Bridging Modalities for Hate Speech Detection in Memes

Fadi Hassan[1,*,†], Evgenii Migaev[1,†], Md Saroar Jahan[1,†] and Frank Mtumbuka[1,†]

[1]*Huawei Finland Research Center*

Abstract

This paper presents our system developed by our research team at FiRC-NLP for the HASOC-meme 2025 shared task, which focuses on detecting hate speech and offensive content in multilingual memes. We explore three distinct approaches: prompt-based inference using Gemini, cross-modality encoders combining image and text, and a text-only modality leveraging Optical Character Recognition (OCR), OCR English translation and image descriptions. Our final ensemble system fuses these modalities, achieving macro F_1-scores up to 67.50 on test sets and top-1 ranking in 3 out of 4 tracks, demonstrating the effectiveness of multimodal fusion and robust training strategies.

Keywords

Hate Speech Detection, Multimodal Classification, Memes, Few-Shot Learning

1. Introduction

Memes pose unique challenges for hate speech detection due to their multimodal nature [1, 2]. Visual elements often carry implicit offensive cues that are not captured by text alone [3, 4]. To address this, we designed a system that explores multiple modalities and fusion strategies, aiming to maximize semantic coverage and classification accuracy. The proliferation of memes on social media has established them as a dominant mode of online communication, capable of conveying ideas through the interplay of text and imagery. While memes are often used for entertainment and cultural commentary, they are increasingly exploited to disseminate offensive and hateful content [5, 6]. Detecting such harmful content presents unique challenges compared to traditional text-based hate speech detection. This difficulty arises from the multimodal nature of memes: textual components may appear benign in isolation but gain offensive meaning when combined with visual cues, and conversely, images may implicitly reinforce or alter the semantics of overlaid text. As a result, effective detection requires approaches that can account for the rich cross-modal interactions inherent in memes [7].

Recent advances in multimodal machine learning have motivated research into systems capable of integrating textual and visual signals for classification tasks. Prior work has explored multimodal fusion for hate detection [8], including CLIP-based approaches [9], and vision-language pretraining [10]. However, challenges like linguistic diversity in datasets such as MUTE [11], and language-specific resources such as BanglaAbuseMeme [12] persist. Designing robust systems for hate speech detection in memes remains an open problem due to issues such as noisy or stylized embedded text, linguistic diversity, implicit visual symbolism, and the limited availability of annotated datasets. Related efforts on hate mitigation in Indic languages, such as SafeSpeech [13], further highlight the importance of addressing linguistic and cultural diversity. To address these challenges, the HASOC-meme 2025 shared task provides a benchmark for evaluating multimodal approaches to hate speech and offensive content detection in memes.

In this paper, we present our system developed for the shared task, which combines multiple modalities and fusion strategies to maximize semantic coverage. Specifically, we explore three complementary approaches: (i) prompt-based inference with large multimodal models, (ii) cross-modality encoders

Forum for Information Retrieval Evaluation, December 17-20, 2025, India

*Corresponding author.

†These authors contributed equally.

fadi.hassan@huawei.com (F. Hassan); evgenii.migaev2@huawei.com (E. Migaev); saroar.jahan@huawei.com (M. S. Jahan); frank.mtumbuka@h-partners.com (F. Mtumbuka)

jointly processing images and text, and (iii) text-only pipelines enhanced with optical character recognition (OCR), translation into English, and image description generation. We further investigate ensemble strategies that integrate these modalities, demonstrating that their combination yields significant performance improvements over single-modality baselines.

Our contributions are threefold:

- We design and evaluate distinct modality-specific pipelines for meme classification.
- We propose a fusion strategy that leverages the complementary strengths of prompt-based, cross-modal, and text-enriched approaches.
- We provide an empirical analysis of system performance on the HASOC-meme dataset, achieving top-1 ranking in three out of four tracks of the competition.

These findings underline the importance of multimodal integration for robust hate speech detection in memes and suggest promising directions for future research, such as adaptive fusion strategies and meme-specific captioning models.

2. Task Description

The HASOC shared task series (Hate Speech and Offensive Content Identification) has reached its seventh edition in 2025. The HASOC-meme 2025 [7] task extends previous editions by focusing specifically on memes, which combine textual and visual modalities, thereby presenting unique challenges for automatic classification. Unlike purely textual corpora, memes often encode offensive cues through both explicit and implicit multimodal signals such as sarcastic phrasing, symbolic imagery, or subtle cultural references.

The 2025 edition introduces a five-part classification challenge, requiring participants to analyze memes across multiple dimensions of offensive communication:

Sentiment Detection

- **Positive:** The meme conveys support, humor, or appreciation.
- **Neutral:** The meme is neither overtly positive nor negative.
- **Negative:** The meme expresses hostility, mockery, or criticism.

Sarcasm Detection

- **Sarcastic:** The meme implies the opposite of its literal meaning, often mocking or ridiculing.
- **Non-Sarcastic:** The meme conveys its message directly, without irony.

Vulgarity Detection

- **Vulgar:** The meme contains explicit or offensive words, gestures, or depictions.
- **Not Vulgar:** The meme does not include such content.

Abuse Detection

- **Abusive:** The meme includes harmful, derogatory, or offensive elements targeting individuals or groups.
- **Non-Abusive:** The meme lacks abusive or derogatory content.

Target Community Identification

- **Gender:** Mentions gender identities (male, female, non-binary, transgender).
- **Religion:** References to religious beliefs, practices, or symbols.
- **Individual:** Mentions or depicts a specific person.
- **Political:** Refers to political parties, ideologies, politicians, or policies.
- **National Origin:** Targets groups by country or ethnicity.
- **Social Sub-groups:** Refers to socio-economic, occupational, or cultural groups.
- **Others:** Any community not covered above.
- **None:** No explicit target community.

Through this multi-dimensional design, HASOC-meme 2025 emphasizes the complexity of abusive content detection, encouraging participants to go beyond binary classification and address fine-grained aspects such as sarcasm, sentiment, and vulgarity.

3. Dataset and Preprocessing

We used the official HASOC-meme dataset. To evaluate and compare different strategies, we employed 5-fold cross-validation, partitioning the training data into five subsets and iteratively using four folds for training and one for validation.

Table 1
Dataset distribution across languages.

Language	Training Data	Test Data
Bangla	2693	1821
Bodo	378	254
Gujarati	889	604
Hindi	1141	769
Total	5101	3448

While the organizers provided baseline OCR text, we hypothesized that a more detailed extraction of visual and textual elements would improve performance [14, 15]. We developed an advanced feature extraction pipeline using the cost-effective and fast **Gemini 2.5 Flash-Lite [16]** model, which involved the following steps:

1. **Image Decomposition:** Each meme image was programmatically divided into several distinct sub-images to isolate different visual components.
2. **Detailed Extraction:** For each sub-image, we prompted the model to perform OCR and generate a concise description of the visual content.
3. **Aggregation and Translation:** The extracted texts and descriptions from all sub-images were aggregated. The combined OCR text was also translated into English to create a unified linguistic representation for subsequent retrieval tasks.

This process yielded a rich set of features for each meme: the original image, a comprehensive OCR text (in its native language), an English translation of the OCR, and a structured description of the visual elements.

4. System description

One of the key challenges in this work was building a strong classifier from limited training data in low-resource languages. To address this, we unified all datasets into a single multilingual corpus and explored two complementary strategies: (1) prompt-based inference using large multimodal foundation models, and (2) fine-tuning compact multi- and single-modal encoders. Our system integrates multiple components, detailed in the following sections:

4.1. Prompt-Based Inference

Our core methodology revolved around prompt-based inference using a powerful multimodal foundation model. We explored both zero-shot and few-shot paradigms to leverage the model's capabilities.

4.1.1. Foundation Model Selection

Our initial step was to select the most suitable foundation model. We conducted a preliminary evaluation on a small subset of 20 training samples, testing OpenAI's GPT-4, Anthropic's Claude 3 Opus, and Google's **Gemini 2.5 Flash**. Each model was prompted to classify the meme images across the four target labels. In this comparison, **Gemini 2.5 Flash** demonstrated consistently superior performance, leading us to select it for all subsequent experiments.

4.1.2. Zero-Shot Classification

Our first strategy was a **zero-shot** approach, designed to test the model's intrinsic understanding of the task without any prior examples. The model was provided with the raw meme image and a carefully crafted prompt instructing it to classify the image according to the four labels (sentiment, abuse, vulgar, sarcasm). The prompt 1 also required the model to provide a brief explanation for its decision, which offered valuable qualitative insights into its reasoning process. This approach used no samples from our prepared training set.

4.1.3. Few-Shot In-Context Learning via Retrieval

Our second, more advanced strategy was a **few-shot** approach that provided the model with relevant examples from the training set to guide its prediction. This method, often referred to as Retrieval-Augmented Generation (RAG), relied on our extracted features rather than the raw images. The process was as follows:

1. **Vector Database Creation:** We used a Gemini text embedding model to generate vector embeddings for the **English-translated OCR** and the **image descriptions** of all samples in the training set. These embeddings were stored in a Chroma vector database.
2. **Dynamic Example Retrieval:** For each new sample, we performed two similarity searches against the vector database to retrieve the top 5 training examples with the most similar translated OCR and the top 5 examples with the most similar image description.
3. **Prompt Engineering:** The retrieved 10 examples, along with their ground-truth labels, were concatenated and formatted into the prompt. The model was then tasked with classifying the target sample based on this contextual information. Prompt 2

4.2. Cross-Modality Encoders

We fused the vision encoder from google/siglip-base-patch16-224 (SigLIP) [17] with XLM-RoBERTa-large (XLM-R) [18] in a similar way to Figure 4.3, leveraging the strengths of both modalities: SigLIP provides robust visual representations, while XLM-R excels in handling multilingual text, especially in low-resource languages. This hybrid design allowed us to build a more effective cross-modal encoder capable of capturing nuanced relationships between meme images and their associated text across diverse linguistic contexts.

4.3. Text Modality via OCR and Description

We used the OCR of the images and applied Gemini (using Prompt 3) to generate detailed image descriptions and translate the content into English. The outputs from Gemini and the OCR process were then fed into two language models: XLM-R and l3cube-pune/bengali-bert (l3cube-bangla) [19], for classification, as illustrated in Figure 4.3.

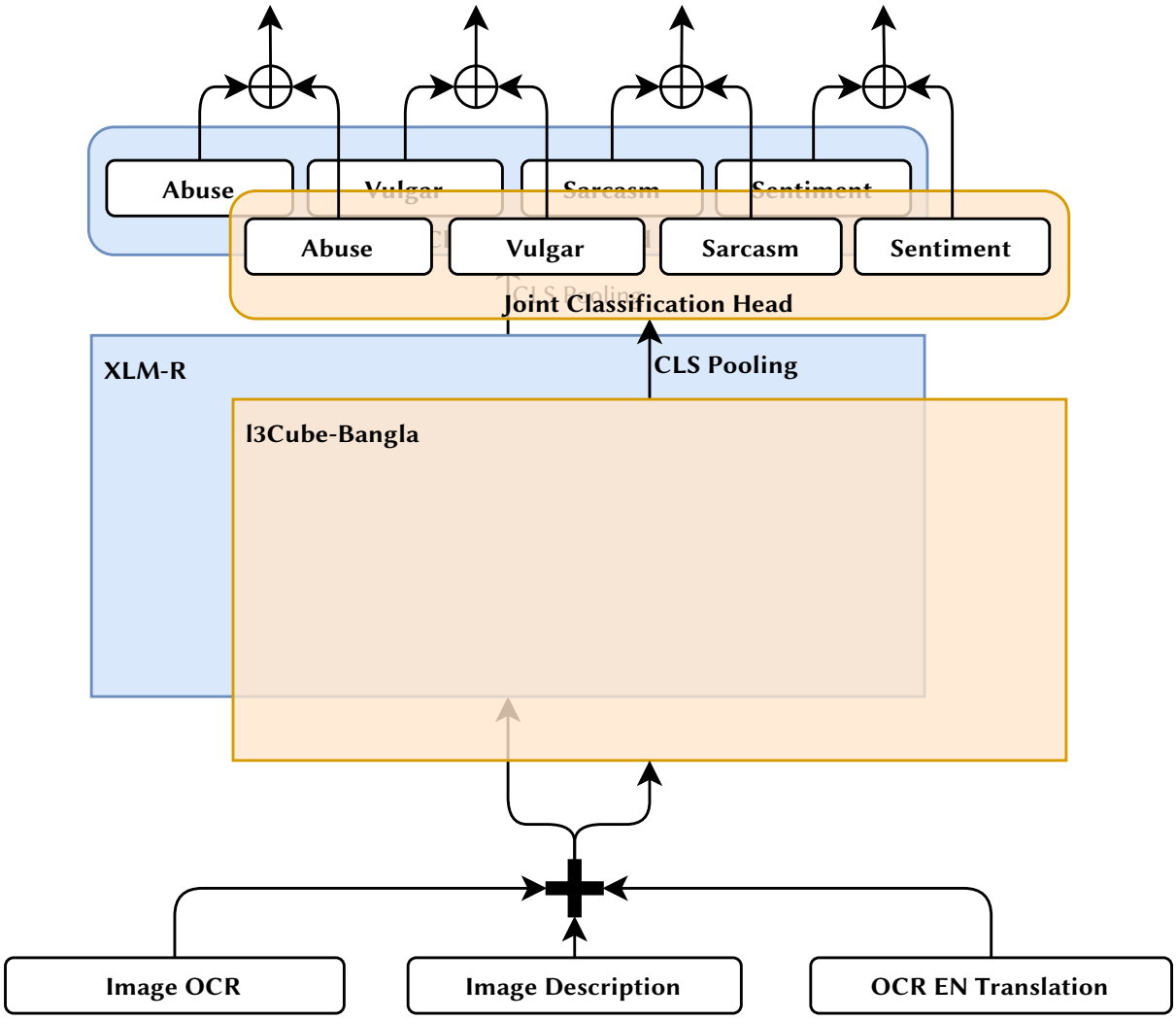

Figure 1: Architecture of text modality encoders used for multilingual meme classification.

4.4. Fusion Strategy for Final Submission

For the final submission, we adopted a **hybrid selection strategy** to maximize the overall macro F_1-score. For each of the 16 tasks (4 languages × 4 labels), we compared the performance of all our methods, including zero-shot LLM, few-shot LLM, and our encoder-based models in the internal validation set.

The approach that yields the highest validation F_1 score for a specific task was selected to generate the final prediction for the test set. For encoder-based models, we further optimized the F_1 score by setting a prediction threshold for each label based on its positive class distribution in the training data. This ensured that the predicted label ratios were aligned with the observed data. This strategic combination allowed us to take advantage of the distinct strengths of each modeling technique.

5. Experimental Setup

For the fine-tuned models, we employed 5-fold cross-validation, where four splits were used for training and one split for validation (Valid) in each fold. At the end of training, predictions from the five folds were combined into an ensemble to score the test set. All models were implemented using the HuggingFace Transformers framework and optimized with early stopping to prevent overfitting. A constant learning rate schedule with warmup was applied to enhance training stability and overall performance.

We trained the models on four out of the five classes (sentiment, sarcasm, vulgarity, and abusive), excluding the target community class since it was not requested in the final submission and, in preliminary experiments, its inclusion degraded performance due to potentially low-quality annotations.

6. Results and Discussion

We evaluate the performance of multiple approaches across Bangla, Bodo, Gujarati, and Hindi, reporting macro F_1-scores on both validation and test sets. Table 2 presents the performance comparison of different approaches.

Language →			**Bangla**		**Bodo**		**Gujarati**		**Hindi**	
Approach ↓	Backbone	Training set	Valid	Test	Valid	Test	Valid	Test	Valid	Test
Prompt-Based (Zero-Shot)	Gemini	-	58.10	59.45	45.90	46.28	58.71	59.48	58.72	59.18
Prompt-Based (Few-Shot)	Gemini	all	52.34	54.04	55.71	50.53	57.02	59.20	56.40	59.62
Cross-Modality Encoder	XLM-R + SigLIP	all	55.95	56.48	48.36	52.81	62.09	63.96	60.00	61.31
Text Modality (OCR only)	XLM-R + l3cube-Bangla	all	58.56	59.48	55.04	55.55	64.27	65.16	**61.56**	63.93
Text Modality (OCR + Image Description + OCR Translation)	XLM-R + l3cube-Bangla	all	**61.02**	59.91	53.34	54.45	**64.33**	65.84	61.18	64.95
		Bodo	-	-	**61.02**	57.39	-	-	-	-
Combined Ensemble	-	all	-	**62.75**	-	**62.22**	-	**67.50**	-	**65.71**

Table 2
Macro F_1-scores per language on validation and test sets for each approach.

6.1. Prompt-Based Methods

Zero-shot prompt-based approaches using Gemini perform surprisingly well for the test set of the Bangla language, outperforming the fine-tuned models. However, when applied to other languages, its performance tends to be more modest. On the other hand, few-shot prompting occasionally outperforms zero-shot prompting but not consistently. This variability makes it challenging to draw a clear conclusion about its overall effectiveness.

6.2. Text-Modality Encoders

Approaches leveraging text-only encoders (XLM-R + l3cube-Bangla) generally outperform prompt-based strategies, particularly for Gujarati and Hindi. However, their performance varies across languages, with weaker results observed for Bodo. The performance on Bodo was notably low when the model was trained jointly with other languages. We hypothesize that this may be due to inconsistencies in the annotation of Bodo data, warranting further investigation of the annotation process. Similar low-resource challenges were also discussed in prior shared task analyses [20]. To address this, we trained a monolingual classifier for Bodo, which resulted in improved performance.

6.3. Cross-Modality Fusion

Models combining multimodal cues (OCR, image descriptions, and OCR translations) consistently outperform text-only methods. This fusion strategy leads to the strongest results, indicating that meme-based hate speech detection benefits substantially from integrating multimodal signals.

6.4. Combined Ensemble

The ensemble approach achieves the best overall macro F_1-score. These results confirm that ensemble-based strategies, which leverage diverse models and representations, provide robust improvements across all languages.

7. Conclusion

Our system demonstrates that combining prompt-based reasoning, cross-modality encoding, and enriched text-only pipelines yields strong performance in meme-based hate speech detection.

Future work will focus on exploring dynamic fusion strategies and fine-tuning captioning models for meme-specific semantics. In addition, we plan to investigate encoder fusion using cross-attention mechanisms, as the current implementation relies solely on late aggregation (where inputs are independently processed through encoders and fused only at the end). Cross attention has the potential to capture richer cross-encoder interactions and improve overall system effectiveness.

8. Acknowledgments

We would like to thank the HASOC team and all participants for their efforts in organizing and contributing to the shared task.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4-based tools in order to: Grammar and spelling check. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The hateful memes challenge: Detecting hate speech in multimodal memes, CoRR abs/2005.04790 (2020). URL: https://arxiv.org/abs/2005.04790. arXiv:2005.04790.

[2] M.-H. Van, X. Wu, Detecting and mitigating hateful content in multimodal memes with vision-language models, 2025. URL: https://arxiv.org/abs/2505.00150. arXiv:2505.00150.

[3] P. Kapil, A. Ekbal, A transformer based multi task learning approach to multimodal hate speech detection, Natural Language Processing Journal 11 (2025) 100133. URL: https://www.sciencedirect.com/science/article/pii/S2949719125000093. doi:https://doi.org/10.1016/j.nlp.2025.100133.

[4] C. Yang, F. Zhu, Y. Liu, J. Han, S. Hu, Uncertainty-aware cross-modal alignment for hate speech detection, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 16973–16983. URL: https://aclanthology.org/2024.lrec-main.1475/.

[5] E. L. T. Tchokote, E. F. Tagne, Effective multimodal hate speech detection on facebook hate memes dataset using incremental pca, smote, and adversarial learning, Machine Learning with Applications 20 (2025) 100647. URL: https://www.sciencedirect.com/science/article/pii/S2666827025000301. doi:https://doi.org/10.1016/j.mlwa.2025.100647.

[6] C. Koutlis, M. Schinas, S. Papadopoulos, Memefier: Dual-stage modality fusion for image meme classification, in: Proceedings of the 2023 ACM International Conference on Multimedia Retrieval, ICMR '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 586–591. URL: https://doi.org/10.1145/3591106.3592254. doi:10.1145/3591106.3592254.

[7] K. Ghosh, M. Das, M. Narzary, S. Saha, S. Barman, A. Mukherjee, S. Modha, D. Ganguly, U. Garain, S. Jaki, T. Mandl, Overview of the hasoc track at fire 2025: Abusive meme identification — shadows behind the laughter, in: K. Ghosh, T. Mandl, S. Pal, S. Majumdar, A. Chakraborty (Eds.), Forum for Information Retrieval Evaluation (Working Notes) (FIRE 2025), December 17–20, Varanasi, India, CEUR-WS.org, 2025.

[8] R. Prabhu, V. Seethalakshmi, A comprehensive framework for multi-modal hate speech detection in social media using deep learning, Scientific Reports 15 (2025) 13020. URL: https://doi.org/10.1038/s41598-025-94069-z. doi:10.1038/s41598-025-94069-z.
[9] G. Arya, M. K. Hasan, A. Bagwari, N. Safie, S. Islam, F. R. A. Ahmed, A. De, M. A. Khan, T. M. Ghazal, Multimodal hate speech detection in memes using contrastive language-image pre-training, IEEE Access 12 (2024) 22359–22375. doi:10.1109/ACCESS.2024.3361322.
[10] Y. Chen, F. Pan, Multimodal detection of hateful memes by applying a vision-language pre-training model, PLOS ONE 17 (2022) e0274300. URL: https://doi.org/10.1371/journal.pone.0274300. doi:10.1371/journal.pone.0274300.
[11] E. Hossain, O. Sharif, M. M. Hoque, MUTE: A multimodal dataset for detecting hateful memes, in: Y. Hanqi, Y. Zonghan, S. Ruder, W. Xiaojun (Eds.), Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop, Association for Computational Linguistics, Online, 2022, pp. 32–39. URL: https://aclanthology.org/2022.aacl-srw.5/. doi:10.18653/v1/2022.aacl-srw.5.
[12] M. Das, A. Mukherjee, Banglaabusememe: A dataset for bengali abusive meme classification, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 15498–15512.
[13] K. Ghosh, N. K. Singh, J. Mahapatra, et al., Safespeech: a three-module pipeline for hate intensity mitigation of social media texts in indic languages, Social Network Analysis and Mining 14 (2024). URL: https://doi.org/10.1007/s13278-024-01393-9. doi:10.1007/s13278-024-01393-9.
[14] Y. Zhou, Z. Chen, H. Yang, Multimodal learning for hateful memes detection, in: 2021 IEEE International Conference on Multimedia Expo Workshops (ICMEW), 2021, pp. 1–6. doi:10.1109/ICMEW53276.2021.9455994.
[15] A. Anaissi, J. Akram, K. Chaturvedi, A. Braytee, Detecting and understanding hateful contents in memes through captioning and visual question-answering, 2025. URL: https://arxiv.org/abs/2504.16723. arXiv:2504.16723.
[16] R. Cao, R. K.-W. Lee, W.-H. Chong, J. Jiang, Prompting for multimodal hateful meme classification, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 321–332. URL: https://aclanthology.org/2022.emnlp-main.22/. doi:10.18653/v1/2022.emnlp-main.22.
[17] X. Zhai, B. Mustafa, A. Kolesnikov, L. Beyer, Sigmoid loss for language image pre-training, 2023. arXiv:2303.15343.
[18] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.
[19] R. Joshi, L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages, arXiv preprint arXiv:2211.11418 (2022).
[20] K. Ghosh, S. Saha, T. Mandl, S. Modha, Findings from shared tasks on hate speech detection: Performance patterns for low-resource languages, Pattern Recognition Letters (2025). URL: https://www.sciencedirect.com/science/article/pii/S0167865525003150. doi:10.1016/j.patrec.2025.09.004.

A. Prompts in YAML Format

Listing 1: Prompt used for the Zero-Shot approach

```
prompt: >
  You are a helpful content moderator. You need to analyze meme images in Bangla, Hindi,
    Gujarati, or Bodo.
```

```
  You should detect:
    1) sentiment (one from 3 values):
      - positive - The meme conveys a supportive, humorous, or appreciative tone.
      - neutral -  The meme is neither overtly positive nor negative in tone.
      - negative - The meme expresses hostility, mockery, or criticism.
    2) sarcasm (True or False):
      - True - The meme presents statements or visuals that imply the opposite of their literal
    meaning, often to mock or ridicule.
      - False - The meme directly conveys its message without sarcasm or irony.
    3) vulgar (True or False):
      - True - The meme contains explicit or offensive words, gestures, or depictions.
      - False - The meme does not include any such content.
    4) abuse (True or False):
      - True - The meme includes offensive, harmful, or derogatory language, imagery, or
    implications targeting an individual or a group.
      - False - The meme does not contain any offensive, harmful, or derogatory content.
    5) target - (one from 8 values):
      - "gender" - Any reference to male, female, non-binary, or transgender identities.
      - "religion" - Mentions or imagery related to any religious belief, deity, or practice.
      - "individual" - Specifically mentions or portrays a particular person.
      - "political" - Targets political ideologies, parties, politicians, or policies.
      - "national" - Targets people based on their country or ethnicity.
      - "social subgroups" - Groups based on socio-economic status, occupation, cultural
    identity, or other affiliations.
      - "other" - Any target that does not fall into the above categories.
      - "non-targeted" - If the meme does not target any specific community, no target label is
    assigned.
    6) description - finally provide a short explanation of why you make these predictions.
```

Listing 2: Prompt used for the Few-Shot approach

```
prompt_few_shots: >
  You are a helpful content moderator. You need to analyze meme images in Bangla, Hindi,
    Gujarati or Bodo.
  I will provide you description of every part of the meme (images and texts)
  You should detect:
    1) sentiment (one from 3 values):
      - positive - The meme conveys a supportive, humorous, or appreciative tone.
      - neutral -  The meme is neither overtly positive nor negative in tone.
      - negative - The meme expresses hostility, mockery, or criticism.
    2) sarcasm (True or False):
      - True - The meme presents statements or visuals that imply the opposite of their literal
    meaning, often to mock or ridicule.
      - False - The meme directly conveys its message without sarcasm or irony.
    3) vulgar (True or False):
      - True - The meme contains explicit or offensive words, gestures, or depictions.
      - False - The meme does not include any such content.
    4) abuse (True or False):
      - True - The meme includes offensive, harmful, or derogatory language, imagery, or
    implications targeting an individual or a group.
      - False - The meme does not contain any offensive, harmful, or derogatory content.

  Use the examples below as a guide:

  {examples}

  Meme description to analyze:
  {meme_description}
```

```
description_template: >
  Image description: {image_description}

  Text description: {text_description}

description_ocr_template: >
  OCR from a meme: {ocr}

example_template: >
  -- {description}

  Ground Truth: {ground_truth}
```

Listing 3: Prompt used for Image OCR + Image Description + OCR Translation extraction

```
ocr_prompt: >
  You are an expert at analyzing and describing internet meme images in Bangla, Hindi, Gujarati,
     or Bodo. Your task is to extract the text and provide a concise description of the image
    content for each distinct part of a meme.
  You need to return a list of descriptions of every part of the meme. For every part (subimage)
    , return the location of the part, image_description and text in the original language it
    exists on this part and translated_text of this text to the English language (if the
    original text is English, then return the same text). Your example output:
    [{"location": "Top", "image_description": None, "text": "work in IT", "translated_text": "
    work in IT"}, {"location": "right bottom image", "image_description": "picture of software
    developer working at night", "text": "I will work only when I want", "translated_text": "I
    will work only when I want"}, {"location": "left bottom image", "image_description": "
    picture of tired and angry software developer", "text": "And also when my boss will ask me",
     "translated_text": "And also when my boss will ask me"}]
```