EthosAI@CHiPSAL2026:Hate and Sentiment Understanding in Low-Resource Memes using a Multimodal Approach

Vinayak Bansal[1], **Deepawali Sharma**[2],**Aakash Singh**[1], **Vivek Kumar Singh**[1]
[1]Department of Computer Science, University of Delhi, Delhi, India
[2]School of Computer Science Engineering and Technology (SCSET),
Bennett University, Greater Noida, India
bansal.vinayak2002@gmail.com, deepawali121@gmail.com,
asingh@cs.du.ac.in, vivek@cs.du.ac.in

Abstract

Memes have become a popular way for people to share opinions and emotions on social media, but they are also often used to spread hate and negative sentiments. In this paper, we present our multimodal approach to the CHiPSAL 2026 shared task on multimodal hate and sentiment detection in Nepali memes, which includes two subtasks: hate detection and sentiment analysis. Since memes usually combine both text and images, we first experimented with different unimodal models for text and images separately. After identifying the top two best-performing text and image models, combined them using different fusion techniques. The results show that multimodal models outperform unimodal ones, highlighting that both textual and visual information are important for understanding the context of memes. The multimodal model, which combines sentence-transformers/LaBSE for text and ResNet-18 for image using weighted Fusion technique, achieved a macro F1 score of 0.6614 for Subtask A and sentence-Transformers/LaBSE for text and deit-Base for image using simple Fusion technique, achieved a macro F1 score of 0.4839 for SubTask B, on the test dataset.

Keywords: Hate speech Detection, Low-Resource language, Meme Classification, Multimodal Data, Sentiment analysis

1. Introduction

In the era of social media, memes have become a powerful medium for expressing public sentiment within online communities. For teenagers in particular, memes play a central role in communication.According to a report by YPulse (2019)[1], 75% of individuals between the ages of 13 and 36 share memes. Memes often serve as a means of spreading joy and happiness. However, some people misuse memes to spread hate, especially against vulnerable communities (Mukhtar et al., 2024). Such hateful content can contribute to social unrest and may even lead to violence (Parihar et al., 2021; Bhandari et al., 2023; Singh et al., 2026). Several studies (Kiela et al., 2020; Arya et al., 2024; Das et al., 2020) have focused on detecting hate speech in English- language memes. In contrast, only a limited number of studies (Singh et al., 2024; Nagaraju and Shashirekha, 2025), address hate detection in memes in low-resource languages, and for some languages, no prior work exists. Nepali is one such low-resource language(Sharma et al., 2025). Nepali is the official and most widely spoken language of Nepal, with approximately 19 million native speakers worldwide[2].

The CHiPSAL 2026 task (Thapa et al., 2026) is introduced to specifically focus on detecting hate in this low-resource language setting. The task is divided into two subtasks: (A) identifying hate in memes and (B) identifying sentiment expressed in memes. This task aims to bridge the gap in hate detection and sentiment analysis for the Nepali language. Preventing the spread of hateful content on social media is essential for creating safer and healthier digital platforms for people from all backgrounds.

In this paper, we analyze multimodal architectures and show that both textual and visual information are essential for accurately identifying hate and sentiment in memes. Since memes combine images and text to convey meaning, understanding their full context requires integrating both modalities. To examine this, we evaluate several state-of-the-art (SOTA) models for text and image processing, apply different fusion strategies, and compare their performance. The results demonstrate that multimodal approaches outperform unimodal models for both subtasks.

The subsequent sections of this paper are structured as follows. Section 2 reviews the relevant literature on hateful meme detection and sentiment analysis. Section 3 presents the dataset description. Section 4 describes implemented unimodal and multimodal architectures. Section 5 outlines the experimental setup, while Section 6 discusses the results. Finally, Section 7 concludes the paper

[1]YPulse report 3 Stats That Show What Memes Mean to Gen Z & Millennials https://www.ypulse.com/article/2019/03/05/3-stats-that-show-what-memes-mean-to-gen-z-millennials/

[2]"Nepali language," Wikipedia, last modified 2026, https://en.wikipedia.org/wiki/Nepali_language.

Proceedings of the Second workshop on Challenges in Processing South Asian Languages (CHiPSAL2026) @ LREC, pages 316–323
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

by summarizing our key contributions and insights.

2. Related Work

With the rapid growth of social media platforms, the need to monitor and detect harmful online content has become increasingly important. Early research in hate speech detection has focused primarily on text-based analysis; studies utilized natural language processing for the identification of hate speech in comments, posts, etc. (Schmidt and Wiegand, 2017; Davidson et al., 2017; Fortuna and Nunes, 2018; Sharma et al., 2024). Studies have been done on the emotional connection of speech, which laid the foundation for sentiment analysis (Plaza-Del-Arco et al., 2021; Nandwani and Verma, 2021; Sadat et al., 2022; Singh et al., 2025). These works demonstrate the interlink between hate speech and sentiment analysis, since negative emotions frequently precede hate speech.

With the rapid rise of internet memes, there has been a shift toward multimodal analysis, as memes usually combine both images and text to convey meaning. Understanding them through only one modality is often not enough. The Hateful Memes Challenge (Kiela et al., 2020) was one of the early efforts that highlighted this issue, showing that memes require models to jointly interpret visual and textual information due to their complex and subtle nature. Later developments such as CLIP further emphasized the importance of aligning images and text representations for better cross-modal understanding (Arya et al., 2024) . Other studies have also explored contextual meme analysis for hate detection, though many of these efforts remain preliminary (Das et al., 2020). While most early research focused on English memes, recent work has started paying attention to code-mixed content, especially Hindi–English memes. These studies suggest that for low-resource languages to benefit from multimodal models, linguistic diversity and language mixing must be considered (Singh et al., 2024). Therefore, there is a need to develop resources that include diverse Indian languages, helping create safer and more inclusive social media environments (Nagaraju and Shashirekha, 2025).

The integration of Large Language Models (LLMs) represents a significant paradigm shift in computational social science, offering advanced reasoning capabilities for complex content analysis (Thapa et al., 2025a). Current frameworks leverage prompt-based approaches and retrieval-augmented methods to handle the scarcity of data and the linguistic intricacies of code-mixed, low-resource memes (Thapa et al., 2025b,c). Despite these advancements, a comprehensive survey of South Asian languages reveals that no research in Nepali specifically concerning meme-based hate and sentiment detection (Sharma et al., 2025). This gap highlights the need for a multimodal dataset for hate detection in the Nepali language. The CHiPSAL2026 (Thapa et al., 2026) shared task addresses this issue by introducing a dataset for detecting hate and analyzing sentiment in low-resource Nepali memes . In this paper, we analyze multimodal models to detect the hate and sentiment analysis within the Nepali memes.

3. Dataset Description

The dataset used in this study is provided by the shared task on Multimodal Hate and Sentiment Understanding in Low-Resource Memes (Thapa et al., 2026) in the CHiPSAL 2026 workshop (Sarveswaran et al., 2026). The task consists of two subtasks. SubTask A is for the identification of Hateful Memes in Nepali. SubTask B is for sentimental analysis of Hateful Memes in Nepali.

3.1. SubTask A

This Subtask is for the identification of Hateful Memes in Nepali. It consists of 1068 memes for training (720 of Hate and 348 of Non Hate), 133 memes for validation (98 of Hate and 35 of Non Hate), and 134 memes for testing, as shown in Table 1.

Table 1: Dataset distribution for SubTask A

Class	Train	Val	Test
Hate	720	98	-
Non-Hate	348	35	-
Total	1068	133	134

3.2. SubTask B

This SubTask is for sentimental analysis of Hateful Memes in Nepali. It consists of 1061 memes for training (341 Negative, 473 Neutral, and 247 Positive), 133 memes for validation (39 Negative, 65 Neutral, and 29 Positive), and 133 memes for testing, as shown in Table 2.

Table 2: Dataset distribution for SubTask B

Sentiment	Train	Val	Test
Negative	341	39	-
Neutral	473	65	-
Positive	247	29	-
Total	1061	133	133

4. Methodology

The dataset consists of memes comprising two modalities: image and text. In this section, we in-

troduce the methodology used in the study to handle both modalities and their fusion. The problem statement is defined as follows:

$$I \in \mathbb{R}^{3\times224\times224}$$
$$T \in \mathbb{R}^{1\times256}$$

where I is a 3-channel image of height 224 and width 224, and T is a 256-dimensional embedding representation.

Subtask A is defined as:

$$y = \begin{cases} 0 & \text{Hate} \\ 1 & \text{Non-Hate} \end{cases}$$

The classifier HC_1 is defined as:

$$HC_1 : (I, T) \rightarrow y, \quad y \in \{0, 1\}$$

Subtask B is defined as:

$$y = \begin{cases} 0 & \text{Negative Sentiment} \\ 1 & \text{Neutral Sentiment} \\ 2 & \text{Positive Sentiment} \end{cases}$$

The classifier HC_2 is defined as:

$$HC_2 : (I, T) \rightarrow y, \quad y \in \{0, 1, 2\}$$

4.1. Unimodal Model

In the unimodal approach, the classifier is designed to process output given only one of the two modalities (e.g., either image or text). This individual contribution serves as a baseline for each modality.

4.1.1. Image

In this section, we summarize the image model analysis. We test models $I_1, I_2, \ldots, I_{10}$ as shown in the image model Table 3. We select the top two individual image models, I_1 and I_2, to be tested further for fusion, as shown in section 4.2.

Table 3: Abbreviations of Image Model Names

Tag	Model Name
I1	Deit-Base
I2	Resnet18
I3	Beit-Base
I4	Resnet50
I5	Vgg16
I6	Efficientnet-B4
I7	Efficientnet-B0
I8	Convnext-Base
I9	Vit-Large
I10	Vit-Base

1. I_1**: DeiT-Base** is a transformer-based model introduced by (Touvron et al., 2021). This model follows the Vision Transformer pipeline with a key modification: knowledge distillation through a distillation token. In this model, an additional distillation token is introduced, which learns from a CNN teacher model. Here, the model is supervised by both the ground truth and the teacher predictions. This allows the model to perform better on smaller datasets where data is limited for parameter tuning.

2. I_2**: ResNet-18** is a convolution-based (CNN) model introduced by (He et al., 2016). This model handles the degradation problems in CNNs with the help of the residual connection method, which is mathematically expressed as:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \{\mathbf{W}_i\}) + \mathbf{x}$$

where $\mathbf{x}$ is the input and $\mathcal{F}(\mathbf{x}, \{\mathbf{W}_i\})$ represents the output of the residual mapping. This residual network helps reduce the vanishing gradient problem, enabling the model to efficiently handle degradation without additional parameters, which helps the model train and perform better on smaller datasets.

4.1.2. Text

In this section, we summarize the text extraction from images and the text models. For text extraction, we use EasyOCR and the Qwen model. EasyOCR text was provided by CHiPSAL; we also use Qwen2-VL-7B-Instruct to generate image descriptions alongside the image OCR. As neither EasyOCR nor Qwen OCR is fully reliable on its own, we extracted text from both to improve overall OCR accuracy. Beyond OCR, the Qwen model is used to generate further details of the image. Listings 1 and 2 provide the prompt templates for Subtask A and Subtask B, respectively.

```
messages = [
    {
        "role": "user",
        "content": [
            {"type": "image", "image": image_path},
            {"type": "text", "text": (
                "Task: 1. OCR Nepali/English text. "
                "2. Describe visual/textual meaning. "
                "3. Classify: 'Hate' or 'Not Hate'. "
                "4. Provide justification.\n"
                "Output: OCR_TEXT, MEANING, LABEL, JUSTIFICATION"
            )}
        ]
    }
]
```

Listing 1: Prompt Template for Subtask A: Hate Speech Detection

```
messages = [
    {
        "role": "user",
        "content": [
            {"type": "image", "image": image_path},
            {"type": "text", "text": (
                "Task: 1. OCR Nepali/English text. "
                "2. Describe meme meaning. "
                "3. Determine Sentiment: Positive, Neutral, or Negative. "
                "4. Explain sentiment reason.\n"
                "Output: OCR_TEXT, MEANING, SENTIMENT, SENTIMENT_REASON"
            )}
        ]
    }
]
```

Listing 2: Prompt Template for Subtask B: Sentiment Analysis

Table 4: Abbreviations of Text Model Names

Tag	Model Name
T1	Sentence-Transformers/LaBSE
T2	Bert-Base-Multilingual-Cased
T3	Distilbert-Base-Multilingual-Cased
T4	Sentence-Transformers/Paraphrase-Multilingual-Mpnet-Base-V2
T5	Sentence-Transformers/Paraphrase-Multilingual-MiniLM-L12-V2
T6	Xlm-Roberta-Base
T7	Ai4Bharat/Indic-Bert
T8	Microsoft/Mdeberta-V3-Base
T9	Google/Muril-Base-Cased
T10	Google/Canine-C

We test models $T_1, T_2, \ldots, T_{10}$ as shown in the text model Table 4. We select the top two individual text models, T_1 and T_2, to be tested further for fusion in section 4.2.

- T_1**: Sentence-Transformers/LaBSE** is a multilingual sentence embedding model introduced by (Feng et al., 2022). This model is based on Multilingual BERT (mBERT) with 12 transformer layers. The model uses a shared semantic vector space for different languages; it provides a similar vector space for both the Nepali text and its English translation. This helps in maintaining the complete text constructed in the preceding sections within a similar vector space, allowing the model to better understand the context of the meme.
- T_2**: Bert-Base-Multilingual-Cased (mBERT)** is a multilingual language model trained on Wikipedia data from 104 languages. This model aligns Nepali and English using cross-lingual alignment, which happens implicitly through the shared training process. The model has 12 transformer layers, and its cross-lingual alignment helps it work on unlabeled Nepali data based on fine-tuning on English.

4.2. Multimodal Model

In the previous subsection 4.1, we established a baseline for how image and text performed individually. Based on a single modality, meme content cannot always be reliably classified as hate or non-hate, nor can we perform sentiment analysis; therefore, we perform different types of fusion on the top two performing image models (I_1, I_2) and the top two performing text models (T_1, T_2).

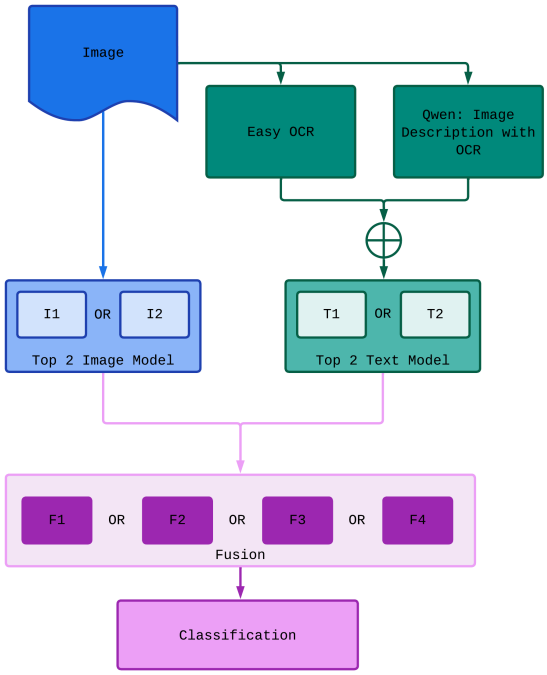

Figure 1: Proposed Multimodal Design

Figure 1 presents our proposed multimodal design. In the fusion of the two modalities, the modalities are first projected into a common dimension. Let $\mathbf{t} \in \mathbb{R}^{d_t}$ denote the text feature vector and $\mathbf{v} \in \mathbb{R}^{d_v}$ denote the image feature vector.

$$\tilde{\mathbf{t}} = W_t \mathbf{t} + b_t, \quad W_t \in \mathbb{R}^{256 \times d_t}, \quad b_t \in \mathbb{R}^{256} \tag{1}$$

$$\tilde{\mathbf{v}} = W_v \mathbf{v} + b_v, \quad W_v \in \mathbb{R}^{256 \times d_v}, \quad b_v \in \mathbb{R}^{256} \tag{2}$$

Such that

$$\tilde{\mathbf{t}}, \tilde{\mathbf{v}} \in \mathbb{R}^{256}.$$

- F_1: **Simple Fusion** In this fusion, the projected vectors $\tilde{\mathbf{t}}$ and $\tilde{\mathbf{v}}$ are concatenated and passed through a fully connected layer for classification.

$$\mathbf{x} = [\tilde{\mathbf{v}}; \tilde{\mathbf{t}}] \in \mathbb{R}^{512} \tag{3}$$

  The classifier is defined as:

$$\hat{y} = \sigma(W_c \mathbf{x} + b_c) \tag{4}$$

  where $W_c \in \mathbb{R}^{1\times512}, b_c \in \mathbb{R}$.

- F_2: **Weighted Fusion** In weighted fusion, we introduce learnable scalar weights α and β:

$$\mathbf{x} = \alpha\tilde{\mathbf{t}} + \beta\tilde{\mathbf{v}}, \quad \mathbf{x} \in \mathbb{R}^{256} \tag{5}$$

  where $\alpha, \beta \in \mathbb{R}$ are trainable parameters. The classifier is defined as:

$$\hat{y} = \sigma(W_c \mathbf{x} + b_c), \quad W_c \in \mathbb{R}^{1\times256}. \tag{6}$$

- F_3: **Cross-Attention Fusion** We employ a single-head cross-attention mechanism with text features as the query and image features as the key and value:

$$Q = \tilde{\mathbf{t}}, \quad K = \tilde{\mathbf{v}}, \quad V = \tilde{\mathbf{v}} \tag{7}$$

  The attention mechanism is defined as:

$$\mathbf{x} = \text{softmax}\left(\frac{QK^\top}{\sqrt{256}}\right)V \tag{8}$$

  where $\mathbf{x} \in \mathbb{R}^{256}$. The final prediction is:

$$\hat{y} = \sigma(W_c \mathbf{x} + b_c), \quad W_c \in \mathbb{R}^{1\times256}. \tag{9}$$

- F_4: **Element-Wise Fusion** Element-wise fusion is performed via direct addition:

$$\mathbf{x} = \tilde{\mathbf{t}} + \tilde{\mathbf{v}}, \quad \mathbf{x} \in \mathbb{R}^{256} \tag{10}$$

  The classifier output is:

$$\hat{y} = \sigma(W_c \mathbf{x} + b_c), \quad W_c \in \mathbb{R}^{1\times256}. \tag{11}$$

5. Experimental Setup

In this section, we provide the details of the experimental setup used in Section 4 for model training. All experiments are performed using an NVIDIA RTX 5000 Ada (32 GB GPU VRAM) paired with 256 GB system RAM and an Intel Xeon W7-2475X processor as the hardware platform. The primary software libraries employed for this study are PyTorch, Pandas, and OpenCV.

Table 5 summarizes the hyperparameters used in our experiments. All models are trained using a batch size of 8 to balance performance and memory usage. Training is performed for a maximum of 50 epochs. To avoid overfitting, early stopping is applied based on the validation loss, with a patience value of 5. We use ReduceLROnPlateau as the learning rate scheduler, with a reduction factor of 0.5 and a patience of 2. For Subtask A, BCEWithLogitsLoss is used as the loss function. For Subtask B, CrossEntropyLoss is applied. The AdamW optimizer is used to update the model parameters during training.

Table 5: Training Hyperparameters

Hyperparameter	Value
Batch Size	8
Epochs	50
Early Stopping Patience	5
Learning Rate	1×10^{-5}
Weight Decay	1×10^{-4}
Optimizer	AdamW
Scheduler	ReduceLROnPlateau
Scheduler Factor	0.5
Scheduler Patience	2
Loss Function	BCEWithLogitsLoss / CrossEntropyLoss

6. Results

Table 6 presents the performance of unimodal models on the validation dataset. T_1 and T_2 are the best-performing text models. T_1 achieves the best validation macro F1 score of 0.605 for Subtask A and 0.4443 for Subtask B, followed by T_2 with a validation macro F1 score of 0.597 for Subtask A and 0.4342 for Subtask B. For image models, I_1 and I_2 are the top performers. I_1 achieves the best validation macro F1 score of 0.5735 for Subtask A and 0.4606 for Subtask B, followed by I_2 with a validation macro F1 score of 0.5929 for Subtask A and 0.4525 for Subtask B.

Table 7 presents the results for the multimodal models on the validation dataset. The results indicate that for Subtask A, the top-performing multimodal model is F_3 using T_1 and I_1 (macro F1 of 0.6568), followed by F_3 using T_1 and I_2 (macro F1 of 0.6502), and F_2 using T_1 and I_2 (macro F1 of 0.6425). For Subtask B, the top-performing multimodal model is F_2 using T_1 and I_1 (macro F1 of 0.4755), followed by F_3 using T_1 and I_1 (macro F1 of 0.4747), and F_1 using T_1 and I_1 (macro F1 of 0.4613).

We select the top three performing multimodal models on the validation dataset for evaluation on the final test dataset. Table 8 provides the results for the top three models on the test dataset; notably, F_2 with the T_1 and I_2 model performs the best with a macro F1 of **0.6614**. Table 8 provides the results

for the test dataset, where the F_1 fusion on the T_1 and I_1 model performs the best with a macro F1 of **0.4839**.

We attempted to detect hate speech and analyze meme sentiment based on the direct output of the Qwen model; however, the model was unable to successfully classify all the memes. Consequently, we were required to pass the Qwen descriptions along with the OCR to the text models. A closer look at the results reveals a bottleneck caused by the count of randomly initialized learnable parameters introduced in the fusion layers. The results make it evident that the model is unable to completely train these randomly initialized fusion parameters.

Simple and weighted fusion methods are performing better than cross-attention-based fusion. For Subtask A, weighted fusion performs the best; however, for Subtask B, which has a smaller dataset size per class, simple fusion is performing better than even weighted fusion. These results demonstrate that for a smaller dataset, the introduction of untrained parameters results in the addition of more noise. We also observe that element-wise fusion leads to a loss of essential information, resulting in lower performance than unimodal approaches. The model is unable to retain the unimodal information in that configuration. Hence, simple concatenation, providing a balance between dataset size and the introduction of new parameters, yields the best performing results.

Table 6: Performance Comparison of Unimodal Models (Validation Dataset)

Mod.	Tag	Model Name	Sub A (F1)	Sub B (F1)
Text (Unimodal)				
Text	**T1**	**Sentence-Transformers/ LaBSE**	**0.605**	**0.4443**
Text	**T2**	**Bert-Base-Multilingual-Cased**	**0.597**	**0.4342**
Text	T3	DistilBert-Base-Multilingual-Cased	0.5535	0.4241
Text	T4	Sent.-Trans./ Para.-Multiling.-Mpnet-Base-V2	0.562	0.4057
Text	T5	Sent.-Trans./ Para.-Multiling.-MiniLM-L12-V2	0.5793	0.3921
Text	T6	XLM-Roberta-Base	0.4242	0.2189
Text	T7	Ai4Bharat/ Indic-Bert	0.4242	0.324
Text	T8	Microsoft/ MDeBerta-V3-Base	0.4242	0.2189
Text	T9	Google/ MuRIL-Base-Cased	0.4242	0.2189
Text	T10	Google/ Canine-C	0.4242	0.2189
Image (Unimodal)				
Image	**I1**	**DeiT-Base**	**0.5735**	**0.4606**
Image	**I2**	**ResNet18**	**0.5929**	**0.4525**
Image	I3	BeiT-Base	0.581	0.4388
Image	I4	ResNet50	0.5822	0.413
Image	I5	VGG16	0.5794	0.3504
Image	I6	EfficientNet_B4	0.5694	0.4485
Image	I7	EfficientNet_B0	0.5427	0.3997
Image	I8	ConvNext_Base	0.5393	0.4371
Image	I9	ViT-Large	0.5654	0.3946
Image	I10	ViT-Base	0.5347	0.4149

Table 7: Performance Comparison of Multimodal Models on Subtasks A and B (Validation Dataset)

Text	Image	Fusion	SubTask A Macro F1	SubTask B Macro F1
T_1	I_1	F_4	0.5186	0.4579
T_1	I_1	F_3	**0.6568**	**0.4747**
T_1	I_1	F_1	0.6171	**0.4613**
T_1	I_1	F_2	0.6252	**0.4755**
T_1	I_2	F_4	0.5395	0.4336
T_1	I_2	F_3	**0.6502**	0.4546
T_1	I_2	F_1	0.6094	0.4365
T_1	I_2	F_2	**0.6425**	0.3946
T_2	I_1	F_4	0.5423	0.3924
T_2	I_1	F_3	0.5574	0.4201
T_2	I_1	F_1	0.5130	0.4275
T_2	I_1	F_2	0.5405	0.3793
T_2	I_2	F_4	0.6094	0.4337
T_2	I_2	F_3	0.5585	0.4301
T_2	I_2	F_1	0.5485	0.3847
T_2	I_2	F_2	0.5970	0.4089

Note: F_1: Simple Fusion, F_2: Weighted fusion, F_3: Element-Wise fusion, F_4: Cross-Attention Fusion.

Table 8: Testing Dataset Results for Multimodal Fusion

Text	Image	Fusion	Macro F1	A	P	R
Subtask A						
T1	I1	F3	0.6229	0.6791	0.6324	0.6275
T1	I2	F3	0.6080	0.6866	0.6321	0.6040
T1	**I2**	**F2**	**0.6614**	**0.7164**	**0.6743**	**0.6553**
Subtask B						
T1	I1	F3	0.4336	0.4586	0.4611	0.4349
T1	**I1**	**F1**	**0.4839**	**0.4962**	**0.4824**	**0.4857**
T1	I1	F2	0.4670	0.5038	0.4975	0.4583

A: Accuracy; **P**: Precision (Macro); **R**: Recall (Macro).

7. Conclusion

Detecting hate and sentiment in memes is important for building safe and inclusive online spaces. Memes spread ideas quickly and often hide harmful intent, which makes automatic detection necessary, especially for low-resource languages like Nepali, where data and tools are limited. In this paper, we presented our multimodal approach to the CHiPSAL 2026 shared task on multimodal hate and sentiment detection in Nepali memes. Various text-based and image-based unimodal models are implemented and evaluated. The top-performing models from both modalities are then combined using different fusion techniques. The multimodal approach outperforms the unimodal approach. The best-performing model combines T_1: Sentence-Transformers/LaBSE for text and I_2: ResNet-18 for image using F_2: Weighted Fusion technique, and the reported macro F1-scores of 0.6614 for Subtask A and T_1: Sentence-Transformers/LaBSE for text and I_1: Deit-Base for image using F_1: Simple Fusion technique, and the reported macro F1-scores of $0.0.4839$ for SubTask B, respectively. This work serves as a valuable baseline and resource for the research community and supports the development of more robust and effective computational tools for multimodal content understanding in low-resource languages.

Code Availability

The code is available at the following link https://github.com/vinayakbansal-2002/EthosAI-Chipsal

Acknowledgements

We acknowledge the financial assistance and academic support provided to one of the authors under the Visvesvaraya PhD Scheme by the Ministry of Electronics and Information Technology, Government of India, in the form of the Visvesvaraya Fellowship.

8. Bibliographical References

Greeshma Arya, Mohammad Kamrul Hasan, Ashish Bagwari, Nurhizam Safie, Shayla Islam, Fatima Rayan Awad Ahmed, Aaishani De, Muhammad Attique Khan, and Taher M Ghazal. 2024. Multimodal hate speech detection in memes using contrastive language-image pre-training. *IEEe Access*, 12:22359–22375.

Aashish Bhandari, Siddhant B Shah, Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. Crisishatemm: Multimodal analysis of directed and undirected hate speech in text-embedded images from russia-ukraine conflict. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1994–2003.

Abhishek Das, Japsimar Singh Wahi, and Siyao Li. 2020. Detecting hate speech in multi-modal memes. *arXiv preprint arXiv:2012.14891*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 878–891.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *Acm Computing Surveys (Csur)*, 51(4):1–30.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624.

Shahira Mukhtar, Qurat Ul Ain Ayyaz, Sadaf Khan, Atiya Muhammad Nawaz Bhopali, Muhammad Khalid Mehmood Sajid, Allah Wasaya Babbar, et al. 2024. Memes in the digital age: A sociolinguistic examination of cultural expressions and communicative practices across border. *Educational Administration: Theory and Practice*, 30(6):1443–1455.

Rachana Nagaraju and Hosahalli Lakshmaiah Shashirekha. 2025. Towards safer social media: Multimodal hate speech detection in memes across diverse indian languages.

Pansy Nandwani and Rupali Verma. 2021. A review on sentiment analysis and emotion detection from text. *Social network analysis and mining*, 11(1):81.

Anil Singh Parihar, Surendrabikram Thapa, and Sushruti Mishra. 2021. Hate speech detection using natural language processing: Applications and challenges. In *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 1302–1308. IEEE.

Flor Miriam Plaza-Del-Arco, M Dolores Molina-González, L Alfonso Ureña-López, and María Teresa Martín-Valdivia. 2021. A multi-task learning approach to hate speech detection leveraging sentiment analysis. *IEEE Access*, 9:112478–112489.

Anwar Sadat, Herman Lawelai, and Ansar Suherman. 2022. Sentiment analysis on social media: Hate speech to the government on twitter. *PRAJA: Jurnal Ilmiah Pemerintahan*, 10(1):69–76.

Kengatharaiyer Sarveswaran, Surendrabikram Thapa, Ashwini Vaidya, Tafseer Ahmed Khan, and Bal Krishna Bal. 2026. Findings of the second workshop on challenges in processing south asian languages (chipsal 2026). In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10.

Deepawali Sharma, Tanusree Nath, Vedika Gupta, and Vivek Kumar Singh. 2025. Hate speech detection research in south asian languages: A survey of tasks, datasets and methods. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 24(3):1–44.

Deepawali Sharma, Aakash Singh, and Vivek Kumar Singh. 2024. Thar-targeted hate speech against religion: A high-quality hindi-english code-mixed dataset with the application of deep learning models for automatic detection. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Aakash Singh, Vinayak Bansal, Muskan Saini, Deepawali Sharma, and Vivek Kumar Singh. 2026. Safeplay-x: A comprehensive gameplay video dataset for violence detection with explainable deep learning applications. *Expert Systems with Applications*, page 131724.

Aakash Singh, Deepawali Sharma, and Vivek Kumar Singh. 2024. Mimic: Misogyny identification in multimodal internet content in hindi-english code-mixed language. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Aakash Singh, Deepawali Sharma, and Vivek Kumar Singh. 2025. Emogif: A multimodal approach to detect emotional support in animated gifs. *IEEE Transactions on Computational Social Systems*.

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Surabhi Adhikari, Hariram Veeramani, Mehwish Nasim, and Usman Naseem. 2025a. Large language models (llm) in computational social science: prospects, current state, and challenges. *Social Network Analysis and Mining*, 15(1):4.

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Kritesh Rauniyar, Laxmi Thapa, Surabhi Adhikari, Kristina T. Johnson, Kengatharaiyer Sarveswaran, Bal Krishna Bal, and Usman Naseem. 2026. Multimodal hate and sentiment understanding in low-resource text-embedded images for online safety and digital well-being. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*.

Surendrabikram Thapa, Hariram Veeramani, Liang Hu, Qi Zhang, Wei Wang, and Usman Naseem. 2025b. A multimodal prompt-based framework for analyzing code-mixed and low-resource memes. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 1913–1923.

Surendrabikram Thapa, Hariram Veeramani, Imran Razzak, Roy Ka-Wei Lee, and Usman Naseem. 2025c. Cross platform multimodal retrieval augmented distillation for code-switched content understanding. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2042–2051.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. 2021. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10347–10357. PMLR.