **XLNet Architecture:** Utilizes XLNet to extract contextual embeddings, capturing long-range dependencies and subtle linguistic nuances that traditional models might miss.
- **Numerical Representation:** Converts semantic meaning into dense feature vectors suitable for fusion with visual data.

5. Image Preprocessing and Feature Extraction (ViT)

Visual components are processed using a state-of-the-art Vision Transformer to decode spatial and contextual patterns.

- **Normalization:** Images are resized and normalized to ensure consistent input dimensions for the transformer layers.
- **ViT:** Employs the ViT model to divide images into patches and extract deep visual features, capturing contextual relationships across the entire image area.
- **Spatial Semantics:** Transforms visual patterns into numerical vectors representing the "visual language" of the meme.

6. Feature Fusion and Representation

A critical stage where the framework integrates the two distinct modalities into a unified intelligence vector.

- **Multimodal Integration:** Combines the contextual text embeddings from XLNet with the visual feature vectors from ViT.
- **Enhanced Understanding:** The fusion process allows the system to understand memes where the text alone might be neutral, but the combination with a specific image becomes offensive.
- **Unified Input:** The combined vector serves as the master input for the classification suite.

7. ML Classification Models

The framework employs a diverse ensemble of classifiers to provide a comparative analysis of the fused features.

- **LR:** Performs linear classification based on calculated feature probabilities.
- **DTC:** Uses hierarchical rules to capture complex decision-making logic within the meme data.
- **KNN:** Classifies content based on its similarity to neighboring data points in the feature space.
- **SLIM (Proposed Model):** A robust, Random Forest-based ensemble model designed for high-accuracy and stable multimodal predictions.

8. Prediction Results and Output Generation

The system translates complex transformer logic into interpretable safety indicators.

- **Dual Classification:** Generates a final label indicating whether the content is "Offensive" or "Non-Offensive."
- **Confidence Scores:** Provides numerical scores representing the model's certainty in its prediction.
- **User Interpretation:** Results are displayed in a structured format on the UI, allowing for easy review by content moderators.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

9. Model Evaluation and Visualization

A diagnostic layer is integrated to quantify the precision and reliability of the detection system.

- **Core Metrics:** Evaluates performance using Accuracy, Precision, Recall, and F1-score.
- **Visual Analytics:** Generates confusion matrices and ROC curves to analyze model performance across both classes.
- **Comparative Benchmark:** Allows users to see how the proposed SLIM model performs against traditional baselines.

10. Prediction Workflow

The framework supports a flexible operational flow to handle various types of input.

- **Flexible Entry:** Supports text-only, image-only, or integrated multimodal inputs.
- **Automated Pipeline:** Data automatically flows through the preprocessing, feature extraction (XLNet/ViT), and classification stages without manual intervention.
- **Real-Time Execution:** Designed to provide rapid feedback, suitable for high-volume moderation environments.

11. Model Adaptation and Improvement

The system is engineered for longevity in the fast-moving landscape of internet trends.

- **Continuous Evaluation:** Monitors accuracy over time as meme trends and linguistic variations evolve.
- **Retraining Mechanism:** Supports the ingestion of new datasets to retrain the transformer backbones and ensemble classifiers.
- **Robustness:** This iterative learning process ensures the system remains resilient against new forms of subtle or evolving offensive content.

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.

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.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

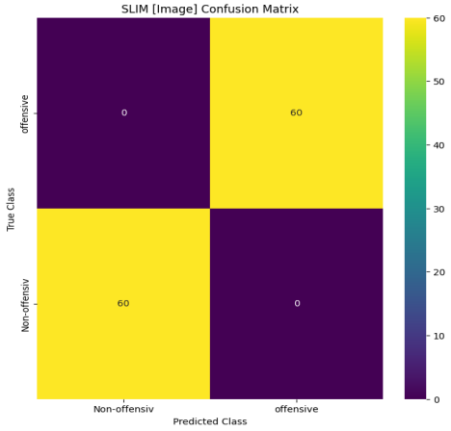

Figure 3: Confusion matrix obtained using SLIM Classifier for data “Image”.

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.

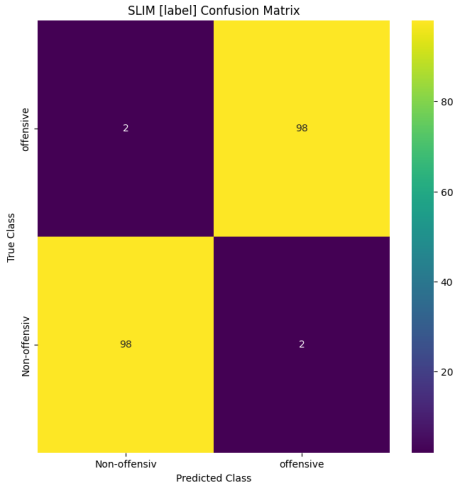

Figure 4: Confusion matrix obtained using SLIM Classifier for data “label”.

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.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

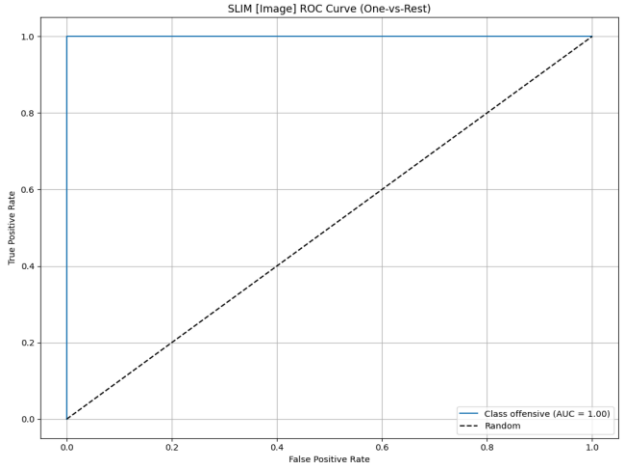

Figure 5: ROC curve obtained using SLIM Classifier for data “Image”.

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.

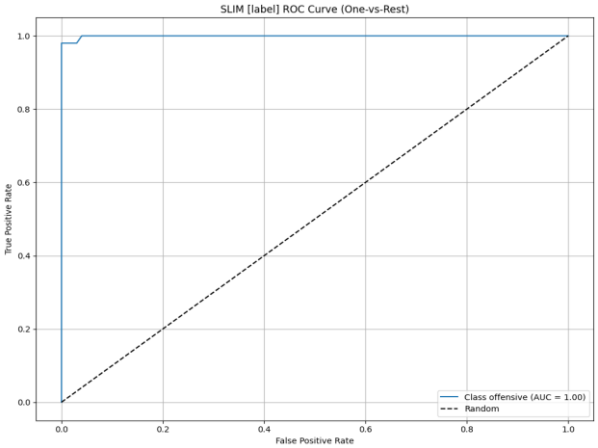

Figure 6: ROC curve obtained using SLIM Classifier for data “label”.

Figure 7 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.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

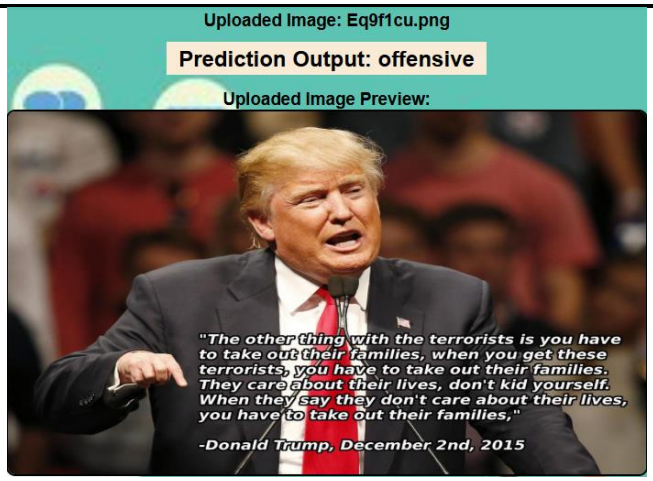

Figure 7: Prediction of Image output.

Figure 8 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.

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 8: File Prediction.

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

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

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 such as 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 such as 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 study introduces a robust multimodal framework for automatically classifying memes as offensive or non-offensive. By combining visual feature extraction using ViT with textual representation through XLNet, the framework effectively captures both image-based and linguistic information within the MultiOFF dataset. The integration of these two modalities allows for a deeper understanding of complex contextual relationships often present in social media memes. A comparative evaluation was performed using several machine learning models, including LR, DTC, KNN, and SLIM, where the SLIM Classifier achieved the best performance, reaching 100% accuracy on image features and 98% accuracy on textual data. These outcomes demonstrate the strength of combining ensemble-based methods with deep feature extraction techniques. Furthermore, applying data balancing methods improved class distribution, while joblib-based caching enhanced computational efficiency and minimized processing time. The results underline the significance of integrating transformer-based models with traditional ML techniques for precise hate speech detection. The proposed framework not only boosts classification accuracy but also ensures scalability and efficient deployment. Overall, this work highlights the effectiveness of multimodal learning in tackling challenges in content moderation and establishes a solid base for future research in automated social media content analysis.

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.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

[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.