eGrantha.ai@CHiPSAL 2026: Stochastic Image Captioning for Robust Hate Speech Detection in Low-Resource Nepali Memes

Anish Thapaliya
Institute of Science and Technology, Tribhuvan University
Kathmandu, Nepal
anish.805522@sms.tu.edu.np

Abstract

This paper presents a system for hate speech detection in low-resource Nepali memes, submitted as part of Subtask A of the Shared Task on Multimodal Understanding at CHiPSAL 2026. Detecting hateful memes is particularly challenging due to the combination of images, text, and emojis used to portray humor, satire, or sociopolitical commentary, as well as the low-resource nature of the Nepali language. We investigate a range of unimodal and multimodal modeling strategies, including text-only, vision-text, and caption-based approaches. For caption generation, the Gemini family of models (Gemini 2.X and Gemini 3.X) was used to produce contextually rich captions, which are publicly released as **NeMeme-CAP** on Hugging Face. Caption-based modeling leverages stochastic caption augmentation to address class imbalance and Test-Time Augmentation (TTA) to reduce prediction variance and improve model robustness. The best-performing system fine-tunes an encoder-only transformer model, RoBERTa-base, on the generated captions, achieving **third** place on the official leaderboard with a macro-averaged F1-score of **0.7397**. The code is publicly available at https://github.com/thapaliya123/LREC-CHiPSAL-2026.

Keywords: multimodal hate speech detection, Nepali memes, low-resource Nepali NLP, vision-language model, stochastic caption augmentation

1. Introduction

Social media platforms such as Facebook, Instagram, Twitter (currently X) have popularized memes as a form of multimodal communication that combines images, text, and emojis to portray humor, satire, or sociopolitical commentary (as shown in Figure 1). Despite the popularity of memes, their multimodal nature facilitates the proliferation of hate speech targeting individuals and communities on the basis of race, gender, religion, or nationality (Kiela et al., 2020). Online hate speech has become a growing concern, prompting increasing research efforts toward automated detection methods (Parihar et al., 2021). However, since it is challenging to manually keep up with everything being shared, prior research and shared tasks have primarily focused on developing automated systems for high-resource languages such as English (Thapa et al., 2025a); comparatively limited attention has been given to low-resource languages such as Nepali, a Devanagari-script language spoken by more than 20 million people worldwide.

To address this gap, a shared task (Thapa et al., 2026) was organized as part of the CHiPSAL 2026 workshop (Sarveswaran et al., 2026). Subtask A was framed as a binary classification problem, where each meme was labeled as either hate or non-hate. The official leaderboard ranking was based on macro-averaged F1-score as the primary metric.

To address this task, several transformer-based strategies were explored. The best-performing solution uses the **Gemini model variants** (Comanici et al., 2025; Google DeepMind, 2025b,a) to generate context-aware image captions, leveraging their strong image understanding capabilities. The RoBERTa-base (Liu et al., 2019) model was fine-tuned on the generated captions for binary classification (non-hate or hate). Additionally, Test-Time Augmentation (TTA) (Moshkov et al., 2020) was applied, which further improved the system performance.

Figure 1: Examples of meme data.

The main contributions are as follows:

- A systematic comparison of pre-trained transformer encoders fine-tuned on Gemini-generated contextually rich captions for meme hate speech detection.
- A demonstration of the efficacy of a stochastic caption regeneration strategy for class-imbalance mitigation, where TTA via caption diversity further enhanced the test F1-score at no additional training cost.

Proceedings of the Second workshop on Challenges in Processing South Asian Languages (CHiPSAL2026) @ LREC, pages 308–315

Class	Train	Validation	Test
non-hate	348	35	-
hate	720	98	-
Total	**1068**	**133**	**134**

Table 1: Class-wise distribution of samples for Subtask A.

- **NeMeme-CAP** [1], a publicly released dataset of Gemini-generated captions for Nepali memes.
- The system ranked **third** on the official CHiPSAL 2026 leaderboard with a macro F1-score of **0.7397**.

,

2. Dataset and Task

The dataset from Subtask A of the Shared Task (Thapa et al., 2026) was utilized, where each meme was categorized as either hate or non-hate. The dataset was compiled from prior works, including NeMeme (Thapa et al., 2025b), ENeMeme (Thapa et al., 2025c), and CrisisHateMM (Bhandari et al., 2023).

The class-wise distribution of training, validation, and test splits is reported in Table 1.

3. Methodology

The system includes a three-stage pipeline: **(i)** image caption generation; **(ii)** stochastic caption augmentation; and **(iii)** transformer fine-tuning.

Formally, given a meme image I and a structured prompt P that instructs the model to describe visible content, the vision-language model (VLM), acting as a caption generator G, produces context-aware image captions in natural language:

$$\hat{y} = C\big(G(I,\, P)\big) \tag{1}$$

where C denotes the downstream text classifier and $\hat{y}$ the predicted hate-speech label. The caption $T = G(I, P)$ is the only textual input to the text classifier, converting the multimodal problem into a standard text classification problem and eliminating the need for a separate OCR pipeline, cross-modal fine-tuning, or Nepali-specific vision–language pretraining.

Figure 2 illustrates the overall training and inference pipeline of the experiment.

3.1. Image Caption Generation

The **Gemini 3 Flash** (Comanici et al., 2025; Google DeepMind, 2025a) was accessed via the Google Generative AI API using a structured prompt to generate context-aware image captions, which are publicly released as **NeMeme-CAP** on Hugging Face. The full structured prompt is provided in Appendix A.

[1] https://huggingface.co/datasets/Anish/nepali-meme-captions

Class	Before Aug.	After Aug.
non-hate	348	696
hate	720	720
Total	**1068**	**1416**

Table 2: Class-wise distributions of the training set before and after applying stochastic caption augmentation (Aug.).

3.2. Stochastic Caption Augmentation

To mitigate class imbalance observed in the training set (see Table 1), a data augmentation strategy leveraging the stochastic nature of VLM decoding was explored, in which Gemini 3 Flash was requeried at **temperature 1.0**, producing diverse descriptive captions of the same meme image. Concretely, non-hate class training images received one additional caption, doubling minority-class coverage and yielding a more balanced training distribution (see Table 2).

At inference time, the same stochastic decoding mechanism was applied as **TTA**. A diverse family of Gemini model variants – **Gemini 2.5 Flash**, **Gemini 2.5 Pro** (Comanici et al., 2025), **Gemini 3 Flash** (Google DeepMind, 2025a), and **Gemini 3 Pro** (Google DeepMind, 2025b) – generated five independent captions for each test image, and each sample was passed through the trained classifier separately. The resulting class-probability distributions were averaged before argmax was taken to produce the final prediction (see Figure 2).

3.3. Transformer Fine-Tuning

Caption classification was framed as standard sequence classification employing transformer architectures (Vaswani et al., 2017). The generated caption T was tokenized and encoded by a pretrained transformer encoder, yielding a contextualized `[CLS]` token representation:

$$\mathbf{h} = \mathrm{Encoder}(T) \in \mathbb{R}^d, \tag{2}$$

where d denotes the hidden dimension of the encoder.

The `[CLS]` token representation was projected to a binary hate-speech prediction through a linear classification head:

$$\hat{y} = \mathrm{softmax}(\mathbf{W}\mathbf{h} + \mathbf{b}), \tag{3}$$

where $\mathbf{W} \in \mathbb{R}^{2\times d}$ is the learned projection matrix and $\mathbf{b} \in \mathbb{R}^2$ is the bias term.

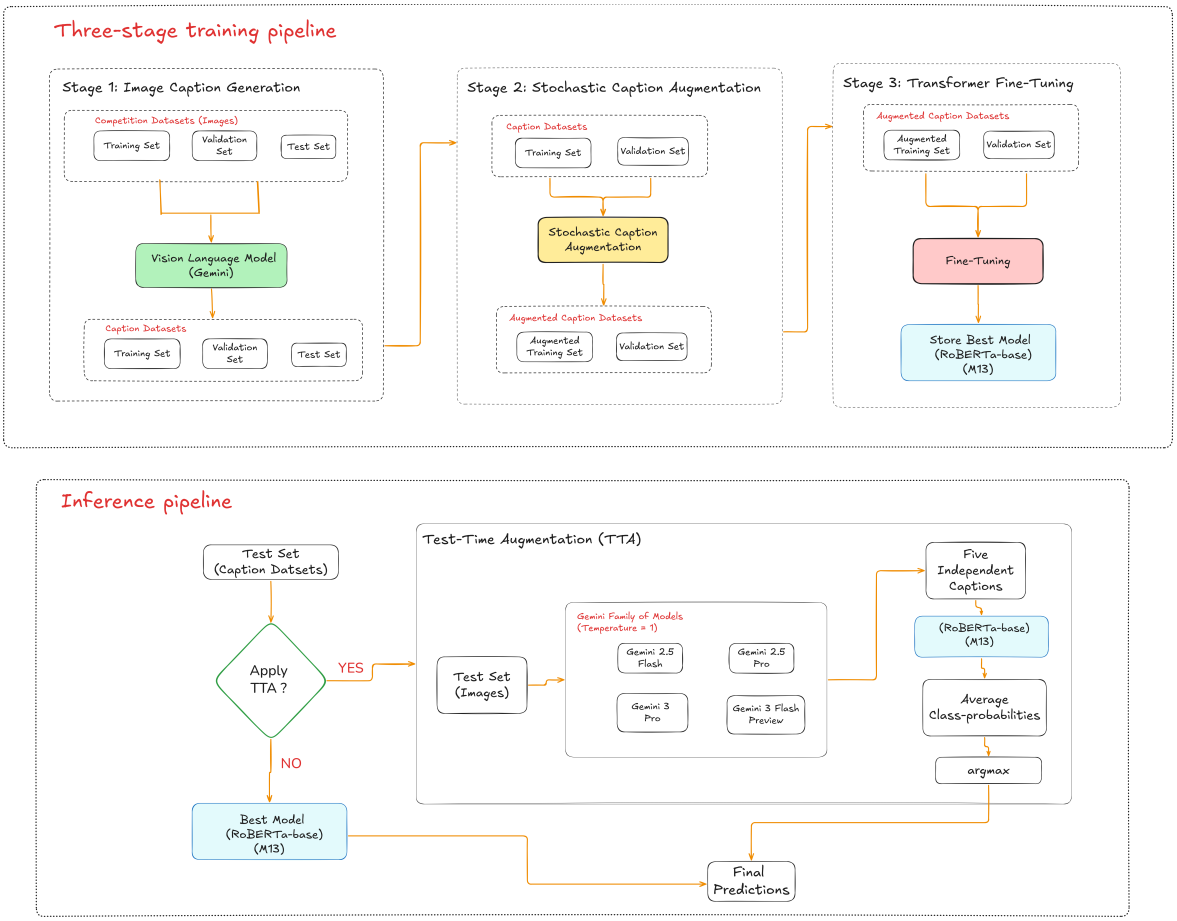

Figure 2: Three-stage training (top) and inference (bottom) pipelines. Stage 1 converts meme images into natural language captions using Gemini 3 Flash (Google DeepMind, 2025a). Stage 2 applies stochastic caption augmentation with a temperature hyperparameter set to 1.0 to address class imbalance. Stage 3 fine-tunes a pre-trained transformer encoder on the augmented caption dataset and stores the checkpoint with the highest validation macro F1. During inference, a total of five independent captions per test image are generated using a family of Gemini model variants (Comanici et al., 2025; Google DeepMind, 2025b) at temperature 1. The resulting class probability distributions are averaged, and the final prediction is obtained via the argmax operation.

The following encoder-only transformer architectures were evaluated, spanning general-purpose, domain-specific, multilingual, and modernized variants.

3.3.1. General-purpose encoder

The study examined multiple general-purpose encoder-only models for text representation. Specifically, BERT, RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2021) were employed. In each case, the standard base configuration was utilized to process caption data.

3.3.2. Domain-specific encoder

HateBERT: HateBERT (Caselli et al., 2021), a model pre-trained on English abusive-language corpora obtained from banned Reddit communities, was included to assess the effectiveness of a domain-specific encoder.

3.3.3. Multilingual encoder

MuRIL: MuRIL (Khanuja et al., 2021), pre-trained on 16 Indian languages and English, was included because of its outstanding results for hate speech detection in Devanagari script languages (Ale et al., 2025).

3.3.4. Modernized encoder

ModernBERT: ModernBERT (Warner et al., 2025), trained on 2 trillion tokens, optimizing the original BERT with modern architectural enhancements was incorporated to assess the impact of large-scale pre-training on caption-based text classification.

3.4. Comparison Baselines

To isolate the contribution of context-aware image captions, the following comparison baselines were established:

3.4.1. Prompting-based approach

Zero-shot prompting was used to establish an initial baseline using the Gemini 3 Pro model (Google DeepMind, 2025b), which demonstrated the strong image understanding capabilities among other Gemini variants (Comanici et al., 2025). The full structured prompt is provided in Appendix B.

3.4.2. Text-only modality

A text-only modality where two multilingual encoder-only transformer models, MuRIL and XLM-RoBERTa-base (Conneau et al., 2019), were evaluated using the OCR-extracted Devanagari meme text.

3.4.3. Vision-text modality

To incorporate visual information, an early fusion technique was implemented in which input images were encoded using Vision Transformer (ViT) (Dosovitskiy et al., 2021), while multilingual encoder-only transformers (Khanuja et al., 2021; Pires et al., 2019; Pudasaini et al., 2023) were used to encode the OCR-extracted text. The obtained image and text representations were concatenated and passed to a linear classification head.

An image-caption modality was also explored to assess whether the encoded image representation provides additional discriminative information beyond the Gemini caption.

3.5. Hyperparameters and Training Configuration

- **Hardware:** NVIDIA GeForce RTX 3090 GPU (24GB VRAM)
- **Optimizer:** AdamW (lr=2e-5, weight decay=0.01, β_1=0.9, β_2=0.999, ε=1e-8)
- **Max sequence length:** 512
- **Batch Size:** 16
- **Epoch:** 8 (early stopping)
- **Random seed:** 0

4. Results and Discussion

The Shared Task was hosted on the **Codabench** [2] platform, where predictions were submitted as binary output files (0 or 1). The submissions were evaluated using macro-averaged precision, recall, F1-score, and accuracy, with leaderboard rankings based on macro F1-score. Table 3 reports the computed metrics using the test split across all modalities, including their respective IDs.

4.1. Zero-shot prompting

As a baseline (see Table 3), the Gemini 3 Pro, chosen for its competitive image understanding capabilities (Comanici et al., 2025; Google DeepMind, 2025b), was used under a zero-shot prompting paradigm (M1), providing no training examples and relying entirely on the model's pre-trained knowledge. This approach yielded the weakest performance across all configurations, with a macro F1-score of 0.3728, precision of 0.3187, recall of 0.4485, and accuracy of only 0.3060. These results highlight the limitations of relying on a large general-purpose VLM without domain-specific fine-tuning and confirm that task-specific training is essential for hate speech detection in Nepali memes.

4.2. Text-only models

Two multilingual text encoders, namely MuRIL (M2) and XLM-RoBERTa-base (M3), were fine-tuned on OCR-extracted meme text. Both models achieved identical test-set performance, with macro F1-score, precision, recall, and accuracy of 0.4018, 0.3358, 0.5000, and 0.6716, respectively. The low macro F1-score indicates that the text-only modality is insufficient, motivating the exploration of multimodal strategies such as text–image fusion.

4.3. Vision-text models

A multimodal strategy combining text and image representations using an early fusion technique was investigated to combine both textual and visual information. All evaluated combinations, including MuRIL + ViT (M4), NepaliBERT + ViT (M5), and mBERT + ViT (M6), significantly outperformed the text-only baseline, achieving F1-scores ranging from 0.6199 to 0.6266, with M4 yielding the best performance.

4.4. Caption-based models

The competitive performance of the vision–text modality motivated the exploration of a caption-based experiment. The experiments were con-

[2] https://www.codabench.org/competitions/12090/

Category	ID	Text Model	Vision Model	Augmentation Type	F1 Score	Precision	Recall	Accuracy
Zero-Shot prompting	M1	Gemini 3 Pro		None	0.3728	0.3187	0.4485	0.3060
Text Only	M2	MuRIL	–	None	0.4018	0.3358	0.5000	0.6716
	M3	XLM-RoBERTa-base	–	None	0.4018	0.3358	0.5000	0.6716
Text + Image	M4	MuRIL	ViT	None	0.6266	0.6535	0.6210	0.7015
	M5	NepaliBERT	ViT	None	0.6194	0.6670	0.6149	0.7090
	M6	mBERT	ViT	None	0.6119	0.6692	0.6091	0.7090
Caption Only	M7	MuRIL	–	None	0.4242	0.3684	0.5000	0.7368
	M8	HateBERT	–	None	0.5991	0.6094	0.6240	0.6119
	M9	BERT-base-uncased	–	None	0.6511	0.6521	0.6503	0.6940
	M10	RoBERTa-base	–	None	0.6694	0.6949	0.6606	0.7313
	M11	DeBERTa-base	–	None	0.4453	0.6730	0.5172	0.6791
	M12	ModernBERT-base	–	None	0.5213	0.5405	0.5520	0.5563
	M13	RoBERTa-base	–	Stochastic	0.7288	0.7241	**0.7472**	0.7463
	M14	**RoBERTa-base**	–	**TTA**	**0.7397**	**0.7612**	0.7285	**0.7836**
Caption + Image	M15	RoBERTa-base	ViT	–	0.6059	0.6553	0.6035	0.7015

Table 3: Performance of all experimental configurations on the CHiPSAL 2026 Subtask A test set. All metrics are macro-averaged. **Bold** = best value per column. TTA = Test-Time Augmentation.

ducted under two settings: with and without stochastic caption augmentation.

4.4.1. Without Stochastic Caption Augmentation

Six models were evaluated: MuRIL (M7), HateBERT (M8), BERT-base-uncased (M9), RoBERTa-base (M10), DeBERTa-base (M11), and ModernBERT-base (M12). Among these, RoBERTa-base performed the best, achieving an F1-score of 0.6694 without augmentation. MuRIL performed the worst with an F1-score of 0.4242, suggesting that multilingual models are less effective for caption-based classification. Despite being trained on domain-specific data or using modernized architectures, HateBERT and ModernBERT-base underperformed, with F1-scores of 0.5991 and 0.5213, respectively, compared to the vision–text modality. This is likely due to domain discrepancies and the limited amount of fine-tuning data available for this low-resource Nepali task.

Additionally, the Caption + Image modality was evaluated using RoBERTa-base + ViT (M15), achieving an F1-score of 0.6059, which is lower than the best-performing RoBERTa-base model under the caption-only category.

4.4.2. With Stochastic Caption Augmentation

Caption-level data augmentation strategies leveraging the stochastic nature of Gemini variants (Comanici et al., 2025; Google DeepMind, 2025a,b) were studied. In this setting, multiple captions were generated per image by sampling at **temperature 1.0**, effectively doubling the minority non-hate class training samples. The RoBERTa-base model trained with this augmentation outperformed the model without augmentation, achieving an F1-score of 0.7288. Applying TTA further improved the F1-score to 0.7397 and secured third place on the CHiPSAL 2026 final leaderboard.

5. Conclusion and Future Work

This study highlights the potential of contextually rich captions in multimodal hate speech detection in the Devanagari script. The systematic evaluation of unimodal (text-only), multimodal (image-text), and caption-based modeling with stochastic caption augmentation and TTA demonstrated the effectiveness of caption-based modeling, achieving **third** place on the CHiPSAL 2026 shared task leaderboard.

Future work may explore contrastive pre-training on large-scale Devanagari meme datasets to align images with context-aware captions using a contrastive learning framework (Chen et al., 2020).

6. Limitations

Despite strong image understanding capabilities, Gemini and its variants (Comanici et al., 2025; Google DeepMind, 2025a,b) lack sufficient knowledge of the Nepali cultural context, idiomatic expressions, and nuanced humor. Over 75% of the generated captions begin with phrases such as "The image is...", "This meme features...", or "The image displays". Such redundant lexical patterns across the dataset may have impacted the learned latent representations. The example memes and generated captions are provided in Appendix C.

7. Acknowledgements

The author thanks the organizers of the CHiPSAL 2026 shared task for providing the annotated datasets and organizing the competition.

8. Bibliographical References

Prabhat Ale, Anish Thapaliya, and Suman Paudel. 2025. MDSBots@NLU of Devanagari script languages 2025: Detection of language, hate speech, and targets using MURTweet. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 308–313, Abu Dhabi, UAE. International Committee on Computational Linguistics.

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.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (ALW), co-located with EMNLP 2021*, pages 17–25. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Google DeepMind. 2025a. Gemini 3 flash: frontier intelligence built for speed. https://blog.google/products-and-platforms/products/gemini/gemini-3-flash/. Official Google AI blog post. Accessed: 2026-03-01.

Google DeepMind. 2025b. Gemini 3 pro model card. Technical report, Google DeepMind. Model card for Gemini 3 Pro. Accessed: 2026-03-01.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. MuRIL: Multilingual representations for Indian languages. *arXiv preprint arXiv:2103.10730*.

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. In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624. Curran Associates, Inc.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Nikita Moshkov, Botond Mathe, Attila Kertesz-Farkas, Reka Hollandi, and Peter Horvath. 2020. Test-time augmentation for deep learning-based cell segmentation on microscopy images. *Scientific reports*, 10(1):5068.

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.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Shushanta Pudasaini, Subarna Shakya, Aakash Tamang, Sajjan Adhikari, Sunil Thapa, and Sagar Lamichhane. 2023. Nepalibert: Pre-training of masked language model in nepali corpus. In *2023 7th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pages 325–330.

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)*.

Surendrabikram Thapa, Siddhant Bikram Shah, Kritesh Rauniyar, Shuvam Shiwakoti, Surabhi Adhikari, Hariram Veeramani, Kristina T. Johnson, Ali Hurriyetoglu, Hristo Tanev, and Usman Naseem. 2025a. Multimodal hate, humor, and stance event detection in marginalized sociopolitical movements. In *Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts*, pages 20–31, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547.

A. Image Caption Generation Prompt

(1) **Visual and textual content** — Describe all visible elements and transcribe any embedded Devanagari or English text verbatim.

(2) **Target and framing** — Identify the subject of the meme and note whether any person or group is portrayed with respect to gender, race, religion, caste, or organizational status.

(3) **Tone and intent** — Describe the emotional register, noting whether the content expresses affection, humor, support, criticism, mockery, insult, or glorification of violence, and whether any humor appears benign or malicious in intent.

(4) **Cultural context** — Identify culturally specific symbols, political allusions, or references relevant to the Nepali context.

Do not classify the meme; only describe what is present.

B. Zero-shot Hate Speech Detection Prompt

You are an expert content moderator for Nepali social media content. Analyze the provided meme and classify it according to the exact annotation guidelines used in the NeMeme dataset.

Binary Classification Task

Label 0 (Non-Hate): The meme does NOT contain hate speech. Label 1 (Hate): The meme contains hate speech.

Definition of Hate Speech (Per NeMeme Annotation Guidelines)

A meme is HATEFUL (1) if it meets ANY of these criteria:

- Vilifies, denigrates, bullies, insults, or mocks a subject - Targets based on protected characteristics: gender, race, religion, caste, or organizational status - Contains hateful symbols - Glorifies violence

A meme is NON-HATE (0) if it contains:

- Affection, gratitude, support, motivation - Humor without malicious intent - Constructive criticisms (even if critical of public figures or policies).

C. Example Memes and Generated Captions

Caption: *This image is a satirical side-by-side comparison of two titan figures in Nepalese politics, centered on a linguistic pun regarding their names and the quest for political power. On the left, former Prime Minister K.P. Sharma Oli is shown with a stern expression and a raised hand as if making a point or refusing a request; the Nepali text above him reads (*Dinna Ba*), which translates to "I won't give, Father." On the right, fellow multi-term former Prime Minister Sher Bahadur Deuba is pictured smiling with his hands in a traditional *Namaste* gesture, positioned under the text (*Deu Ba*). The humor stems from the wordplay on the word "Ba" (a respectful term for "Father" used by supporters for both men) and their surnames. "Deu Ba" translating to a plea of "Give, Father," whereas "Dinna Ba" portrays Oli as the obstinate leader refusing to yield.*

Caption: *This humorous Nepali meme plays on a common childhood myth regarding the dangers of swallowing gum. The top text in Nepali translates to "Swallowed a lot of chewing gum when I was young," referencing the widespread (and false) childhood fear that doing so could be fatal or cause permanent damage. The visual features two men in stylish attire, with the focal point being a man on the right wearing a cowboy hat and sunglasses, gesturing in a cool, nonchalant manner. Centered over the scene is the phrase "THUG LIFE" accompanied by sunglasses emojis and various "SARCASM Nepal" watermarks. The bottom text delivers the punchline: "But well, it seems I haven't died even until today."*

Caption: *This informational graphic is divided into two equal vertical halves, used to clarify Nepali terminology for different forms of rice. The left side features a bold red header containing white Nepali text that translates to "Let's call cooked rice 'Bhaat'.". The right side of the graphic has a solid blue header with white Nepali text that translates to "Call puffed rice 'Bhuja'." Together, the two sides serve as a linguistic guide, advocating for the specific use of the word "Bhaat" for steamed rice and "Bhuja" specifically for puffed rice.*

Caption: *This Nepali-language meme humorously references the devastating 2015 earthquake in relation to a popular television character. At the top, bold text translates to "In such a massive earthquake," setting a serious backdrop for an ironic question. On the left side of the image is a classic black-and-white "thinking face" internet meme character, looking upward inquisitively with its hand on its chin. To the right is a photograph of the character "Padey" from the well-known Nepali sitcom *Bhadragol*, recognizable by his signature grey beanie with pointed ears, thick-rimmed glasses, and white-painted beard. The bottom caption asks, "I wonder if Padey Solti's 'Cottage' survived or not?" This is a joke directed at fans of the show, as Padey frequently and boastfully refers to his very humble, ramshackle hut as a "Cottage," leading the meme creator to wonder if such a flimsy structure could have possibly withstood the massive tremors.*