A Multimodal Prompt-based Framework for Analyzing Code-Mixed and Low-Resource Memes

Surendrabikram Thapa[1], **Hariram Veeramani**[2], **Liang Hu**[3, 4], **Qi Zhang**[3, 4], **Wei Wang**[5], **Usman Naseem**[6]

[1]Virginia Tech, Blacksburg, Virginia, USA
[2]University of California Los Angeles, California, USA
[3]DeepBlue Academy of Sciences, China
[4]Tongji University, Shanghai, China
[5]Shenzhen MSU-BIT University, Shenzhen, China
[6]Macquarie University, Sydney, New South Wales, Australia
usman.naseem@mq.edu.au

Abstract

The emergence of social media has led memes to become a powerful mode of communication, blending text, images, and emojis. However, this surge in meme usage has also seen a rise in offensive material. With manual content moderation proving impractical due to the sheer volume of data, there's a pressing need for automated methods to identify harmful memes. Yet, existing research predominantly targets high-resource languages such as English, neglecting low-resource ones like Nepali. To bridge this gap, we introduce the first Nepali meme dataset annotated for hate speech and sentiment. Our contributions are threefold: (1) We create and release **NeMeme**, a unique dataset featuring Nepali and code-mixed Nepali memes (combining Nepali and English). (2) We evaluate NeMeme using cutting-edge unimodal and multimodal models to establish initial performance benchmarks. (3) We introduce **MemeNePAL**, a novel multimodal framework employing prompt-assisted learning to effectively categorize Nepali memes. MemeNePAL overcomes the shortcomings of prior state-of-the-art (SOTA) techniques, which were designed for high-resource languages and struggle with Nepali's linguistic differences and cultural subtleties. This work not only promotes inclusivity in content moderation research but also aligns with UN Sustainable Development Goals such as promoting well-being, reducing inequalities, and fostering peace. We adhere to FAIR principles by making the dataset publicly available.

Introduction

The rapid evolution of the digital landscape has given rise to a new form of content that is easily accessible and widely shareable – memes. Memes (as shown in Figure 1), as a form of multimodal content, play a pivotal role in shaping online discourse. Understanding the content and sentiment of memes is of paramount importance, particularly in preventing the spread of hate and harmful ideas. On social media platforms such as Facebook, Twitter (now X), Instagram, and Reddit, memes have become an eminent mode of communication and self-expression. They combine text, images, and often emojis to convey complex ideas and emotions. However, the rise of meme usage has led to an increase in offensive and hateful content, challenging content moderation efforts (Cao et al. 2023). Traditional manual moderation is overwhelmed by the volume, necessitating automated solutions to protect moderators and maintain platform integrity.

(a) Nepali Meme (b) Code-mixed Nepali Meme

Figure 1: Example of monolingual Nepali and code-mixed Nepali-English memes

Researchers are now focusing on analyzing memes beyond text, considering their visual and contextual elements to categorize sentiments accurately. Yet, this research predominantly targets high-resource languages, leaving Nepali and other low-resource languages underexplored. Despite the emergence of code-mixed memes (containing multiple languages), research on Nepali memes, whether monolingual or code-mixed, remains scarce, highlighting a gap in the understanding and moderation of meme content in low-resource language contexts.

Nepali language is spoken by more than 20 million people worldwide. Recently, memes have seen a surge in popularity on social media in Nepal. Despite this, a significant challenge in meme analysis in the Nepali language is the development of algorithms capable of recognizing and analyzing meme information in the Nepali language. The language's script (Devanagari), sociopolitical context, and cultural nuances all contribute to the difficulty of accurately understanding memes using models trained on high-resource

languages. Moreover, memes in the Nepali online ecosystem play a central role in political commentary and public discourse, often conveying subtle and implicit messages that require deep cultural understanding. To address this issue, researchers must develop new algorithms for analyzing multimodal data (image and text). However, limited available data to train these algorithms has impeded progress in automating content detection and moderation in the multimodal meme in Nepali. To overcome this bottleneck, we present a new meme dataset annotated for (i) hate speech and (ii) sentiment analysis in the Nepali language as well as code-mixed Nepali language. This dataset is the first of its kind and addresses the unique challenges of meme understanding in Nepali. To provide a solid foundation for this research, we establish baseline results using various SOTA methods. Furthermore, to effectively categorize hate speech and various sentiments, we introduce **MemeNePAL**, a multimodal framework for categorizing **Meme**s in **Ne**pali language using **P**rompting **A**ided **L**earning. This framework leverages the power of prompting to understand memes in Nepali comprehensively. Our main contributions are:

- We curate and release a novel meme dataset, NeMeme for analysis of Nepali and code-mixed memes in the Nepali language. We labeled the dataset for sentiment (negative, neutral, positive) and hate speech (hate, non-hate). We benchmark our dataset with various models.

  Datasets — https://zenodo.org/records/15164380

- We propose MemeNePAL, a novel approach for Nepali meme analysis by leveraging prompting methods.
- We demonstrate that MemeNePAL outperforms SOTA approaches in both unimodal and multimodal models.

Our research advances multimodal meme analysis in low-resource languages like Nepali while also aligning with several UN Sustainable Development Goals (SDGs). By releasing our dataset publicly, we support the SDG principle of "Leave No One Behind" and encourage inclusive research. Analyzing hate speech promotes mental well-being (SDG-3), reduces inequalities (SDG-10), and supports peace and justice in digital spaces (SDG-16).

Related Works

Existing low-resource monolingual datasets

Unimodal Datasets: Within the domain of low-resource languages, research has explored various aspects of textual online content (Table 1). For instance, Akhtar, Ekbal, and Bhattacharyya (2016) introduced a method and dataset for aspect-based sentiment analysis in Hindi. Similarly, Ali et al. (2021) curated a hate speech detection dataset in Urdu for the analysis of hate speech. Additionally, Romim et al. (2021) developed a dataset for hate speech detection in Bengali language. Several works have also addressed sentiment analysis and hate speech detection in Nepali. Sitaula et al. (2021) introduced a dataset for sentiment analysis of COVID-19-related tweets in Nepali. Similarly, Thapa et al. (2023) presented a dataset of over 13,000 tweets annotated for the presence of hate speech.

Multimodal Datasets: Research on multimodal datasets in low-resource languages is relatively scarce compared to unimodal datasets. Bandyopadhyay et al. (2023) proposed a dataset of Hindi memes annotated for three sarcasm categories (i) Not sarcastic (ii) Mildly sarcastic (iii) Highly sarcastic and 13 fine-grained emotion classes. Similarly, Suryawanshi et al. (2020) proposed the TamilMemes dataset in Tamil for identifying trolls in memes. Recently, Das and Mukherjee (2023) introduced a Bangla abusive meme dataset annotated for the presence of abusive components. Despite works in other low-resource languages, there has been no research on meme analysis in monolingual Nepali.

Existing datasets for code-mixed languages

Unimodal Datasets: As one of the pioneering work in code-mixed Hindi-English language, Bohra et al. (2018) proposed a dataset of tweets for the detection of hate speech. Similarly, Chakravarthi et al. (2020a) proposed a dataset for the sentiment analysis of code-mixed Malayalam-English text. Similarly, Maity, Saha, and Bhattacharyya (2022) introduced a dataset of Hindi-English tweets for the detection of cyberbullying. Additionally, Chakravarthi et al. (2020b) developed a dataset for sentiment analysis of comments in Tamil-English language. Despite works in other low-resource code-mixed languages, there is no research in Nepali code-mixed language.

Multimodal Datasets: Since code-switched memes are also widely shared on the internet, there have been some good works in memes analysis and understanding in code-mixed languages. For instance, Hossain, Sharif, and Hoque (2022a) curated a dataset of Bangla code-mixed memes for the analysis of positive, neutral, and negative sentiment. In addition to that, Hossain, Sharif, and Hoque (2022b) also developed a dataset for the analysis of hate speech in Bangla and code-mixed Bangla-English language. Similarly, Maity et al. (2022) presented a multimodal dataset and framework for analysis of cyberbullying in code-mixed Hindi-English memes. They collected the dataset from Twitter and Reddit platforms. In the case of Dravidian low-resource languages, Kannan and Rajalakshmi (2022) proposed a dataset for troll meme classification in code-mixed Tamil. Additionally, Kannan, Ravikiran, and Rajalakshmi (2022) annotated an interesting dataset in code-mixed Tamil memes. The dataset was annotated for the facial emotions of human subjects in the memes. Despite these efforts in other low-resource code-mixed languages, there is a gap in research on Nepali language. To address this gap, we introduce a comprehensive dataset of *monolingual Nepali* and *code-mixed Nepali* memes, annotated for hate speech and sentiment.

Existing methods for sentiment and hate analysis in low-resource languages

Methods for textual data: Several methods have been developed for sentiment analysis and hate speech detection in textual data in low-resource languages. Jafri et al. (2023) used various machine learning and deep learning models to analyze hate speech in the context of the Indian election and showed transformer-based approach excelled in the

(a) Hateful Code-Mixed Meme | (b) Hateful Nepali Meme | (c) Non-hateful code-mixed meme | (d) Non-hateful Nepali Meme

Figure 2: Examples of Hateful memes in our NeMeme dataset

Work	Nepali	Hate	Sentiment	Multimodal	Source
Jafri et al. (2023)	✗	✓	✗	✗	Twitter
Pereira-Kohatsu et al. (2019)	✗	✓	✗	✗	Twitter
Bhandari et al. (2023)	✗	✓	✗	✓	Twitter, Facebook, Reddit
Waseem and Hovy (2016)	✗	✗	✗	✗	Twitter
Hossain, Sharif, and Hoque (2022b)	✗	✓	✗	✓	Facebook, Twitter, Instagram
Sitaula et al. (2021)	✓	✗	✓	✗	Twitter
Kiela et al. (2020)	✗	✓	✗	✓	Self-generated
Thapa et al. (2023)	✓	✓	✗	✗	Twitter
NeMeme (Ours)	✓	✓	✓	✓	Twitter, Instagram, Reddit, Facebook, Threads

Table 1: Summary of related datasets used in hate speech detection and sentiment analysis.

classification of hate speech in the Hindi language. Apart from this, Chakravarthi et al. (2020a) also showed that the BERT model excelled in the identification of positive, negative, neutral, and mixed sentiment in Malayalam-English text. Similarly, Maity, Saha, and Bhattacharyya (2022) proposed a two-channel CNN model called BERT+VecMap-CNN for detecting code-mixed cyberbullying. One input channel uses the BERT language model, and the other uses bilingual VecMap (Artetxe, Labaka, and Agirre 2018) word embeddings. This dual-input architecture combines semantic knowledge (BERT) and cross-lingual embeddings from VecMap to enhance multilingual cyberbullying detection.

Methods for multimodal data: While research on multimodal data is limited, there is a growing interest in developing methods for analyzing sentiment and hate speech in multimodal content. Various early-fusion and late-fusion of different modalities-based methods are mostly famous. Karim et al. (2022) proposed the fusion of XLM-RoBERTa and DenseNet-161 for the analysis of hate speech in Bengali memes. The method outperforms all the unimodal textual and visual baselines and stands out the best with an F1-score of 0.83. Rajput et al. (2022) proposed a multi-channel CNN + LSTM-based framework for identification of hate-inducing memes in code-switched Hindi-English language. These research suggest that understanding both modalities is important in getting a better representation of memes. Recently, prompt-based methods have been widely used in classification of various sentiments and hate speech in high-resource languages. Motivated by such advancements, we propose a prompt-based learning framework for understanding monolingual Nepali memes and code-mixed memes.

Dataset

In the context of this study, memes are defined as images that include textual content embedded within the visual elements. In this section, we provide an overview of our data collection process, elaborate on the annotation guidelines, and present key data statistics.

Dataset Collection and Deduplication

We obtained our dataset from a variety of social media platforms, including Twitter, Reddit, Instagram, Threads, and Facebook. To ensure data quality, we conducted a manual data collection process, diligently avoiding any low-quality entries. The presence of duplicates can introduce errors into our subsequent analyses and compromise the overall quality of our results. Therefore, we implemented a rigorous deduplication process utilizing two separate deduplication tools. Firstly, we employed dupeGuru[1], a dependable software designed for the identification and removal of duplicate files. Subsequently, we utilized the Duplicate Image Finder, a Python package named difPy[2], to further assist in the re-

[1]https://github.com/arsenetar/dupeguru
[2]https://github.com/elisemercury/Duplicate-Image-Finder

moval of duplicate images. This two-step approach effectively enabled us to eliminate all duplicate entries.

With our dataset now devoid of duplicate content, we extracted the textual information contained within the meme images. This was accomplished using the Google optical character recognition (OCR) Vision API[3], which enabled OCR on the images. This process effectively extracted the text within the memes, helping in subsequent data processing and analysis.

Filtering Criteria

To ensure dataset quality, we applied rigorous filtering criteria. Memes were manually reviewed to confirm they were either monolingual Nepali or code-mixed. Most code-mixed memes included Nepali and English, with occasional Hindi words, as Hindi is also widely used in Nepal. Notably, both Nepali and Hindi use the Devanagari script. For inclusion in our dataset, Nepali memes exclusively contained Nepali language content. Additionally, we excluded memes with poor image quality that hindered accurate OCR text extraction.

Annotation Scheme

The quality of the annotations directly influences the subsequent analyses and model development. Our annotation process was carried out by a team of three annotators, all proficient in English, Hindi, and Nepali languages, and well-versed in the dynamics of memes in social media. These annotators brought diverse educational backgrounds, political perspectives, and life experiences to the task, ensuring a broad spectrum of viewpoints. Recognizing the potential for inaccuracies or inconsistencies in data labeling to distort subsequent analysis and model development, we implemented a three-phase annotation scheme (Li, Yuan, and Li 2023; Bhandari et al. 2023) to familiarize annotators with the annotation process and enhance the quality of their work.

Three-phase Annotation Our three-phase annotation approach is described below:

Pilot Run: In this phase, annotators worked with a set of 50 memes each for both sentiment analysis and hate speech detection tasks. This pilot run was a step to ensure that the annotation instructions were understood by all annotators.

Revision of Instructions: The second phase involved annotating 100 memes using the refined guidelines established after the pilot run. Annotators were provided with revised instructions and tasked with labeling the memes. This phase aimed to make the annotations more explicit and precise.

Consolidation Phase: In the final stage of annotation, each annotator annotated 50 memes for both sentiment analysis and hate speech detection tasks. The discrepancies that emerged during the second phase were addressed through group discussions. This phase ensured that all annotators had a good understanding of the annotation instructions, leading to clearer and less ambiguous guidelines.

Annotation Guidelines

Presence of Hate Speech: Annotators were instructed to identify the presence of hate speech within memes.

[3]https://cloud.google.com/vision/docs/ocr

- **Hateful Memes**: Such memes contained hate speech and were intended to vilify, denigrate, bully, insult, or mock a subject based on characteristics such as gender, race, religion, caste, or organizational status. The memes with hateful symbols and glorified violence were also labeled as hateful memes.
- **Non-hateful Memes**: Such memes did not exhibit hate and conveyed emotions like affection, gratitude, support, motivation, or humor without malicious intent. Constructive criticisms were also labeled as non-hateful.

Figure 2 shows some of the annotated memes for the presence of hate speech in our dataset.

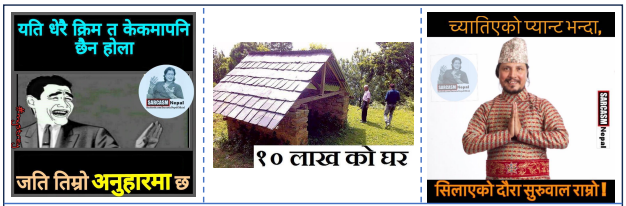
Sentiment Analysis: Sentiment analysis involved categorizing memes into specific sentiment classes.

- **Negative**: Memes with negative sentiment are those that aim to denigrate, insult, or belittle a subject based on their social, personal, or organizational status.
- **Neutral**: Memes with neutral sentiment are those that do not exhibit a clear positive or negative sentiment. They generally offer a more objective perspective without conveying strong emotional tones.
- **Positive**: Memes falling into the positive category include those that express affection, support, gratitude, praise, or motivation.

Figure 3 shows examples of Nepali-English code-mixed memes annotated for various aspects of sentiments. Similarly, Figure 4 shows examples of monolingual Nepali memes annotated for different sentiments.

(a) Negative (b) Neutral (c) Positive

Figure 3: Examples of memes exhibiting different sentiments in Nepali-English code-mixed memes



(a) Negative (b) Neutral (c) Positive

Figure 4: Examples of memes exhibiting different sentiments in monolingual Nepali memes

Furthermore, throughout the annotation process, if annotators encountered any challenges or uncertainties, they conducted group meetings for in-depth discussions. Addition-

ally, they sought guidance and consultation from expert annotators, including postdoctoral researchers and professors, who possessed a profound understanding of the subject matter and annotation guidelines.

Inter-annotator Agreement The inter-annotator agreement scores, Fleiss' Kappa (κ), for code-mixed memes were 0.71 for the presence of hate speech and 0.68 for annotating the aspects of sentiment. Similarly, for Nepali-only memes, the inter-annotator agreement score (κ) was 0.74 for identifying hate speech and 0.69 for sentiment aspects. These scores indicate a high level of consistency.

Dataset Statistics

The data statistics presented in the Table 2 outline the composition of the NeMeme Dataset. Altogether, we have 3,180 code-mixed memes and 3,054 monolingual Nepali memes.

	Problem	Labels	#memes	Avg. Words	Avg. Char.
Code-mixed	Hate Speech	Hate	640	17.43	79.46
		Non Hate	2,540	17.63	76.74
	Sentiment	Neutral	1,754	16.79	76.44
		Positive	817	17.68	78.45
		Negative	609	17.20	77.85
Nepali Only	Hate Speech	Hate	972	16.20	77.85
		Non Hate	2,082	16.11	74.60
	Sentiment	Neutral	1,429	16.45	75.47
		Positive	682	45.46	72.86
		Negative	943	16.29	75.34

Table 2: Statistics of the NeMeme Dataset. The average words and characters are the statistics of text extracted using OCR extraction.

Methodology

Image Captioning Module

In our implementation inspired by Cao et al. (2022), we first utilize the Clipcap model (Mokady, Hertz, and Bermano 2021) to generate captions from our memes. These intermediate captions are processed through the CLIP (Radford et al. 2021) model, enabling us to obtain the intermediate caption embedding from CLIP. Simultaneously, we also extract the CLIP image-text embedding from our meme. Additionally, to incorporate the relevance score between the image and text, calculated by CLIP, as a feature for our feature extraction module (as illustrated in Figure 5), we leverage CLIP-S (Hessel et al. 2021) as a metric. The metric is given by equation 1 where I, c are image and caption, f^I, f^T are CLIP's image and text encoders, and ω is set to 2.5.

$$\text{CLIP-S}(I, c) = \omega \times \max\left(\frac{f^I(I)^\top f^T(c)}{|f^I(I)| \cdot |f^T(c)|}, 0\right) \quad (1)$$

Prompting-based approach

In order to better understand the context of the meme, as shown in Figure 6, we leverage both textual information and image information. We also leverage prompting

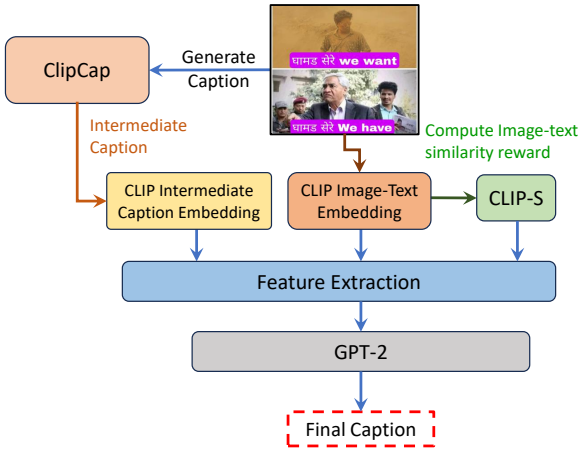


Figure 5: Caption generation model

to provide useful context, expert guidance, and direction to focus the model's multimodal understanding of ambiguous meme content. Research shows that carefully crafted prompts also mitigate data biases, reduce labeling requirements, and enable rapid iteration for improved classification of hate speech in memes (Cao et al. 2022). To operationalize this, we use GPT-2 as our underlying language model due to its open accessibility, efficient performance, and compatibility with prompt-based learning in low-resource settings.

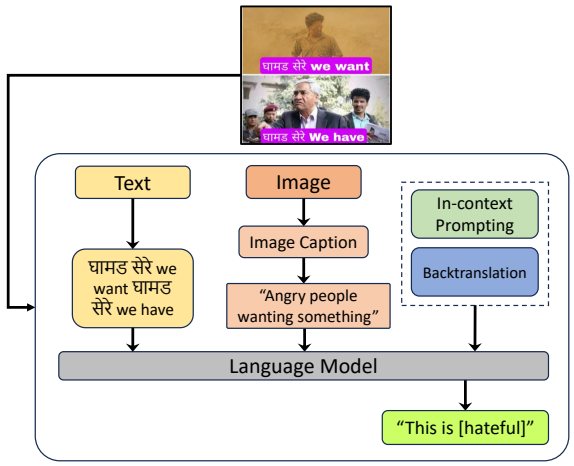


Figure 6: An overview of the proposed framework.

We apply a multi-step approach to enhance model's understanding of memes. We use the text of the memes and randomly mask the words of the text. We do it so for 50% of the cases. For the remaining half, we utilize backtranslation methods on meme text. This allows model to understand the nuances of language more effectively. These three components—unaltered meme text, in-context prompted or backtranslated content, and image captions—are then fed into a language model (LM). We further prompt the LM with

	Models	Sentiment Analysis				Hate Detection			
		Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
Text Only	mBERT-BERT-uncased	0.3819	0.3967	0.3733	0.4064	0.4912	0.4944	0.4759	0.6011
	DeBERTa	0.3863	0.3811	0.3707	0.385	0.5199	0.5038	0.4365	0.6517
	ALBERT-base-v2	0.3822	0.3865	0.3705	0.391	0.5142	0.5037	0.4458	0.6461
	Nepali-BERT	0.4402	0.4234	0.4378	0.4451	0.6185	0.5226	0.48	0.6372
Visual	VGG-16	0.3705	0.3634	0.3325	0.3476	0.4586	0.4823	0.4387	0.6124
	ResNet-50	0.3725	0.3402	0.3308	0.3412	0.4591	0.4733	0.4212	0.6075
	ResNet-101	0.4003	0.4056	0.3758	0.3904	0.5175	0.5075	0.4713	0.6348
Multimodal	VGG-16 + DeBERTa	0.3462	0.3503	0.3459	0.4011	0.3315	0.5	0.3986	0.6629
	VGG-16 + mBERT	0.3072	0.3297	0.3138	0.3797	0.4981	0.4997	0.4234	0.6517
	VGG-16 + NepaliBERT	0.421	0.4053	0.4179	0.4398	0.512	0.5024	0.4509	0.6432
	ResNet-101 + DeBERTa	0.4473	0.4555	0.3927	0.4225	0.6705	0.5249	0.4606	0.6742
	ResNet-101 + mBERT	0.3867	0.4026	0.3667	0.4225	0.622	0.5206	0.4576	0.6685
	ResNet-101 + NepaliBERT	0.4781	0.4711	0.4642	0.4986	0.6249	0.5316	0.4806	0.6348
	Pro-Cap	0.4931	0.4894	0.4912	0.5401	0.6411	0.5602	0.5979	0.6515
	ProHarMeme	0.4811	0.4778	0.4794	0.5393	0.6201	0.5413	0.5780	0.6419
	MultimodalGPT	0.4612	0.4973	0.4789	0.5211	0.5923	0.5096	0.5467	0.6124
	CogVLM	0.4725	0.4871	0.4797	0.5314	0.6052	0.5276	0.5638	0.6287
	LLaVA	0.4543	0.5026	0.4765	0.5158	0.5813	0.5178	0.5478	0.6043
	Proposed Model (MemeNePAL)	**0.5324**	**0.5278**	**0.5301**	**0.5875**	**0.6875**	**0.5983**	**0.6331**	**0.6994**

Table 3: Performance of various models for sentiment and hate analysis in monolingual Nepali dataset

task-specific cues. For instance, in hate speech detection, we prompt the model by presenting a phrase such as 'This is [MASK],' where the model is expected to fill the mask with hate, particularly within hateful memes. The language model is pretrained with masked language model (MLM) objective. It is important to note that our prompting technique uses language model which gives the highest performance with textual data. We call our approach MemeNePAL, a multimodal **meme** analysis framework for **Ne**pali memes with **P**rompt **A**ssisted **L**earning.

Experiments

Experimental Settings

We used uniform experimental settings across all experiments. Consistent with past studies (Ji, Ren, and Naseem 2023), we utilized a 10-fold cross-validation approach to evaluate the performance of baselines and our frameworks.

Baselines

Unimodal Textual Models: We use various text-only models. For our experimentation, we used mBERT-base-uncased (Devlin et al. 2019), DeBERTa-V3 (He et al. 2020), ALBERT-base-V2 (Lan et al. 2019), and Nepali-BERT[4].

Unimodal Visual Models: We use various state-of-the-art visual models as our baselines. We use VGG-16 (Simonyan and Zisserman 2015), ResNet50 (He et al. 2016), and ResNet101 (He et al. 2016).

Mutlimodal Models: We use various combinations of unimodal textual and visual models. The early fusion technique is used to combine visual and textual information. Our model combinations include VGG16+DeBERTa, VGG16+mBERT, VGG16+NepaliBERT, ResNet101+DeBERTa, ResNet101+mBERT, and ResNet101+Nepali-BERT. Apart from these, we use prompt-based methods which include Pro-Cap (Cao et al. 2023) and prompt-based method by ProHarMeme (Ji, Ren, and Naseem 2023). In addition to these, we also evaluate various large vision-language models (LVLMs) like multimodal GPT (Gong et al. 2023), LLaVA-7B (Liu et al. 2023, 2024), and CogVLM-17B (Wang et al. 2023).

[4]https://huggingface.co/Rajan/NepaliBERT

Results

We present the result with the monolingual Nepali dataset, code-mixed dataset, and a combined dataset in Table 3, Table 4, and Table 5 respectively.

Sentiment Analysis: For sentiment analysis in monolingual Nepali text, the Nepali-BERT model exhibited the best performance among text-based models, reflecting its strong grasp on local language intricacies. In contrast, the visual-based ResNet-101 showcased superior performance among image-based models, capturing nuanced visual cues to understand sentiment in the memes. However, the multimodal approach, specifically ResNet-101 combined with NepaliBERT, outperformed all others baselines, reflecting the benefit of leveraging both textual and visual elements in this task. In the monolingual Nepali dataset, the proposed model achieved an F1-score of 0.5301, significantly surpassing the best performing prompt-based baseline (Pro-Cap) at 0.4912.

In the code-mixed dataset, Nepali-BERT continued its dominance in the text-based category, echoing its efficacy in understanding language variations. However, visual models, especially ResNet-101, displayed improved performance compared to the monolingual dataset, possibly due to the visual elements compensating for the linguistic complexities in code-mixed data. In multimodal analyses, the ResNet-

	Models	Sentiment Analysis				Hate Detection			
		Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
Text Only	mBERT-BERT-uncased	0.3894	0.3813	0.3756	0.5923	0.4934	0.4954	0.4891	0.6994
	DeBERTa	0.3519	0.2812	0.2801	0.3846	0.4134	0.4576	0.4295	0.6994
	ALBERT-base-v2	0.3672	0.365	0.3524	0.5772	0.4865	0.4623	0.4652	0.7102
	Nepali-BERT	0.417	0.413	0.4195	0.6541	0.5142	0.509	0.51	0.7235
Visual	VGG-16	0.3865	0.3321	0.3331	0.4692	0.5426	0.5135	0.492	0.7572
	ResNet-50	0.3821	0.3307	0.3328	0.4585	0.5074	0.5128	0.485	0.7297
	ResNet-101	0.3722	0.3433	0.3567	0.5923	0.5172	0.5102	0.5029	0.7225
Multimodal	VGG-16 + DeBERTa	0.3203	0.2735	0.295	0.5077	0.3902	0.5	0.4383	0.7803
	VGG-16 + mBERT	0.3679	0.3487	0.3564	0.6462	0.8924	0.5132	0.4654	0.7861
	VGG-16 + NepaliBERT	0.402	0.4085	0.4138	0.6497	0.5137	0.515	0.5104	0.7764
	ResNet-101 + DeBERTa	0.4012	0.405	0.3959	0.6308	0.4527	0.4782	0.4541	0.7168
	ResNet-101 + mBERT	0.3739	0.3699	0.3696	0.6231	0.4971	0.4987	0.4795	0.7341
	ResNet-101 + NepaliBERT	0.429	0.42	0.4276	0.6922	0.5238	0.516	0.5124	0.7168
	Pro-Cap	0.4211	0.4170	0.4190	0.6816	0.5393	0.5221	0.5306	0.7341
	ProHarMeme	0.4082	0.4021	0.4051	0.6773	0.5316	0.5194	0.5254	0.7317
	MultimodalGPT	0.4012	0.3957	0.3984	0.6638	0.5182	0.5068	0.5124	0.7211
	CogVLM	0.4125	0.3989	0.4056	0.6792	0.5248	0.5112	0.5179	0.7283
	LLaVA	0.4056	0.3976	0.4015	0.6704	0.5205	0.5093	0.5149	0.7248
	Proposed Model (MemeNePAL)	**0.437**	**0.4733**	**0.4651**	**0.7134**	**0.569**	**0.548**	**0.5497**	**0.8015**

Table 4: Performance of various models for sentiment and hate analysis in code-mixed dataset

101 combined with NepaliBERT once again exhibited superior performance, confirming the advantages of multimodal fusion in complex language contexts. In the code-mixed dataset, the F1-score of the proposed **MemeNePAL** model reached 0.4651, exceeding the highest baseline at 0.4276. Across all datasets, the proposed model consistently emerged as the best performer showcasing an F1-score of 0.5489 against the best baseline at 0.5132 (Pro-Cap). This consistent performance superiority across all datasets indicates the robustness and effectiveness of the proposed model in sentiment analysis for diverse language contexts. Our proposed method also surpassed all LVLMs in all the datasets with a good margin.

Hate Speech Detection: In the monolingual Nepali dataset, Nepali-BERT, being a text-based model, exhibited a solid performance, understanding the local language nuances. Contrarily, ResNet-101 among visual models showed good performance, implying the importance of visual cues in detecting hate speech. However, the combined ResNet-101 and NepaliBERT in multimodal models outperformed other models among baselines, emphasizing the benefits of incorporating both visual and textual contexts in hate speech detection. In the monolingual Nepali dataset, the proposed model exhibited an F1-score of 0.6331, markedly outperforming the leading baseline (Pro-Cap) at 0.5979. This shows the effectiveness of our approach.

In the code-mixed dataset, Nepali-BERT maintained its robust performance in text-based models, reiterating its adaptability to language variations. Visual models such as ResNet-101 showed enhanced performance in code-mixed datasets, implying that visual cues compensated for linguistic complexities. Similar to the sentiment analysis task, ResNet-101 paired with NepaliBERT outshone other baseline multimodal models, proving the significance of combining visual and textual elements in hate speech detection. A similar trend is observable in the code-mixed dataset for our proposed model, where it achieved an F1-score of 0.5497 against the best baseline (Pro-Cap) at 0.5306.

However, as compared to the code-mixed language, in the combined dataset, the performance of ResNet-101 diminished slightly, indicating potential limitations when handling multiple languages together. Yet, across combined datasets, the ResNet-101 combined with NepaliBERT consistently offered impressive results among baseline multimodal models. The combined dataset also saw the proposed model displaying a superior F1-score of 0.6785, outstripping the best baseline (Ji, Ren, and Naseem (2023)) at 0.6170. These results demonstrate the proficiency of the proposed model in detecting hate speech across various linguistic and contextual complexities. The proposed method significantly outperforms the recent LVLMs with 3 to 15 % increment in F1-score across different datasets used in our study. This shows that while LVLMs are appropriate in understanding the memes and its sentiment, its ability is mostly limited to high resource languages.

Ablation Analysis

In our first ablation setting, we remove MLM from our model shown in Figure 6. Second, we remove CLIP Intermediate caption embedding component (represented as MemeNePAL - ICE) and CLIP Image-text embedding component (MemeNePAL - CITE) one by one from our caption generation model as shown in Figure 5. Our analysis in Table 6 show that for all datasets, MLM is the most important component in the model, followed by CLIP Intermediate caption embedding and CLIP Image-text embedding in caption generation. This shows that the strength of our model lies in the overall integration of all components.

	Models	Sentiment Analysis				Hate Detection			
		Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
Text Only	mBERT-BERT-uncased	0.4167	0.427	0.4172	0.4949	0.5734	0.5912	0.5714	0.6322
	DeBERTa	0.4389	0.4286	0.4285	0.5436	0.5371	0.5312	0.5316	0.6744
	ALBERT-base-v2	0.4241	0.4195	0.4256	0.5387	0.5439	0.5265	0.5199	0.7044
	Nepali-BERT	0.4731	0.4798	0.4624	0.5902	0.5887	0.6143	0.5854	0.7138
Visual	VGG-16	0.4218	0.4053	0.3977	0.574	0.5535	0.5618	0.5542	0.6367
	ResNet-50	0.4197	0.4112	0.389	0.567	0.543	0.5415	0.5368	0.6017
	ResNet-101	0.4369	0.4517	0.4308	0.5868	0.5187	0.5226	0.5143	0.5911
Multimodal	VGG-16 + DeBERTa	0.4696	0.4885	0.4517	0.55	0.5815	0.5827	0.5821	0.6833
	VGG-16 + mBERT	0.4668	0.4513	0.4571	0.5798	0.5547	0.5604	0.5561	0.6478
	VGG-16 + NepaliBERT	0.4701	0.4713	0.4579	0.5836	0.5798	0.6154	0.5787	0.7064
	ResNet-101 + DeBERTa	0.4064	0.3925	0.3936	0.51	0.5538	0.5314	0.5257	0.71
	ResNet-101 + mBERT	0.4104	0.3984	0.4014	0.5254	0.57	0.5578	0.5605	0.6967
	ResNet-101 + NepaliBERT	0.485	0.4763	0.4628	0.6304	0.5895	0.6135	0.5872	0.7099
	Pro-Cap	0.5164	0.5101	0.5132	0.6710	0.5991	0.6198	0.6093	0.7113
	ProHarMeme	0.4796	0.4667	0.4731	0.6108	0.6121	0.6219	0.6170	0.7090
	MultimodalGPT	0.4432	0.5544	0.4882	0.5701	0.5210	0.5342	0.5061	0.6175
	CogVLM	0.4115	0.5332	0.4981	0.5850	0.5213	0.5326	0.5069	0.6348
	LLaVA	0.4327	0.5174	0.4721	0.6023	0.5125	0.5491	0.4892	0.6157
	Proposed Model (MemeNePAL)	**0.564**	**0.5721**	**0.5489**	**0.7026**	**0.6914**	**0.7257**	**0.6785**	**0.7259**

Table 5: Performance of various models for sentiment and hate analysis in combined dataset

Model	Dataset	Hate Speech	Sentiment
MemeNePAL	Monolingual	**0.5301**	**0.6331**
MemeNePAL - MLM		0.4916	0.5883
MemeNePAL - ICE		0.5112	0.5970
MemeNePAL - CITE		0.5090	0.5921
MemeNePAL	Code-mixed	**0.4651**	**0.5497**
MemeNePAL - MLM		0.4340	0.5116
MemeNePAL - ICE		0.4401	0.5202
MemeNePAL - CITE		0.4388	0.5190
MemeNePAL	Combined	**0.5489**	**0.6785**
MemeNePAL - MLM		0.5100	0.6369
MemeNePAL - ICE		0.5217	0.6398
MemeNePAL - CITE		0.5209	0.6443

Table 6: Ablation analysis of MemeNePAL. The values given are F1-scores.

Figure 7: Some misclassification errors made by the model

Error Analysis

Hate speech classification and sentiment analysis within memes is inherently challenging and more pronounced in low-resource languages. As seen in Figure 7, it can be seen that the figure on the left that the ground truth is hate speech whereas the model predicts it as non-hateful speech. This is possibly because the hate speech is presented in a more subtle way. The model needs to know a lot of background information about politicians shown in the memes to properly understand the context. Similarly, the figure of the left of Figure 7 shows that the meme with negative sentiment is predicted as positive. This can be attributed to the usage of words meaning positive meanings to represent negative sentiment. Future work should incorporate interpretability techniques to better understand model reasoning and patterns in misclassifications, which can guide the development of more culturally aware and explainable multimodal systems.

Conclusion

In conclusion, this work presents a significant step toward inclusive and culturally aware multimodal content moderation by introducing NeMeme—the first annotated dataset for hate speech and sentiment analysis in Nepali and code-mixed memes—and proposing MemeNePAL, a prompt-assisted multimodal framework designed for low-resource and code-mixed language settings. Through a comprehensive exploration of various models, encompassing textual, visual, and multimodal approaches, our findings show the significance of amalgamating both textual and visual information for better meme understanding. While our current study treats sentiment and hate speech detection as separate tasks to establish strong individual baselines, future work can explore joint optimization via multi-task learning, which may further enhance performance. We also plan to expand the dataset and evaluate emerging foundation models.

Limitations

The limited linguistic and cultural coverage, challenges with nuanced or context-specific memes, variable performance across domains, high computational demands, and ethical concerns around bias and censorship in automated moderation should be noted. Future work should address these by expanding the dataset, enhancing model robustness, and exploring ethical implications.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (Granted No. 62276190).

References

Akhtar, M. S.; Ekbal, A.; and Bhattacharyya, P. 2016. Aspect based Sentiment Analysis in Hindi: Resource Creation and Evaluation. In Calzolari, N.; Choukri, K.; Declerck, T.; Goggi, S.; Grobelnik, M.; Maegaard, B.; Mariani, J.; Mazo, H.; Moreno, A.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2703–2709. Portorož, Slovenia: European Language Resources Association (ELRA).

Ali, M. Z.; Rauf, S.; Javed, K.; Hussain, S.; et al. 2021. Improving hate speech detection of Urdu tweets using sentiment analysis. *IEEE Access*, 9: 84296–84305.

Artetxe, M.; Labaka, G.; and Agirre, E. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 789–798.

Bandyopadhyay, D.; Kumari, G.; Ekbal, A.; Pal, S.; Chatterjee, A.; and BN, V. 2023. A Knowledge Infusion Based Multitasking System for Sarcasm Detection in Meme. In Kamps, J.; Goeuriot, L.; Crestani, F.; Maistro, M.; Joho, H.; Davis, B.; Gurrin, C.; Kruschwitz, U.; and Caputo, A., eds., *Advances in Information Retrieval*, 101–117. Cham: Springer Nature Switzerland. ISBN 978-3-031-28244-7.

Bhandari, A.; Shah, S. B.; Thapa, S.; Naseem, U.; and Nasim, M. 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*, 1993–2002.

Bohra, A.; Vijay, D.; Singh, V.; Akhtar, S. S.; and Shrivastava, M. 2018. A Dataset of Hindi-English Code-Mixed Social Media Text for Hate Speech Detection. In Nissim, M.; Patti, V.; Plank, B.; and Wagner, C., eds., *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, 36–41. New Orleans, Louisiana, USA: Association for Computational Linguistics.

Cao, R.; Hee, M. S.; Kuek, A.; Chong, W.-H.; Lee, R. K.-W.; and Jiang, J. 2023. Pro-cap: Leveraging a frozen vision-language model for hateful meme detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, 5244–5252.

Cao, R.; Lee, R. K.-W.; Chong, W.-H.; and Jiang, J. 2022. Prompting for Multimodal Hateful Meme Classification. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 321–332. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Chakravarthi, B. R.; Jose, N.; Suryawanshi, S.; Sherly, E.; and McCrae, J. P. 2020a. A Sentiment Analysis Dataset for Code-Mixed Malayalam-English. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, 177.

Chakravarthi, B. R.; Muralidaran, V.; Priyadharshini, R.; and McCrae, J. P. 2020b. Corpus Creation for Sentiment Analysis in Code-Mixed Tamil-English Text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, 202–210.

Das, M.; and Mukherjee, A. 2023. BanglaAbuseMeme: A Dataset for Bengali Abusive Meme Classification. *arXiv preprint arXiv:2310.11748*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Gong, T.; Lyu, C.; Zhang, S.; Wang, Y.; Zheng, M.; Zhao, Q.; Liu, K.; Zhang, W.; Luo, P.; and Chen, K. 2023. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Identity Mappings in Deep Residual Networks. *Computer Vision–ECCV 2016*, 630–645.

He, P.; Liu, X.; Gao, J.; and Chen, W. 2020. DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION. In *International Conference on Learning Representations*.

Hessel, J.; Holtzman, A.; Forbes, M.; Le Bras, R.; and Choi, Y. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7514–7528.

Hossain, E.; Sharif, O.; and Hoque, M. M. 2022a. MemoSen: A Multimodal Dataset for Sentiment Analysis of Memes. In Calzolari, N.; Béchet, F.; Blache, P.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Thirteenth Language Resources and Eval-*

uation Conference, 1542–1554. European Language Resources Association.

Hossain, E.; Sharif, O.; and Hoque, M. M. 2022b. MUTE: A Multimodal Dataset for Detecting Hateful Memes. In Hanqi, Y.; Zonghan, Y.; Ruder, S.; and Xiaojun, W., 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*, 32–39. Online: Association for Computational Linguistics.

Jafri, F. A.; Siddiqui, M. A.; Thapa, S.; Rauniyar, K.; Naseem, U.; and Razzak, I. 2023. Uncovering Political Hate Speech During Indian Election Campaign: A New Low-Resource Dataset and Baselines. In *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media.*

Ji, J.; Ren, W.; and Naseem, U. 2023. Identifying Creative Harmful Memes via Prompt based Approach. In *Proceedings of the ACM Web Conference 2023*, 3868–3872.

Kannan, R.; and Rajalakshmi, R. 2022. Multimodal Code-Mixed Tamil Troll Meme Classification using Feature Fusion. In Chakravarthi, B. R.; Murugappan, A.; Chinnappa, D.; Hane, A.; Kumeresan, P. K.; and Ponnusamy, R., eds., *Proceedings of the First Workshop on Multimodal Machine Learning in Low-resource Languages*, 1–8. IIIT Delhi, New Delhi, India: Association for Computational Linguistics.

Kannan, R. R.; Ravikiran, M.; and Rajalakshmi, R. 2022. MMOD-MEME: A Dataset for Multimodal Face Emotion Recognition on Code-Mixed Tamil Memes. In *International Conference on Speech and Language Technologies for Low-resource Languages*, 335–345. Springer.

Karim, M. R.; Dey, S. K.; Islam, T.; Shajalal, M.; and Chakravarthi, B. R. 2022. Multimodal hate speech detection from bengali memes and texts. In *International Conference on Speech and Language Technologies for Low-resource Languages*, 293–308. Springer.

Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; and Testuggine, D. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems*, 33: 2611–2624.

Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; and Soricut, R. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*.

Li, J.; Yuan, J.; and Li, Z. 2023. TP-FER: An Effective Three-phase Noise-tolerant Recognizer for Facial Expression Recognition. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(3): 1–17.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Maity, K.; Jha, P.; Saha, S.; and Bhattacharyya, P. 2022. A multitask framework for sentiment, emotion and sarcasm aware cyberbullying detection from multi-modal code-mixed memes. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1739–1749.

Maity, K.; Saha, S.; and Bhattacharyya, P. 2022. Cyberbullying Detection in Code-Mixed Languages: Dataset and Techniques. In *2022 26th International Conference on Pattern Recognition (ICPR)*, 1692–1698. IEEE.

Mokady, R.; Hertz, A.; and Bermano, A. H. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Pereira-Kohatsu, J. C.; Quijano-Sánchez, L.; Liberatore, F.; and Camacho-Collados, M. 2019. Detecting and monitoring hate speech in Twitter. *Sensors*, 19(21): 4654.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Rajput, K.; Kapoor, R.; Rai, K.; and Kaur, P. 2022. Hate me not: detecting hate inducing memes in code switched languages. *arXiv preprint arXiv:2204.11356*.

Romim, N.; Ahmed, M.; Talukder, H.; and Saiful Islam, M. 2021. Hate speech detection in the bengali language: A dataset and its baseline evaluation. In *Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCACI 2020*, 457–468. Springer.

Simonyan, K.; and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR 2015)*. Computational and Biological Learning Society.

Sitaula, C.; Basnet, A.; Mainali, A.; Shahi, T. B.; et al. 2021. Deep learning-based methods for sentiment analysis on Nepali COVID-19-related tweets. *Computational Intelligence and Neuroscience*, 2021.

Suryawanshi, S.; Chakravarthi, B. R.; Verma, P.; Arcan, M.; McCrae, J. P.; and Buitelaar, P. 2020. A Dataset for Troll Classification of TamilMemes. In Jha, G. N.; Bali, K.; L., S.; Agrawal, S. S.; and Ojha, A. K., eds., *Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation*, 7–13. European Language Resources Association (ELRA). ISBN 979-10-95546-67-2.

Thapa, S.; Rauniyar, K.; Shiwakoti, S.; Poudel, S.; Naseem, U.; and Nasim, M. 2023. NEHATE: Large-Scale Annotated Data Shedding Light on Hate Speech in Nepali Local Election Discourse. In *26th European Conference on Artificial Intelligence*.

Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; Song, X.; et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Waseem, Z.; and Hovy, D. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, 88–93.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes
   (e) Did you describe the limitations of your work? Yes
   (f) Did you discuss any potential negative societal impacts of your work? Yes
   (g) Did you discuss any potential misuse of your work? Yes
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing: Not Applicable
3. Additionally, if you are including theoretical proofs: Not Applicable
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, please visit https://github.com/therealthapa/memenepal where we will upload the codes. We have made data available in Zenodo.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? No
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? No
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Not Applicable
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Not Within Scope
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? Not Applicable
   (b) Did you mention the license of the assets? Not Applicable
   (c) Did you include any new assets in the supplemental material or as a URL? Yes
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Not Applicable
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? Yes
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? Zenodo repository includes answers to the required questions.
6. Additionally, if you used crowdsourcing or conducted research with human subjects: Not Applicable

Ethics Statement

Our work upholds ethical standards by promoting inclusivity and well-being. Annotators were fairly compensated and given clear guidelines to reduce bias. However, we are mindful of the risks associated with automated content moderation, including potential bias and over-censorship. To mitigate these risks, we advocate for a human-in-the-loop approach, where automated systems are supplemented with human oversight to ensure context-sensitive and fair decision-making.