International Journal of Computer Sciences and Engineering
Vol.**13**, Issue.**7**, pp.**01-09**, July **2025**
ISSN: 2347-2693 (Online)
Available online at: www.ijcseonline.org

ISSN: 2347-2693 (E)

Research Article

Multimodal Deep Learning for the Detection of Racist Content Online

Rewanshu S. Bopapurkar[1*], A.G Phakatkar[2]

[1,2]Dept. of Computer Science, PICT College, Pune, 411043, Maharashtra, India

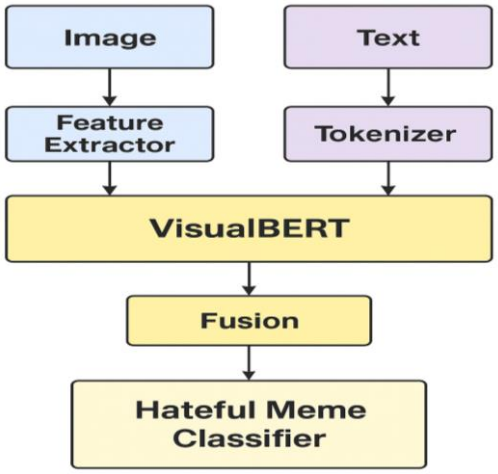
**Corresponding Author:* ✉

Received: 18/May/2025; **Accepted**: 20/Jun/2025; **Published**: 31/Jul/2025. **DOI**: https://doi.org/10.26438/ijcse/v13i7.19

Abstract: This paper responds to the challenges of online racist content, especially insidious, context-dependent forms such as memes and image-text pairs, which tend to elude traditional unimodal content moderation. A multimodal deep learning model that is targeted at detecting this kind of content by jointly considering both textual and visual information. Our suggested approach combines VisualBERT, a vision-language representation-based transformer model, with a Vision Transformer (ViT) for high-level visual feature extraction. This combined model allows the system to capture context-dependent racist cues and successfully discern between offensive and non-offensive ones. The system was thoroughly tested on the Hateful Memes dataset, which contains more than 10,000 meme instances where multimodal understanding is required for proper classification. The model reported a validation accuracy of 0.79, with all recall values higher than 0.79 and an F1-score higher than 0.75 in training. Performance on unseen test data validated the model's strong generalization ability, with accuracy between 0.76 and 0.79 and high recall at 0.70. These findings underscore its high suitability for improving content moderation as well as furthering safer online communities.

Keywords: Content Moderation, Deep Learning, Multimodal Racism Detection, Social Media Analytics, Text and Image Analysis, VisualBERT, Vision Transformer

Graphical Abstract- A multimodal deep learning framework is introduced to detect racist content in online memes, integrating text and image analysis. It leverages VisualBERT for combined vision-language representation and a Vision Transformer for visual feature extraction. Text and image data are converted into embeddings and patch-based vectors, respectively. These modalities are fused in VisualBERT to identify subtle, hateful cues. This significantly enhances content moderation by effectively detecting implicit racism in multimodal social media.



1. Introduction

1.1 Background

Social media sites like Facebook, Twitter, and Reddit have transformed global communication by enabling people to communicate and connect with varied groups. But with these advantages have come potential rich soil for hate speech, racism, and cyberbullying. The dynamic nature of online racism particularly in meme and image-text pair form, provides serious challenges to conventional content moderation models [1].

Digital platforms' pervasive nature has inadvertently allowed the proliferation of detrimental and hostile discourse. Despite fostering global connectivity and free expression, these online environments now frequently host hate speech and racially charged material. Internet memes, which often employ humor or sarcasm to obscure bigoted messages, are a notable example of this trend. The interplay of visual and textual elements in such memes complicates the identification of hateful content far beyond what's seen in purely text-based communications.

1.2 Problem Definition

Although existing research has succeeded in identifying such hate when it appears in text forms, contemporary racist material increasingly takes on subtle, encoded, or contextually embedded forms. For example, a seemingly harmless image paired with a sarcastic or coded caption can convey a racially insensitive message that evades traditional keyword-based or unimodal classifiers [2].

A core challenge arises from unimodal methods' struggle to discern the underlying meaning of content, where hate is frequently camouflaged by humor, irony, or cultural references. This shortcoming introduces considerable hazards to online safety and the mental health of affected groups. Examining visual or textual elements separately often leads to incorrect classifications or overlooked harmful material.

1.3 Motivation

This study is driven by increasing societal worries regarding online racism and hate speech, especially when presented in ways that bypass easy identification. With the ongoing evolution of the digital realm, there's an urgent demand for smart moderation systems that protect users from overtly animosity while also recognizing subtle, hidden forms of prejudice. The rising complexity of online hateful expressions necessitates detection systems that are equally sophisticated, capable of grasping the relationship between images and text. This work aims to provide technological answers that promote healthier online engagement and reinforce inclusivity and respect within digital communities.

1.4 Research Objective and Proposed Solution

This paper focuses on the detection of racist content in multimodal online media, particularly those combining text and images. Memes have emerged as a prominent medium for conveying racism in subtle or implicit forms. In many cases, the offensive nature of a meme is revealed only through the interplay between its textual and visual components; analyzing either modality in isolation often results in misclassification or false negatives. These challenges underscore the importance of multimodal analysis for the accurate identification of hateful content on the internet [3].

To address this gap, a deep learning architecture is introduced that combines VisualBERT, a vision–language transformer model, with a Vision Transformer (ViT) for abstract-level image feature representation. This integrated framework enables the system to capture context-dependent racist cues and to more accurately distinguish between offensive language and explicitly racist content. The central aim of this study is to construct a sophisticated, multimodal deep learning system designed to identify racist content embedded contextually within both visual and textual elements. Utilizing vision-language models, this endeavor seeks to surmount the inherent weaknesses of current unimodal detection systems, offering a more complete and precise approach for online content moderation.

1.5 Key Contributions

A multimodal racism detection system is proposed, which jointly analyzes textual and visual content to address the limitations of unimodal approaches.

The proposed approach integrates VisualBERT for joint vision–language representation learning and a Vision Transformer (ViT) for high-level image feature extraction, enabling the system to effectively capture nuanced and context-dependent indicators of racism.

The system is evaluated on real-world datasets to assess its generalization performance and effectiveness in detecting subtle racism in online memes.

By developing multimodal racism detection, this research makes contributions not just to the realm of artificial intelligence but also to more general social efforts towards creating safer, more respectful digital spaces.

1.6 Organization

This paper is structured as follows: Section 1 provides an introduction, covering the background, problem statement, motivation, research objectives, and key contributions of the present study. Section 2 offers a review of related literature, summarizing prior methodologies involving traditional machine learning, deep learning, and multimodal approaches for detecting racism and hate speech. Section 3 delineates the proposed methodology, specifically detailing the integration of VisualBERT and the Vision Transformer (ViT) for multimodal analysis of textual and visual information. Section 4 outlines the experimental setup, including dataset specifications, chosen evaluation metrics, and the training procedure. Section 5 presents empirical results and a comprehensive discussion, emphasizing the model's performance, a comparative analysis with existing models, and insights derived from evaluation on unseen data. Finally, Section 6 concludes the paper by summarizing the principal findings, acknowledging the limitations of the current work, and proposing avenues for future research. Additionally, the manuscript incorporates author statements concerning data availability, potential conflicts of interest, funding sources, individual contributions, and acknowledgements.

2. Related Work

Online platforms are seeing a rise in hate speech, cyberbullying, and aggressive language across all social media platforms and requires the evolution of advanced detection systems. This section follows the development history of these automated techniques, starting with some root traditional machine learning models, progressing to deep learning architectures and transformer-based approaches, and finally ending with sophisticated hybrid and multimodal models.

2.1 Traditional Machine Learning Approaches

Early hate speech identification was based on conventional machine learning (ML) techniques utilizing manually designed features such as linguistic signals or user information. ML algorithms are capable of learning from such information to label content as discriminatory or not. Support Vector Machines (SVM) have also exhibited

excellent performance, especially in combination with TF-IDF features, since they can define optimal hyperplanes across multi-dimensional spaces. K-Nearest Neighbors (KNN), while simple in its implementation, performs worse on sparse, high-dimensional text since it relies significantly on calculating distances.

Naive Bayes (NB), particularly the Multinomial version, is still a well-used option for text classification owing to its simplicity and sparsity efficiency. The Radial Basis Function (RBF) kernel, often employed with SVM, improves classification by being able to create non-linear decision boundaries, so it can be used for identifying intricate hate speech patterns. These traditional models serve as a baseline foundation for more sophisticated deep learning and hybrid methods in racism detection studies.

SVM has been applied to detect hate speech by classifying anti-Semitic content from Yahoo and American Jewish Congress websites, achieving 0.94 accuracy and demonstrating its strength in text classification. However, later studies indicated that more advanced models would outperform SVM in complex situations [4].

Radial Basis Function (RBF), when utilized as a kernel for SVM, has demonstrated excellent performance in multiclass categorization. In this research experiment conducted on Arabic tweets, RBF posted 0.60 across metrics like accuracy, precision, recall, and F1-score, outperforming conventional classifiers MLP, KNN, and Naive Bayes, because it can represent non-linear trends very effectively [5].

KNN was used to detect cyberbullying, analyzing a sample of 47,692 tweets. After preprocessing and feature extraction, KNN achieved 0.90 accuracy, though it encountered issues with memory usage and overfitting. In comparison, SVM and deep learning models reached 92% and 0.96 accuracy, respectively, highlighting KNN's relative shortcomings [6].

Naive Bayes (NB) has also been applied for hate speech detection, e.g., the UCI hate speech dataset. Although Naive Bayes offers computational efficiency and strong performance with high-dimensional datasets, NB performed poorly (0.798) than SVM (0.993), while previous work documented NB performance at 0.73 - 0.76, showing its weakness in this area [7].

2.2 Deep Learning Architectures

Deep learning has greatly improved hate speech and racism detection by enabling models to directly extract complex features from unprocessed data without hand-engineering features. Convolutional Neural Networks (CNNs) excel at identifying localized patterns, like toxic word combinations in brief social media posts. Meanwhile, Long Short-Term Memory (LSTM) networks are designed to capture sequential relationships, and they are appropriate for learning context in more complex sentences, which is important for identifying implicit racisms.

Hierarchical Attention Networks (HANs) build on these by using attention mechanisms at word and sentence levels, enabling the model to concentrate on the most salient text portions. This design honors the hierarchical structure of language and enhances performance and interpretability, especially for inputs of longer lengths.

Researchers have explored various deep learning models for identifying hate speech. For instance, studies using datasets like ArHS, comprising 10,000 Arabic tweets, found that Convolutional Neural Networks (CNNs) achieved the best accuracy in binary classification, reaching 0.81, while for multiclass classification—spanning racism, misogyny, and religious discrimination, the BiLSTM-CNN model demonstrated superior performance over others with 0.67 accuracy [5]. Similarly, Badjatiya et al. explored how CNN and LSTM models performed architectures on Twitter data, finding that while CNNs are effective at capturing local patterns, LSTM models are superior in modeling contextual dependencies across longer text sequences [8].

In another study, Hierarchical Attention Networks (HANs) were employed to classify Reddit posts using word2vec embeddings. HAN reached an F1-score of 0.896, surpassing the results of various traditional machine learning models such as SVM (0.893) and Random Forest (0.844), AdaBoost (0.835), and Naive Bayes (0.796). The results demonstrate HAN's effectiveness in incorporating both contextual and sentiment-level features for improved hate speech classification [9].

Despite these advantages, models like Hierarchical Attention Networks (HAN) generally necessitate substantial amounts of labeled data, and they may face challenges with generalization when applied to unseen datasets.

2.3 Transfer Learning and Transformer-based Models

The emergence of transfer learning has significantly revolutionized Natural Language Processing (NLP), allowing models pretrained on extensive general-purpose datasets to be adapted for specialized tasks. This development paradigm has proven particularly effective in scenarios with limited task-specific training data. Early implementations of transfer learning in NLP focused on the use of pre-trained word embeddings such as Word2Vec, GloVe, and FastText [10].

BERT marked a major advancement in NLP by enabling bidirectional context modeling of words within text. BERT has since demonstrated superior performance across a diverse range of applications, such as identifying hate speech, where it has consistently shown superior performance compared to conventional machine learning and prior deep learning architectures [11].

AraBERT, a transformer architecture that was pre-trained using Arabic text, was fine-tuned with a Multitask Learning (MTL) approach to identify offensive language and hate speech within the OSACT4 dataset (10K tweets). MTL improved generalizability and addressed data imbalance, where only 5% were labeled as hate speech. The Multitask AraBERT model achieved macro-F1 scores of 0.90 for offensive content and 0.82 for hate speech, surpassing the performance of single-task models and also multi-label

variants. Farasa segmenter and custom preprocessing ensured compatibility with AraBERT's input format.[12].

In the Spanish language context, the BETO model—a BERT architecture pre-trained exclusively on Spanish texts—was applied to detect racist and xenophobic content in a balanced dataset of 2,000 tweets. BETO achieved an accuracy of 0.85, outperforming the multilingual mBERT, which recorded 0.82 precision. This result highlights the importance of developing native language-specific models for improved performance in transfer learning tasks [13].

MuRIL and LaBSE, transformer-based models, were evaluated using the Hindi Hate Speech Dataset (HHSD) within a single-task learning setup. MuRIL achieved 0.84 accuracy and 0.84 F1 on hateful language detection (Layer A), while LaBSE attained 0.82 accuracy and 0.82 F1. For implicit/explicit hate (Layer B), MuRIL scored 0.70 accuracy, outperforming LaBSE. In multi-label tagging (Layer C), MuRIL had 0.53 exact matches. Both models showed promise for multilingual hate speech detection. [10].

2.4 Hybrid and Multimodal Approaches

The integration of diverse model architectures and data modalities has significantly improved the effectiveness and resilience of hate speech detection systems. Empirical studies have demonstrated that combining multiple deep learning architectures yields superior results compared to single-model approaches. Hybrid architecture such as CNN+LSTM and CNN+GRU have consistently outperformed standalone CNN and LSTM models in hate speech detection tasks. Specifically, the LSTM+GRU hybrid achieved a 2–3% performance improvement across five benchmark datasets when compared to single LSTM or CNN implementations [14].

A cascaded framework combining a pre-trained transformer (ArabicBERT), a deep learning sequence model (BiLSTM), and a traditional classifier (Radial Basis Function, RBF) has been proposed for multiclass Arabic offensive language classification in [5].

This model classifies Arabic tweets into categories including bullying, insult, racism, obscene content, and non-offensive text. The cascaded pipeline feeds the output of ArabicBERT, augmented with Word2Vec embeddings, into BiLSTM, which is subsequently passed to the RBF classifier. The model achieved exceptional performance: 0.984 accuracy, 0.982 precision, 0.928 recall, and a 0.984 F1-score. It outperformed previous benchmarks by a considerable margin, surpassing AraBERT by 18% in hate speech detection and SVM by 31.4% in racism detection.

With the growing presence of offensive content shared via multimedia (i.e., text and images), recent efforts have shifted towards multimodal approaches for hate speech classification. One of the early efforts in this space was by Gomez et al., who utilized the MMHS150K dataset comprising both text and image data from Twitter. Models that jointly processed both modalities attained a mean accuracy of 0.683, an AUC of 0.732, along with an F1-score of 0.703. However, these early multimodal approaches did not consistently outperform text-only unimodal models [15].

DisMultiHate, a novel multimodal framework designed for hateful meme classification. The model aims to disentangle and explicitly represent hate speech entity terms such as race and gender. DisMultiHate integrates modules for data preprocessing, text representation learning, and visual feature extraction (leveraging VGG16 for image features). Trained on Facebook Hateful Memes (FHM) and MultiOFF datasets, it achieved 0.758 accuracy and 0.828 AUROC on FHM, and 0.646 weighted F1 on MultiOFF. It has consistently outperformed existing unimodal and multimodal baselines on benchmark datasets such as MultiOFF [16].

3. Methodology

Figure 1 shows the proposed multimodal hate speech classification system processes both textual and visual components.

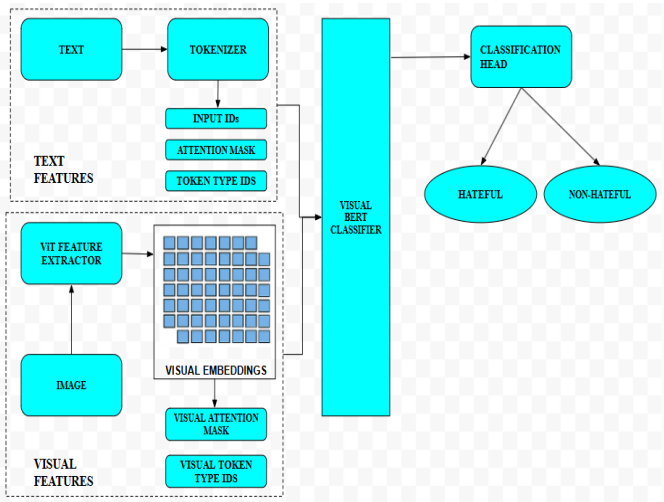


Figure 1. Illustrates the entire architecture of the suggested multimodal hate speech classification system, visually tracing the data flow and the integration points where text and visual processing merge within the mode

3.1 Text Features

The textual content, usually taken from the caption or inserted overlay text. Then the raw TEXT goes through the BERT-based tokenizer, where the readable input is translated into numerical forms appropriate for deep learning models. The tokenizer does subword segmentation through WordPiece encoding and returns three essential components: Input IDs, which are integer indices for every token; Attention Mask, a binary mask in which 1 is a valid token and 0 is padding; and Token Type IDs, which are employed to identify between segments (mostly for sentence-pair tasks).

These three outputs make up the text features that are then supplied to the VisualBERT model. The input IDs retain the semantic structure of the sentence, the Attention Mask directs the transformer to attend to the appropriate tokens, and the Token Type IDs provide segment-level differentiation. Within the VisualBERT model, these text features are

embedded and contextualized with self-attention mechanisms. The model picks up subtle linguistic patterns, implicit prejudice, sarcasm, or racial insults that are typically identifying characteristics of hate speech. The context embedding of the [CLS] token that covers the whole sentence is particularly important as it gets deployed downstream in the classifier for label prediction. Together with image features, this enables the model to recognize hate.

3.2 Visual Features

To encode visual semantics of images, the system utilizes a Vision Transformer (ViT) for feature extraction from images. As indicated in Figure 1, it starts at the IMAGE block where raw input images are read and normalized to RGB format. These images are fed into the ViT Feature Extractor, which applies normal transformations such as resizing, normalization, and patch embedding. This module makes image data compatible with transformer architectures by mapping each image to a sequence of 16×16-pixel patches. The output of this block, VISUAL EMBEDDINGS, is derived from the ViT model's final hidden layer with the shape (197, 768), where 197 is the number of patches plus a class token, and 768 is the hidden dimension.

These visual embeddings encode the high-level abstract feature vectors such as objects, faces, scene context, and demographic information that are important in understanding the kind of visual cue that may signal hate, especially when text alone is ambiguous. A linear transformation is applied to these embeddings to match the modality dimensions before fusion with text features in VisualBERT Classifier. Additional visual attention masks and visual token type IDs are also generated to facilitate attention mechanisms. The processed image features, when combined with their textual counterparts, enable the model to detect hateful content more accurately by considering the interaction between visual and textual modalities.

3.3 Visual BERT Classifier

The VisualBERT Classifier is primarily a multimodal component that forms the core architecture and integrates the interpretation of textual as well as visual information in detecting hate content. It takes two primary inputs, first that are extracted by a BERT tokenizer, namely Input Ids, Attention Mask, and Token Type Ids, while the visual embedding is obtained from the Vision Transformer with their corresponding Visual Attention Mask and Visual Token Type Ids. These inputs are fed into the VisualBERT model, prior to fusion, the visual embeddings undergo a linear transformation to match the dimensional space required for multimodal integration.

Inside the model, VisualBERT applies multi-head self-attention and transformer layers to jointly process both modalities, learning deep contextual alignments between textual tokens and image patches. The output is then fed to a dropout layer, incorporated to mitigate overfitting. Training of the model employed a class-weighted cross-entropy loss function, which was meant to handle dataset imbalance, having weights derived from the label distribution. The output of the Visual BERT Classifier is a probabilistic distribution across the categories.

3.4 Classification Head

The concluding phase of the multimodal design is the Classification Head, which is responsible for mapping the fused multimodal representations to discrete output labels: HATEFUL and NON-HATEFUL. Following text and visual embedding fusion through the VisualBERT model, the pooled output yields a [CLS] token representation. The [CLS] token is positioned at the start of the combined input sequence (text tokens + image embeddings). Its final representation captures information from both modalities due to the model's self-attention mechanism, which allows all tokens—textual and visual—to attend to each other. This vector is initially passed through a dropout layer for the purpose of reducing overfitting and enhancing generalization. Subsequently, a fully connected linear layer maps the pooled embedding into a logit vector of 2 dimensions, where each dimension represents a classification score for one of the two classes.

These logits are unnormalized, raw scores that indicate the model's confidence in each class. A softmax activation is used on these logits to transform them into probabilities so that the output values range from 0 to 1 and sum to 1. The final class label is decided by taking the class with maximum probability. The model utilizes a weighted cross-entropy loss function during training to address class imbalance, with weights inversely proportional to class frequencies specified beforehand. This classification mechanism allows the model to effectively differentiate between diffuse hateful signals and harmless content by incorporating both linguistic and visual contexts.

4. Experimental Setup

4.1 Dataset

The primary dataset employed in this study is the Hateful Memes dataset, openly released by Facebook AI. This dataset was specially curated to support the creation and testing of multimodal hate speech detection models. The dataset contains more than 10,000 unique instances of memes, each combining both an image and relevant text content. Importantly, each meme is carefully annotated with a label indicating whether the content is hateful, using a clear ground truth for classification [17].

One salient feature of this data set is its design, with many instances where the hateful intent of a meme cannot be properly established by examination of the image or text separately. This inherent multimodal uncertainty calls for advanced models that can understand the intermodal synergy between visual and textual content. Every record in the dataset includes the raw text content, the relevant image file path, and its final hateful/non-hateful label. For robust model building and testing, the dataset is divided into distinct portions for training, validation, and testing purposes. Importantly, it provides unseen splits, which are especially useful for strictly evaluating the generalization power of the model to new and unseen material.

Figure 2. Example of 'Hateful' meme from dataset.
The images feature a compilation of assets, with contributions from ©Getty Images.

Figure 3. Example of 'non-hateful' meme from dataset.
The images feature a compilation of assets, with contributions from ©Getty Images.

4.2 Evaluation Parameters

To assess the efficacy of the presented multimodal hate speech detection model, experiments are conducted on both observed and unobserved validation splits to assess its generalizability. The primary performance metrics encompass accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUROC), which collectively offer a thorough understanding of the model's proficiency in identifying racist material. These indicators are especially crucial when dealing with imbalanced data, as they provide insights into both the correctness and consistency of predictions.

Accuracy indicates the ratio of correct predictions (both positive and negative) to the total number of predictions. While it offers a general measure of correctness, it can be deceptive when dealing with imbalanced datasets.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Precision quantifies the proportion of actual positive instances among all instances classified as positive. This metric demonstrates the model's dependability when making a positive class prediction.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

Recall assesses the model's capacity to detect every relevant instance belonging to the positive class. It is calculated as the proportion of correctly identified positive cases to all existing positive instances.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

The F1-score, a harmonic mean of precision and recall, offers a balanced measure of performance. This metric is particularly valuable when class distributions are skewed, and both false positives and false negatives carry significant implications.

$$\text{F1-score} = 2\times\frac{Precision\times Recall}{Precision+Recall} \quad (4)$$

AUROC assesses a model's discriminative power between classes by illustrating the relationship between the true positive rate and the false positive rate at varying thresholds. A higher AUROC suggests superior classification effectiveness across diverse threshold settings.

With TP, TN, FP, and FN representing true positives, true negatives, false positives, and false negatives, in that order.

The integration of textual and visual modalities enables the proposed multimodal system to capture nuanced, context-dependent signals—such as sarcasm and embedded bias—that unimodal models typically overlook. As a result, the proposed framework demonstrates superior performance compared to unimodal baselines, while also addressing concerns related to scalability, fairness, and generalizability, thereby supporting its practical and ethical deployment in real-world content moderation systems.

5. Results and Discussion

To address class imbalance in the Hateful Memes dataset, class weights were computed and integrated into a weighted cross-entropy loss function. The dataset consisted of a total of 10,576 samples, split into training (8,500), validation (1,040), and test (1,036) sets (80% dedicated to training, 10% to validation, and the remaining 10% to testing). Subsequently, the model was trained for 75 epochs using the AdamW optimizer, which is well-suited for transformer-based architecture. A batch size of 24 was primarily used, although it varied between 8 and 32 depending on GPU memory constraints. To mitigate overfitting, dropout layers and early stopping were employed. PyTorch Lightning was used to decouple the training process, while it provided real-time experiment tracking and hyperparameter tuning where a grid search method was applied to optimize the learning rate, weight decay, and other parameters such as dropout probability.

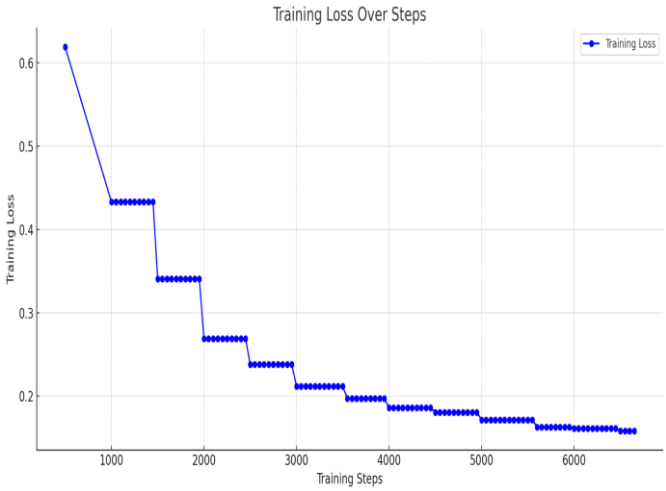


Figure 4. Depicts the evolution of training loss across the various training iterations. This provides a clear visual insight into the model's learning trajectory and its approach to convergence.

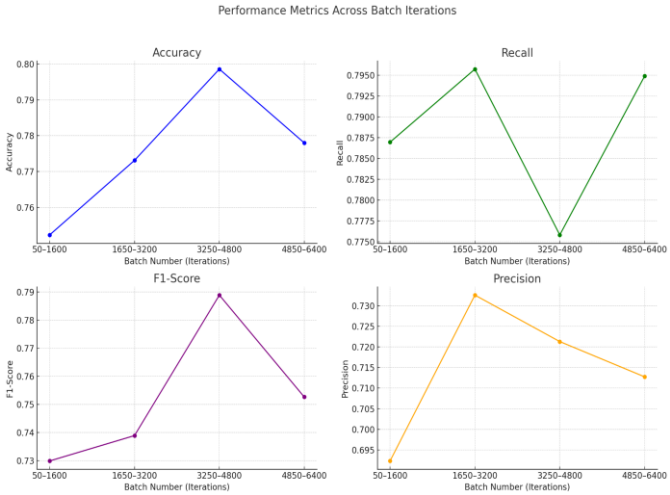


Figure 5. Displays the progression of essential performance metrics—namely Accuracy, Precision, Recall, and F1-Score—over multiple training epochs. This offers insight into the model's increasing proficiency in accurately discerning hateful content as training progresses.

5.1 Training Progress and Loss Evaluation

The training was evaluated using AUROC as the primary metric. As shown in the training loss graph (Figure 4), the model began with a high loss of 0.619, a typical consequence of random weight initialization. A sharp decline was observed early, with the loss reaching 0.433 by step 1000, indicating rapid convergence in the initial phase. The loss continued to decrease gradually, falling to 0.238 by step 2500 and ultimately reaching 0.161 by step 6500, demonstrating that the model effectively optimized its parameters.

The consistent reduction in training loss across more than 5000 training steps signifies that the model generalizes well to hateful and non-hateful meme classification without overfitting. The application of class-weighted loss ensured balanced learning across both categories, reinforcing the steadfastness and consistency of the model during the entire training process.

5.2 Validation Performance and Key Metrics

The model performed well under various measures of evaluation, with significant improvement in recall and F1-score, which are crucial indicators of the durability of content moderation systems. At training, the best validation accuracy of 0.797 was recorded at epoch 3000, which indicates its ability to generalize well over time with minimal overfitting. A consistently high recall of greater than 0.79 across epochs indicates the model's efficacy in classifying true positive instances of racist content, which is critical in minimizing false negatives in online safety uses. The F1-score exceeded 0.75 during subsequent stages of training, reflecting symmetrical performance between precision and recall, but precision scores were consistent, reflecting robust specificity in predictions.

5.3 Evaluation on Unseen Data

Performance with unseen test data also confirmed the model's capacity for generalization. As shown in Table 1, four different batch intervals (50 to 6400 iterations), accuracy varied between 0.762 and 0.798, and recall was in the high 0.70. These observations highlight the model's resilience when applied to practical multimodal content that can differ considerably in style and semantics.

Table 1. Provides a detailed comparison of the model's performance at different training intervals on unseen test data. The variation across batch numbers helps in understanding the model's stability and consistency

Batch Number	Accuracy	Recall	F1-Score	Precision
50–1600	0.75234	0.786953	0.72990	0.69232
1650–3200	0.77312	0.795736	0.73893	0.73258
3250–4800	0.79856	0.775805	0.78893	0.72130
4850–6400	0.77800	0.794899	0.75268	0.71274

In addition, AUROC scores were 0.68 to 0.72, which suggests adequate discriminative ability at classification thresholds. The studies from the literature emphasize the increasing effectiveness of multimodal and attention-based models for hate speech detection, consistent with trends in performance observed here. The proposed model demonstrates highly useful applicability for scalable, precise, and context-aware racism detection within real-world online settings.

5.4 Comparative Analysis with Prior Models

This study addresses the challenge of detecting implicit and contextually embedded racism in text and image. Prior models such as those evaluated on the MMHS150K and DisMultiHate frameworks reported F1-scores below 0.74 and often struggled to generalize across datasets. In contrast, the proposed system achieved consistently higher F1-scores, peaking at 0.769, with recall reaching 0.799, and accuracy exceeding 0.798 across various checkpoints. These improvements represent a performance gain of approximately 2–5%. High recall values are particularly critical in content moderation scenarios to minimize false negatives.

However, limitations remain. The Hateful Memes dataset, though multimodal, is affected by limited language diversity. Additionally, evolving hate expressions and cultural context challenge model adaptability. These findings highlight the need for more adaptive, fair, and explainable multimodal hate detection systems.

6. Conclusion and Future Scope

This work introduces a strong multimodal deep learning model for racism detection through the combination of text and vision inputs. Leveraging VisualBERT for end-to-end vision-language representation and Vision Transformer (ViT) for feature extraction on images, the model shows more than 0.79 accuracy and high recall on dataset, reflecting its potential in capturing subtle and context-specific hate speech. In comparison to unimodal systems, the presented system more accurately captures multimodal signals with subtle nuances, displaying great generalizability. This research helps design scalable, responsible AI systems for safer online spaces.

The experimental findings, characterized by an F1-score of 0.769 and a recall of 0.799, underscore the model's robust capability in identifying racist content, particularly in instances where isolated text or image analysis proves insufficient. Validation using the Hateful Memes dataset confirmed the model's practical utility in contexts necessitating the integrated understanding of multiple modalities. The inherent attention mechanisms within VisualBERT and ViT effectively capture cross-modal dependencies, thereby improving the detection of both overtly and implicitly encoded hate speech.

Despite these advancements, several limitations persist. The constrained linguistic and cultural diversity of the current dataset inherently restricts the model's global applicability. Furthermore, the dynamic nature of cultural expressions, the nuanced use of satire, and complex slang present ongoing challenges to consistent detection. A notable practical constraint is the system's substantial computational resource requirement, which currently impedes its real-time deployment in environments with limited resources.

Future research endeavors should prioritize enhancing efficiency through the development of lightweight models or the application of knowledge distillation techniques to mitigate computational demands. Expanding datasets to encompass multilingual and multicultural samples is critical for improving generalization capabilities. Moreover, integrating factors such as gender identities, regional dialects, and n-gram context patterns holds potential for further refining detection accuracy.

Interdisciplinary collaboration with sociologists and digital rights experts is imperative to ensure that AI systems for content moderation align with established ethical standards. Promoting transparency through open-source datasets and code will encourage community-driven improvements and foster innovation. Strengthening the detection of implicit hate speech can be achieved by incorporating advanced techniques like sentiment analysis, sarcasm recognition, and user context modeling.

In conclusion, while the presented multimodal system represents a significant advancement in hate speech detection, sustained research and collaborative efforts are fundamental to developing adaptable, equitable, and highly effective online content moderation solutions.

Data Availability

The dataset employed in this study originates from the openly accessible Hateful Memes dataset released by Facebook AI. This dataset comprises labeled meme instances that combine both image and text modalities, specifically curated for multimodal hate speech detection tasks. It includes clearly annotated labels indicating whether a meme is hateful or not, enabling robust training and evaluation. The dataset is accessible through the official Facebook AI GitHub repository.

Conflict of Interest

The authors affirm that no conflicts of interest exist.

Funding Source

The authors confirm that no financial support, grants, or other assistance were obtained during the writing of this manuscript.

Authors' Contributions

The research efforts were led by Rewanshu S. Bopapurkar, encompassing a thorough literature review, the development of the model, the integration of VisualBERT and the Vision Transformer (ViT), and the execution of all experiments. His responsibilities also included data analysis, visualization, and preparing the initial draft of the manuscript.

Anupama G. Phakatkar offered comprehensive technical supervision throughout the project, providing guidance on aspects such as model architecture, evaluation methodologies, and experimental design. Her contributions were instrumental in conceptual refinement and the critical revision of the manuscript's intellectual content.

The final version of the manuscript was collaboratively reviewed, edited, and approved for submission by both authors.

Acknowledgements

It is my pleasure to present a Research paper on "Multimodal Deep Learning for The Detection of Racist Content Online". The authors wish to extend their heartfelt thanks to the Department of Computer Engineering at SCTR's Pune Institute of Computer Technology, S.P.P.U, for providing the essential support and resources for this research. Appreciation is extended to our faculty mentors and reviewers for their insightful feedback. I am deeply grateful to Mrs. Anupama G. Phakatkar, my guide for this research paper, whose expert guidance and encouragement were invaluable throughout this study.

References

[1] A. Al-Hassan and H. Al-Dossari, "Detection of Hate Speech in Social Networks: A Survey on Multilingual Corpus", *Proceedings of the 6th International Conference on Computer Science and Information Technology (CS & IT – CSCP)*, Vol.**9**, Issue.**2**, pp.**83–100, 2019**.

[2] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language", *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM)*, Vol.**11**, Issue.**1**, pp.**512–515**, **2017**.

[3] J. S. Malik, G. Pang, and A. van den Hengel, "Deep Learning for Hate Speech Detection: A Comparative Study", *International Journal of Data Science and Analytics*, pp.**1–17, 2024**.

[4] W. Warner and J. Hirschberg, "Detecting Hate Speech on the World Wide Web", *Proceedings of the Second Workshop on Language in Social Media (LSM 2012)*, pp.**19–26, 2012**.

[5] A. Mousa, I. Shahin, A. B. Nassif, and A. Elnagar, "Detection of Arabic offensive language in social media using machine learning models", *Intelligent Systems with Applications*, Vol.**22**, Art. No.**200376**, **2024**.

[6] A. Alabdulwahab, M. A. Haq, and M. Alshehri, "Cyberbullying Detection using Machine Learning and Deep Learning", *International Journal of Advanced Computer Science and Applications*, Vol.**14**, No.**10**, pp.**1–7**, **2023**.

[7] D. C. Asogwa, C. I. Chukwuneke, C. C. Ngene, and G. N. Anigbogu, "Hate Speech Classification Using SVM and Naive BAYES", *IOSR Journal of Mobile Computing & Application (IOSR-JMCA)*, Vol.**9**, Issue.**1**, pp.**27–34, 2022**.

[8] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep Learning for Hate Speech Detection in Tweets", *In the Proceedings of the 26th International Conference on World Wide Web Companion (WWW '17 Companion)*, pp.**759–760**, **2017**.

[9] S. Malik, H. Chopra, and A. Vashishtha, "Capturing racial & gender inequities on social media platforms using machine learning", *EAI Endorsed Transactions on Creative Technologies*, Vol.**9**, Issue.**31**, pp.**4**, **2022**.

[10] P. Kapil, G. Kumari, A. Ekbal, S. Pal, and A. Chatterjee, "HHSD: Hindi hate speech detection leveraging multi-task learning", *IEEE Access*, Vol.**11**, pp.**101460–101473**, **2023**.

[11] M. Mozafari, R. Farahbakhsh, and N. Crespi, "Hate speech detection and racial bias mitigation in social media based on BERT model", *PLOS ONE*, Vol.**15**, Issue.**8**, pp.**0237861**, **2020**.

[12] Marc Djandji, Fady Baly, Wissam Antoun, and Hazem Hajj, "Multi-Task Learning using AraBERT for Offensive Language Detection", *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT4)*, pp.**97–101**, **2020**.

[13] J. A. Benítez-Andrades, Á. González-Jiménez, Á. López-Brea, and J. Aveleira-Mata, "Detecting racism and xenophobia using deep learning models on Twitter data: CNN, LSTM and BERT", *PeerJ Computer Science*, Vol.**8**, pp.**906**, **2022**.

[14] Md Saroar Jahan and Mourad Oussalah, "A systematic review of hate speech automatic detection using natural language processing", *Neurocomputing*, Vol.**546**, pp.**126232**, **2023**.

[15] Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas, "Exploring Hate Speech Detection in Multimodal Publications", *Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp.**1459–1467**, **2020**.

[16] Roy Ka-Wei Lee, Rui Cao, Ziqing Fan, Jing Jiang, and Wen-Haw Chong, "Disentangling Hate in Online Memes", *Proceedings of the 29th ACM International Conference on Multimedia (MM '21 Companion)*, pp.**5138–5147**, **2021**.

[17] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine, "The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes", **2020**.

AUTHORS' PROFILE

Rewanshu S. Bopapurkar completed his bachelor's degree in computer science engineering from Priyadarshini College of Engineering (PCE), Nagpur, and is currently pursuing his Master of Engineering from Pune Institute of Computer Technology (PICT), Savitribai Phule Pune University (SPPU), Pune. His primary research interests include Deep Learning and Machine Learning, with a focus on developing intelligent models for real-world applications such as content moderation and natural language understanding. He is dedicated to advancing research in these domains and contributing to innovative technological solutions.

Anupama G. Phakatkar earned her B.E. in Computer Science and Engineering from MIT, Aurangabad, M.E. in Computer Engineering from Pune Institute of Computer Technology (PICT), Pune, and Ph.D. in Computer Engineering from Savitribai Phule Pune University, Pune. She has been working as an Assistant Professor in the Department of Computer Engineering at PICT, Pune since 2002. She is actively involved in guiding postgraduate students and contributing to research in emerging domains. Her main research work focuses on Machine Learning, Recommendation Systems, Information Retrieval, and Image Processing. She has published research papers in reputed journals and conferences. She has over 20 years of teaching experience and substantial research experience in the field of computer engineering.

