Cross Platform MultiModal Retrieval Augmented Distillation for Code-Switched Content Understanding

Cross Platform MultiModal Retrieval Augmented Distillation for Code-Switched Content Understanding

Surendrabikram Thapa
Virginia Tech
Blacksburg, Virginia, USA
surendrabikram@vt.edu

Hariram Veeramani
University of California Los Angeles
Los Angeles, California, USA
hariram@ucla.edu

Imran Razzak
Mohamed bin Zayed University of Artificial Intelligence
Abu Dhabi, UAE
University of New South Wales
Sydney, Australia
imran.razzak@mbzuai.ac.ae

Roy Ka-Wei Lee
Singapore University of Technology and Design
Singapore
roy_lee@sutd.edu.sg

Usman Naseem
Macquarie University
Sydney, New South Wales, Australia
usman.naseem@mq.edu

Abstract

In the era of digital communication and social media, the proliferation of multimodal content, such as code-switched memes, has become a ubiquitous form of expression. This phenomenon is especially significant for low-resource languages like Nepali, where the need for sentiment analysis and hate speech detection remains unmet due to the unavailability of publicly available datasets. To address this gap, we provide ENeMeme, an annotated dataset of 4,211 code-switched memes in the Nepali-English language for sentiment and hate speech. While the previous state-of-the-art methods of meme analysis particularly focus on high-resource language, they fail to perform well in low-resource language. To bridge this gap, our paper also builds on existing literature to adapt a novel multimodal model, MM-RAD, designed to understand code-switched Nepali-English memes, leveraging both textual and visual content. The model's effectiveness is analyzed across various retrieval platforms. Our proposed MM-RAD demonstrates superior performance in sentiment analysis and hate speech detection compared to individual baseline models. The dataset can be availed through https://github.com/therealthapa/crossplatform

CCS Concepts

• **Information systems** → **World Wide Web**; • **Computing methodologies** → *Information extraction*; *Computer vision problems.*

Keywords

Multimodal Processing, Meme Analysis, Code-Switched Languages, Natural Language Processing

NOTE: PAPER CONTAINS OFFENSIVE CONTENTS.

WWW Companion '25, April 28-May 2, 2025, Sydney, NSW, Australia

ACM ISBN 979-8-4007-1331-6/2025/04
https://doi.org/10.1145/3701716.3718389

ACM Reference Format:
Surendrabikram Thapa, Hariram Veeramani, Imran Razzak, Roy Ka-Wei Lee, and Usman Naseem. 2025. Cross Platform MultiModal Retrieval Augmented Distillation for Code-Switched Content Understanding. In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3701716.3718389

1 Introduction

In the era of digital connectivity, internet has revolutionized the way we share ideas, fostering global interactions and the rapid dissemination of content across linguistic and cultural boundaries [22]. In this digital landscape, memes, as a unique form of multimedia expression, have emerged as a powerful medium for conveying humorous, thought-provoking, and often satirical opinions [29]. Memes amalgamate textual and visual elements, bridging the language gap to deliver impactful messages that resonate with diverse audiences worldwide.

Code-switching, the practice of alternating between two or more languages within a single conversation, is a common phenomenon in multilingual societies [27]. Code-switched languages are widely used across social media platforms and have become an integral part of digital communication. In particular, code-switched languages are prevalent in the creation and dissemination of memes, reflecting the dynamic linguistic environment of internet culture. Understanding code-switched memes is essential for several reasons. Memes encapsulate a wide range of sentiments, from humor and satire to expressions of frustration and anger. Therefore, developing algorithms that can automatically analyze the sentiment expressed in code-switched memes is of paramount importance. These algorithms have the potential to enhance our ability to comprehend the nuances of the content being shared and enable more effective engagement with digital communities.

However, the world of code-switched memes is not without its challenges. These multimodal expressions also feature negative sentiments, including hate speech. The rapid spread of hateful content on social media platforms poses a significant threat to online discourse and community well-being. Human moderation

of such content is not only impractical due to the vast volume of data but also presents risks to the mental well-being of human moderators who may be subjected to disturbing and violent content [7]. To address these challenges, significant efforts have been made to develop automated systems that can understand sentiment in code-switched memes and effectively moderate hateful content. However, these algorithms are highly data-dependent, relying on extensive and diverse datasets to achieve optimal performance.

In the context of low-resource languages, such as Nepali, these efforts have been limited [28]. To date, there is no available multimodal datasets for sentiment analysis and hate speech identification in Nepali-English code-switched content. Recognizing this gap, we have embarked on an initiative to bridge this void. We have curated and annotated a collection of 4,211 code-switched memes in the Nepali-English language, where the Nepali language is code-switched with English. This dataset is an invaluable resource for training and evaluating algorithms tailored to the specific linguistic and cultural nuances of the Nepali-speaking digital community.

In addition to providing this unique dataset, we also offer baseline models for both sentiment analysis and hate speech classification in code-switched memes. Moreover, we propose MM-RAD, a novel algorithm designed to enhance the accuracy and efficiency of sentiment analysis and hate speech identification within the Nepali-English code-switched meme ecosystem. Our framework based on teacher-student model addresses the problem of limited labeled datasets for code-switched memes which is very common in low-resource languages like Nepali. Our semi-supervised approach allows the student model to leverage both the labeled and pseudo-labeled data through distillation, leading to enhanced classification performance compared to using just the limited labeled dataset. Our contribution aims to empower researchers, developers, and online platforms in their efforts to maintain a positive and inclusive digital space while respecting linguistic diversity and cultural sensitivity.

2 Related Works

In recent years, there has been a surge in research focused on understanding code-switched language, where individuals alternate between two or more languages or language varieties within a single conversation. Paul et al. [23] studied code-switched Hindi-English language during the COVID-19 pandemic to understand cyberbullying on Twitter. The authors also propose an ensemble-based approach to enhance the model's classification ability in distinguishing between posts containing instances of bullying and those that do not. Similarly, Chakravarthi et al. [5] curated a dataset of 15,744 YouTube comments in Tamil-English code-switched context. The comments were annotated for the presence of specific sentiments like positive, negative, neutral and mixed feelings. The baselines were also established where logistic regression and Random Forest (RF) performed high scores as compared to other baseline algorithms. Similarly, Kannadaguli [17] provided a dataset for Kannada-English language for analysis of sentiment in online comments. Additionally, Salman et al. [26] curated a corpus of 1,030 texts of code-switched English-Urdu language annotated for different propaganda techniques. The corpus was annotated for the presence of 20 different propaganda techniques. The authors tested the classification ability of 11 different algorithms out of which XLM-RoBERTa trained on Roman Urdu dataset performed the best.

While the code-switched language analysis allows for a deeper exploration of multilingual communication patterns, memes offer a rich platform for the study of multimedia-based expressions in the digital age. Most research currently are predominantly focused on understanding hate speech and sentiment in unimodal settings. This tendency has resulted in a gap in our understanding of the intricate interplay between textual and visual elements within digital discourse.

In recent years, as the memes have become increasingly common, there have been efforts in understanding the memes [1]. Kiela et al. [18] developed a manually created set of 8500 memes for the analysis of hateful content in memes. The authors proposed various unimodal and multimodal methods where multimodal pretraining-based methods like ViLBERT CC and Visual BERT COCO performed higher than all unimodal methods. These results suggest that the interplay between both image and text is important to understand in order to classify hate speech in memes. Apart from using multimodal pretraining-based methods, methods involving separate branches for text and visual processing are also in use. Constantin et al. [6] proposed the use of attention and bidirectional LSTM layers for processing text and ResNet-like convolution network for processing visual information.

While these studies provide a nuanced understanding of memes, it is crucial to note that understanding the interplay between code-switched language and memes. Rajput et al. [25] curated a dataset of 1218 memes annotated for hate speech and satire in Hindi-English (Hinglish). They additionally labeled memes for non-offensive speech. Different baseline models were established to classify non-offensive speech, hate speech, and satirical speech. They also proposed a multichannel CNN-LSTM to process both image and text to enable the model to understand both modalities. Similarly, Hossain et al. [14] proposed MemoSen, a dataset in Bangla code-switched, code-mixed and only Bangla memes. The authors annotated the memes for negative, neutral, and positive sentiments. The authors propose using feature-level fusion (early fusion) and decision-level fusion (late fusion) to improve classification performance over the baseline models. Additionally, Hossain et al. [15] shared MUTE, a meme dataset annotated for the presence of hate speech. They also proposed different textual, visual and multimodal baselines. Their analysis showed that VGG16+B-BERT performed the best among all the textual, visual and multimodal baselines. The overarching takeaway from these studies is that the interplay between code-switched language and memes is a complex and evolving area of research. Including linguistic and cultural diversity in meme analysis adds a layer of intricacy that necessitates tailored approaches and models. As we advance in this field, it becomes increasingly clear that the fusion of textual and visual elements in memes requires specialized techniques and deep understanding, making this an exciting and fertile ground for future exploration and innovation. While research is being done in multiple code-switched languages, research in the Nepali language, which is spoken by over 40 million people worldwide, is lacking research. With the widespread use of social media worldwide, work in the field of code-switched NLP in the Nepali language is a pressing need. Establishing language resources is an important task in propelling the Nepali language

Table 1: An overview of related datasets in meme classification and analysis

Work	Language	Sentiment	Hate Speech	Multimodal
Paul et al. [23]	Hindi-English	✗	✗	✗
Chakravarthi et al. [5]	Tamil-English	✓	✗	✗
Salman et al. [26]	Urdu-English	✗	✓	✗
Kannadaguli [17]	Kannada-English	✓	✗	✗
Kiela et al. [18]	English	✗	✓	✓
Rajput et al. [25]	Hindi-English	✗	✓	✓
Ours (ENeMeme)	Nepali-English	✓	✓	✓

[28]. We address this necessity by establishing a comprehensive resource of code-switched memes for the Nepali-English language. We annotated the dataset ENeMeme for both sentiment and the presence of hate speech. See Table 1 for an overview of related datasets, including ours (ENeMeme), in meme classification and analysis.

3 Dataset

Memes, in the context of this research, are images that incorporate textual information within the image itself. The overall schema is described below (also in section A.1).

3.1 Dataset Collection and Deduplication

We initially collected 7,000 memes from the social media platforms Twitter, Reddit, Instagram, and Facebook. As our data collection was a collaborative effort, with various contributors and data sources involved, it was inevitable that we encountered duplicate entries. The presence of duplicates can introduce errors into our subsequent analyses and compromise the overall quality of our results. Therefore, it became imperative to implement a rigorous deduplication process before proceeding with dataset analysis. To tackle this, our deduplication process consisted of two successive steps using two distinct deduplication tools. The first tool we used was dupeGuru[1], a reliable software for identifying and eliminating duplicate files. Following this, we turned to the Duplicate Image Finder, a Python package known as difPy[2], which further aided us in the removal of duplicate images. This approach effectively helped us to remove all duplicates.

With our dataset now clean of duplicate entries, we extract the textual content within the meme images using Google OCR Vision API[3]. This API allowed us to perform optical character recognition on the images, effectively converting the text contained within the memes into machine-readable format. This, in turn, facilitated further data processing and analysis.

3.2 Filtering Criteria

To ensure the integrity and quality of our dataset, we applied a set of stringent filtering criteria. We screened memes to confirm the presence of code-switching between Nepali and English. Additionally, memes with excessively low image quality, to the extent that optical character recognition (OCR) technology could not accurately detect and convert the embedded text into a machine-readable format, were excluded from our dataset. All filtering were done manually by annotators.

3.3 Annotation Schema

The precise annotation of data is critical, as it facilitates the consistency, reliability, and overall validity of the dataset. Further analysis drawn on the dataset would depend on annotation quality.

3.3.1 Three Phase Annotation. Our annotation started with a set of explicit and unambiguous instructions. These instructions underwent iterative revisions until all annotators had a clear understanding. To further mitigate any ambiguity and ensure uniformity, we executed a well-established three-phase annotation approach [2]:

Pilot Run: The initial phase encompassed a pilot run in which 50 images were annotated. The primary goal here was to confirm that all annotators comprehended the annotation instructions. Given the nuanced nature of identifying hate speech and sentiment, some discrepancies surfaced during this phase, particularly concerning the presence of hate speech. Subsequently, we refined the instructions to address these disparities.

Revised Instructions: In the second phase, each annotator independently annotated 200 images, adhering to the revised set of instructions. This phase aimed to validate the clarity and efficacy of the modified instructions. Results from this phase informed further refinements of the instructions and ensured that annotators consistently identified hate speech and sentiments in accordance with the updated guidelines.

Consolidation Phase: The third and final annotation phase involved a collaborative annotation of 50 images during a group meeting, where all annotators discussed any remaining disparities in their annotations from the second phase and worked to reach a consensus. This phase played a pivotal role in enhancing the overall consistency of annotations, making instructions more evident to the annotators, and providing an opportunity to uncover any lingering ambiguities in the instructions.

3.3.2 Guidelines for Presence of Hate. The meme was annotated as containing 'Hate' if it includes content that expresses or incites hatred, discrimination, or harm towards an individual, group, community, or organization based on attributes such as race, religion, ethnicity, gender, sexual orientation, or any other characteristic. Fig. 1 shows some of the examples of memes labeled as hateful memes.

On the contrary, the meme was annotated as 'No Hate' if it does not exhibit any form of hate speech or hateful content. Memes

[1]https://github.com/arsenetar/dupeguru
[2]https://github.com/elisemercury/Duplicate-Image-Finder
[3]https://cloud.google.com/vision/docs/ocr

Figure 1: Examples of memes annotated as containing hate speech. These memes express or incite hatred, discrimination, or harm towards individuals or groups based on attributes like race, religion, ethnicity, gender, or sexual orientation. Translation of memes: Left-> I am thinking of dying because of tension but Bijay Shahi is saying that he will make dead ones alive. Center-> How can you street kid be single as soon as you see a beautiful girl? Right-> You look like an actress from behind but I nearly died of your ugliness when I saw your face.

that convey humor, satire, or opinions without promoting hatred should be labeled as 'No Hate'. Fig. 2 shows some of the examples of memes labeled as non-hateful memes.

3.3.3 Annotation Guidelines for Sentiment. Memes were labeled as 'Neutral' when they do not express any discernible positive or negative sentiment. Memes that are informational, factual, or do not convey any emotional tone were categorized as 'Neutral'.

Similarly, memes were annotated as 'Positive' when they contain content that is uplifting, humorous or promotes a positive outlook. This includes content that generates laughter, joy, or admiration.

Furthermore, memes were labeled as 'Negative' when they evoke negative emotions, such as anger, sadness, frustration, or promote harmful ideas or sentiments. Memes with content that is critical, derogatory, or mocking were annotated as 'Negative'. Fig. 3 shows the examples of memes for presence of different sentiments.

3.3.4 Demographic Profile of Annotators. The annotators included two Master's students and one working professional (with a Bachelor's degree). The annotators were from the field of Computer Science and Social Sciences. The team included two male and one female annotator with an age range between 24 to 30 years. The annotators came from diverse cultural and linguistic backgrounds, which provided a broad perspective on the data. Additionally, all the annotators had prior experience with annotation tasks, bringing valuable expertise to the process.

3.4 Dataset Statistics

To quantitatively assess inter-annotator agreement, we employed Fleiss' Kappa (κ) as a metric of inter-rater reliability. The inter-annotator agreement for the presence of hate speech κ_{Hate} was 0.69, whereas the agreement score for annotation of sentiment $\kappa_{Sentiment}$ was 0.65. The inter-annotator agreement is comparable to the previous studies involving hate speech and sentiment analysis. The dataset statistics are shown in Table 2.

Table 2: Statistics of ENeMeme Dataset

Problem	Labels	#memes	Avg. Characters	Avg. Words
Hate Speech	Hate	1,017	78.84	16.53
	Non Hate	3,194	747.55	16.04
Sentiment	Neutral	2,070	74.47	16.24
	Positive	1,108	74.99	16.10
	Negative	1,033	77.85	17.21

4 Methodology

In this section, we describe the baseline and our proposed method for better understanding hate speech and sentiment in code-switched Nepali-English.

4.1 Baselines

We established baselines with various unimodal and multimodal methods. We used various textual methods like BERT-base [9], RoBERTa [21], HateBERT [4], mBERT [9], and NepaliBERT[4]. In addition to them, we also used unimodal visual methods like ResNet-50 and ResNet-101 [13]. We also used two self-implemented naive multimodal fusion methods namely vanilla ResNet-101+HateBERT and vanilla ResNet-101+NepaliBERT. We also use three multimodal methods which have shown good performance in the task of meme analysis: ProHarMeme [16], Pro-Cap [3], and MOMENTA [24]. Additionally, we also use recent large vision language models (LVLMs) like CogVLM [31], MultimodalGPT [11], and LLaVA [20].

Apart from these baselines, we propose a tailored solution of using MM-RAD to address the problem of hate speech and sentiment analysis.

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

Figure 2: Examples of memes annotated as not containing hate speech. Translation of memes: Left-> Me to my neighbour: Aunty did my mom give you the keys to our house? Center-> Mom: Go and return this item to the shop. Right-> someone: you look cute, what is the secret? Me: shows Snapchat.

Figure 3: Examples of memes displaying different sentiments: negative (left), neutral (center), and positive (right). Translation of memes: Left-> Random guy calls you younger brother while chatting. Gangster Nias: Who are you pussy? Where do you live? Tell me your phone number. Center-> My friends before eating food. Right-> Sandeep!! I cannot live without you. From your love. Love you.**

4.2 Multimodal Retrieval Augmented Distillation (MM-RAD)

Understanding the semantic meaning of memes is very important, and the posts similar to them serve as complementary information. In addition, the comments or replies, collectively referred to as samples, in memes also hint at visual-linguistic relations in memes. To allow the model to understand complementary information better, we propose MM-RAD as shown in Fig. 4. This model, built on retrieval model by Xu and Li [32] makes modifications by adapting it to the context of code-switched language.

4.2.1 Multimodal Content Retrieval. To assess the similarity between posts, we adopt a holistic approach that balances both image and text features. The image feature is acquired using ResNet-152, a pre-trained model on ImageNet [8]. We extract the representation from its final pooling layer. For text features, we employ SimCSE, known for its effectiveness in measuring similarity [10]. To calculate the similarity score (s_i) between a query post and the $i-th$ multimedia post in our dataset, we use the following formula:

$$s_i = \alpha s_i^I + (1-\alpha)s_i^T + (1-\beta)|s_i^I - s_i^T|^2 \tag{1}$$

Here, s_i^I represents image similarity, and s_i^T represents text similarity, with their effects balanced by the parameter α and β is a cyclic balance regularization hyperparameter. The weight α, which balances the contributions of image and text, is empirically estimated using averaged statistics derived from both the wild datasets (for retrieval) and all experimental data (as a query set):

$$\alpha = \frac{T_{mean}}{I_{mean} + T_{mean}} \tag{2}$$

$$I_{mean} = \frac{1}{MK}\sum_{m=1}^{M}\sum_{k=1}^{K} p_{m,k} \tag{3}$$

$$T_{mean} = \frac{1}{MK}\sum_{m=1}^{M}\sum_{k=1}^{K} q_{m,k} \tag{4}$$

Where $p_{m,k}$ and $q_{m,k}$ refer to the image and text similarity between the $m-th$ query and its rank-k most similar post retrieved

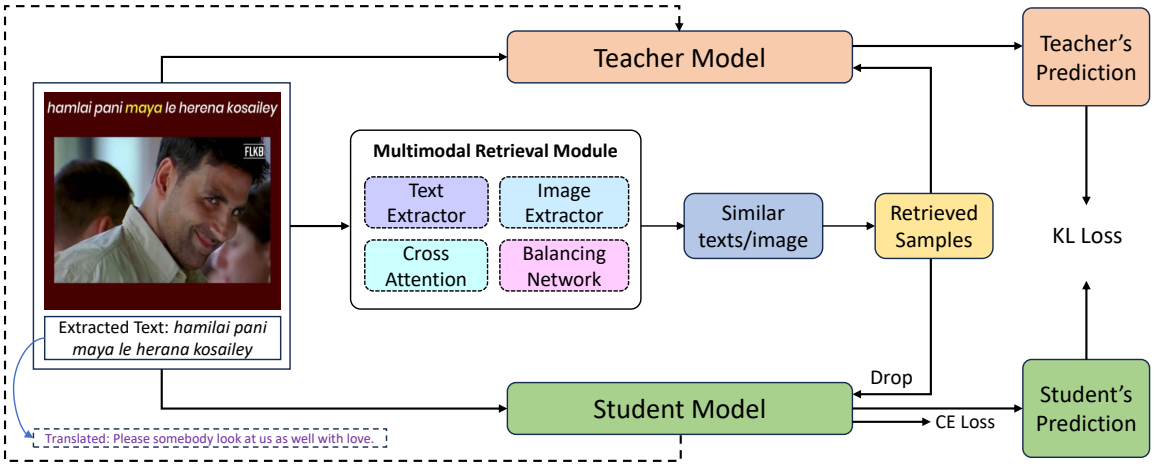

Figure 4: The architecture of our proposed MM-RAD model for code-switched context, adapted from Xu and Li [32]. It consists of a Retrieval Module that finds semantically similar text, image, and comments to the input meme. This retrieved data is used to train a Student Model, guided by the predictions of a Teacher Model trained on labeled data. The Student learns from both labeled and pseudo-labeled retrieved data using cross-entropy and KL divergence loss.

with the respective modality. M represents the query set size, and K is the cut-off number of retrieved posts to be selected. This ranking allows us to retrieve the top-K most similar posts for any given query.

In the case of posts with missing or inaccessible comments, comments from retrieved posts can provide valuable insights into visual-lingual relations. To identify high-quality comments, we shortlist consensus comments from the comment pool (P) collected from all retrieved posts. Consensus comments exhibit higher semantic similarity to other comments in P and are ranked using the following formula:

$$q_i = \frac{1}{|P|} \sum_{p' \in P} Sim(p_i, p') \tag{5}$$

Here, q_i represents the consensus score of the i^{th} comment ($p'_i \in P$), and $Sim(p_i, p')$ measures the similarity between comments using SimCSE. $|P|$ is the size of the comment pool. To bridge the gap between the image and text modalities, for any post, we retrieve the top-N consensus comments to form a set (C). This set helps in enhancing the understanding of the visual-lingual relationships and fosters cross-modal learning.

4.2.2 Multimodal Retrieval Augmented Distillation. The data comprising the retrieved image-text pairs is further explored in the context of multimodal classification, where it is often structured as a labeled parallel dataset, denoted as $L = \{x_i, c_i, y_i\}_{i=1}^{l}$, where x_i represents an image-text pair, c_i signifies the retrieved set of N comments, and y_i denotes a label specific to the classification task at hand.

Labeled datasets like L are frequently constrained in scale, which can give rise to concerns of over-fitting. However, the retrieved posts, similar in nature to the data in L, form an unlabeled set ($U = \{x'_i, c_i\}_{i=1}^{Kl}$). These unlabeled data can be integrated with the labeled dataset to enhance the training process. It's worth noting that x'_i, a retrieved image-text pair corresponding to x_i, shares the same consensus comments c_i with its labeled counterpart. Combining L and U facilitates more robust learning in a semi-supervised fashion.

Our approach adopts self-training based on the widely recognized teacher-student framework. Initially, a teacher model, serving as the classifier, is trained on the labeled data in L to acquire task-specific knowledge. The teacher then pseudo-labels the unlabeled data in U with soft labels, essentially serving as 'teaching samples'. Subsequently, a student model, sharing the same architecture as the teacher, is trained using both the pseudo-labeled U and the labeled L.

During the training of both the teacher and student models, their modality fusion mechanism is fed with comment features. This integration enables the models to explore cross-modal interactions while being aware of the comments associated with each post.

In practice, a specific training strategy is employed. The student model is trained with 50% of the comments randomly dropped, while the teacher model uses the full set of comments. This approach allows the student model to learn from noisy data, promoting better generalization, rather than merely mimicking the behavior of the teacher model.

The teacher model is trained using the cross-entropy loss for classification, and the Kullback-Leibler (KL) divergence [19] loss is additionally incorporated into the student training process. The loss function is defined as:

$$\mathcal{L} = \frac{1}{|L|} \sum_{i \in L} y_i \log(y_i) + \frac{1}{|U|} \sum_{i \in U} \mathrm{KL}(t_i || s_i) \tag{6}$$

Here, $|L|$ and $|U|$ represent the sizes of the datasets L and U, respectively. t_i corresponds to the soft label predicted by the teacher model, while s_i represents the output of the student model. This

training strategy enables the student model to leverage both labeled and pseudo-labeled data, resulting in enhanced classification performance.

Cross-Modal Augmentation: In order to understand the effect of using different sources for augmentation, we use various social media and microblogging sites. We specifically use Twitter, Facebook, Gab, Fox, YouTube and Stormfront.

5 Results

In this section, we describe the baseline results along with the results of our proposed algorithm. We also discuss the strengths and potential pitfalls of our proposed approach with future directions.

5.1 Baseline Results

In Table 4, we present the performance of various baseline models for both Sentiment Analysis and Hate Speech Detection tasks.

For Sentiment Analysis, we observe that the proposed MM-RAD model outperforms all the baseline models, achieving the highest Precision (38.25), Recall (49.72), F1-score (43.24), and Accuracy (60.21). In contrast, BERT-base, RoBERTa, mBERT, HateBERT (A), and NepaliBERT (B) exhibit lower performance across these metrics. We observe impressive performance by multimodal methods like ProHarMeme, Pro-Cap, and MOMENTA. On the other hand, LVLMs like CogVLM, MultimodalGPT, and LLaVA often performed very poorly as compared to the other baselines. Regarding Hate Speech Detection, MM-RAD also excels, obtaining the best Precision (57.33), Recall (74.92), F1-score (64.96), and Accuracy (81.85) compared to the baseline models. The results for the baselines showed similar pattern as the sentiment analysis task where mBERT and ResNet-101 performed the best among textual and visual models. Similarly, ProHarMeme performed the best among multimodal methods whereas CogVLM performed the best among LVLMs. These results signify the superior performance of MM-RAD in both sentiment analysis and hate speech detection.

5.2 Results with MM-RAD

In Table 3, we delve into the performance of MM-RAD using multiple retrieval platforms. We analyze its effectiveness in Sentiment Analysis and Hate Speech Detection across various platforms. In Sentiment Analysis, MM-RAD exhibits the best performance on Twitter with the highest Precision (38.25) and Macro-F1 (43.24) scores, outperforming other platforms. While the performance is slightly lower on Facebook, Fox, Gab, YouTube, and Stormfront, it consistently maintains competitive results across these platforms.

For Hate Speech Detection, MM-RAD again excels on Twitter, achieving the highest Precision (57.33) and Macro-F1 (72.56) scores. Facebook, Fox, Gab, YouTube, Hatecheck, and Stormfront also yield good results, with Facebook and YouTube standing out among them.

5.3 Discussion and Analysis

The results obtained in our study provide valuable insights into the performance of both baseline models and the MM-RAD model with multiple retrieval platforms. In the baseline results, it is evident that the MM-RAD model surpasses the individual baseline models in terms of Sentiment Analysis and Hate Speech Detection. This demonstrates the effectiveness of our proposed model in handling code-switched Nepali-English text and image data. The lower performance of the existing multimodal benchmarks can be attributed to their focus on high-resource languages like English.

Table 3: Analysis of the performance of MM-RAD with multiple retrieval platforms. Prec. represents precision.

Retrieval Platform	Sentiment Analysis		Hate Speech	
	Prec.	Macro-F1	Prec.	Macro-F1
Twitter	**38.25**	**43.24**	**57.33**	**72.56**
Facebook	37.64	42.83	56.15	71.34
Fox	34.28	40.65	51.27	67.18
Gab	33.97	40.2	47.85	61.72
Youtube	37.21	42.56	55.76	70.95
Hatecheck	-	-	54.62	70.83
Stormfront	36.5	41.5	53.08	68.21

When considering the results with MM-RAD, it is clear that the choice of retrieval platform has a notable impact on model performance. Twitter proves to be the most conducive platform for Sentiment Analysis, with the MM-RAD model achieving the highest Precision and Macro-F1 scores. Additionally, Twitter is also the best platform for Hate Speech Detection, reinforcing its significance for this task. In contrast, other platforms like Facebook, Fox, Gab, YouTube, Hatecheck, and Stormfront also provide valuable data, albeit with slightly lower performance. This variability across platforms suggests that the quality and nature of data from different sources can significantly affect the model's performance.

Overall, the MM-RAD model's robustness in handling code-switched Nepali-English text and images, along with the adaptability to various retrieval platforms, makes it a promising tool for sentiment analysis and hate speech detection in multimodal data. The results encourage further exploration and fine-tuning of retrieval sources to enhance the model's capabilities across different social media and microblogging platforms.

5.4 Misclassification of Memes

While our research yielded promising results, there were instances of misclassified memes, which warrant in-depth analysis and consideration for future enhancements. As illustrated in Figure 5, several factors contributed to these misclassification. First, in the leftmost image, the model misclassified a meme featuring a morphed image of the politician Pushpa Kamal Dahal. This misclassification may be attributed to the abundance of images of this prominent figure available on the internet, leading the model to predict a non-hateful sentiment based on the politician's frequency in the dataset. Second, as depicted in the second image from the left, context plays a pivotal role in meme analysis. In this case, the meme's text suggests a seemingly innocuous question, "mention a Newar boy who wants to date me." However, it contains implicit hate, targeting a non-binary gender individual. This subtlety highlights the complexity of meme analysis, where understanding the underlying context is essential for accurate classification. Third, in the third image from the left, the model's ability to provide accurate justifications may be affected by the availability of limited images of certain individuals. In cases where there are fewer pictures associated with a particular

Table 4: Results of various baseline models including MM-RAD

Modality	Models	Sentiment Analysis				Hate Speech Detection			
		Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
Textual	BERT-base	29.50	33.78	31.50	41.95	31.5	45.62	37.27	53.37
	RoBERTa	29.84	34.67	32.07	42.89	33.72	45.99	38.91	55.76
	mBERT	31.09	38.34	34.34	44.63	36.64	55.78	44.22	65.24
	HateBERT (A)	32.17	39.94	35.64	46.84	33.76	49.05	39.99	59.14
	NepaliBERT (B)	33.52	40.67	36.75	48.39	39.76	53.02	45.44	68.53
Visual	Resnet-50	31.28	37.57	34.14	44.23	35.81	50.72	41.98	62.54
	Resnet-101 (C)	32.04	38.45	34.95	45.81	37.62	52.54	43.85	64.95
Multimodal	A + C	33.82	41.36	37.21	50.59	42.21	55.38	47.91	69.83
	B + C	34.36	42.54	38.01	53.76	48.24	58.59	52.91	75.34
	ProHarMeme	36.03	45.98	40.43	57.19	54.44	62.84	58.35	77.14
	Pro-Cap	34.83	43.10	38.55	55.07	50.71	60.60	55.25	76.48
	MOMENTA	34.71	42.98	38.39	54.82	49.96	59.80	54.43	76.31
	CogVLM	31.81	39.06	35.08	46.22	32.71	48.17	38.97	58.66
	MultimodalGPT	28.15	31.23	29.62	40.38	29.17	42.16	34.49	50.78
	LLaVA	27.66	29.91	28.74	38.17	30.20	44.33	35.93	52.34
MM-RAD (Ours)		**38.25**	**49.72**	**43.24**	**60.21**	**57.33**	**74.92**	**64.96**	**81.85**

Actual: Hate **Predicted**: Non-Hate

Actual: Hate **Predicted**: Non-Hate

Actual: Negative **Predicted**: Positive

Actual: Neutral **Predicted**: Negative

Figure 5: Examples of misclassified memes by MM-RAD. Translation of memes (from left): 1. Did you think of flower when you heard the name Pushpa? I am Prachanda! 2. Mention a handsome Newar guy for me. 3. Make such a video that everyone will be waiting for your leaked sex tapes. 4. When I say 'life is like this' to my depressed friends.

entity, the model might struggle to make precise classifications due to the scarcity of contextual information. Fourth, in the rightmost image, the presence of words like "depressed" could influence the model's decision-making process. It may associate such words with negative sentiment, potentially leading to misclassifications. While these initial observations shed light on potential pitfalls, a comprehensive analysis of explainability is crucial to gaining deeper insights into the model's decision-making process. By delving into the technical intricacies and exploring the model's focus and features, we can refine our approach and address these challenges in code-switched meme analysis, further enhancing the accuracy and interpretability of the classification results.

6 Conclusion

Multimodal analysis of code-switched content, particularly in low-resource languages like Nepali, has emerged as a significant and challenging domain in the realm of digital communication. In this study, we have introduced MM-RAD, a novel multimodal model tailored to decode the intricacies of code-switched Nepali-English memes, encompassing both textual and visual elements. The findings from our research underscore the promise of MM-RAD. In the baseline results, it surpassed individual baseline models, affirming its efficacy in sentiment analysis and hate speech detection within the context of code-switched memes. This highlights the model's versatility in understanding and interpreting complex multimodal content. Moreover, our exploration of MM-RAD across various retrieval platforms has illuminated the role of data sources in multimodal analysis. Beyond the technical achievements, our work also addresses a significant need in the development of language resources for code-switched Nepali-English content by providing a resource of 4,211 annotated memes. Future research can explore more nuanced categories of hate, its target and sentiment for Nepali language.

References

[1] Tariq Habib Afridi, Aftab Alam, Muhammad Numan Khan, Jawad Khan, and Young-Koo Lee. 2021. A multimodal memes classification: A survey and open research issues. In *Innovations in Smart Cities Applications Volume 4: The Proceedings of the 5th International Conference on Smart City Applications*. Springer, 1451–1466.

[2] 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*. 1993–2002.

[3] Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang. 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.

[4] Tommaso Caselli, Valerio Basile, Mitrovic Jelena, Granitzer Michael, et al. 2021. HateBERT: Retraining BERT for Abusive Language Detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*. Association for Computational Linguistics, 17–25.

[5] Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020. 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)*. European Language Resources association, Marseille, France, 202–210. https://aclanthology.org/2020.sltu-1.28

[6] Mihai Gabriel Constantin, Dan-Stefan Pârvu, Cristian Stanciu, Denisa Ionascu, and Bogdan Ionescu. 2021. Hateful meme detection with multimodal deep neural networks. In *2021 International Symposium on Signals, Circuits and Systems (ISSCS)*. IEEE, 1–4.

[7] Brandon Dang, Martin J Riedl, and Matthew Lease. 2018. But who protects the moderators? the case of crowdsourced image moderation. *arXiv preprint arXiv:1804.10999* (2018).

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[10] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*. Association for Computational Linguistics (ACL), 6894–6910.

[11] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790* (2023).

[12] Tomáš Hák, Svatava Janoušková, and Bedřich Moldan. 2016. Sustainable Development Goals: A need for relevant indicators. *Ecological indicators* 60 (2016), 565–573.

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

[14] Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. MemoSen: A Multimodal Dataset for Sentiment Analysis of Memes. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 1542–1554.

[15] Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. MUTE: A Multimodal Dataset for Detecting Hateful Memes. In *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.

[16] Junhui Ji, Wei Ren, and Usman Naseem. 2023. Identifying Creative Harmful Memes via Prompt based Approach. In *Proceedings of the ACM Web Conference 2023*. 3868–3872.

[17] Prashanth Kannadaguli. 2021. A Code-Diverse Kannada-English Dataset For NLP Based Sentiment Analysis Applications. In *2021 Sixth International Conference on Image Information Processing (ICIIP)*, Vol. 6. IEEE, 131–136.

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

[19] Solomon Kullback. 1951. Kullback-leibler divergence.

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems* 36 (2024).

[21] 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* (2019).

[22] Pascal Muam Mah, Iwona Skalna, and John Muzam. 2022. Natural Language Processing and Artificial Intelligence for Enterprise Management in the Era of Industry 4.0. *Applied Sciences* 12, 18 (2022), 9207.

[23] Sayanta Paul, Sriparna Saha, and Jyoti Prakash Singh. 2023. COVID-19 and cyberbullying: deep ensemble model to identify cyberbullying from code-switched languages during the pandemic. *Multimedia tools and applications* 82, 6 (2023), 8773–8789.

[24] Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. MOMENTA: A Multimodal Framework for Detecting Harmful Memes and Their Targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. 4439–4455.

[25] Kshitij Rajput, Raghav Kapoor, Kaushal Rai, and Preeti Kaur. 2022. Hate Me Not: Detecting Hate Inducing Memes in Code Switched Languages. (2022).

[26] Muhammad Umar Salman, Asif Hanif, Shady Shehata, and Preslav Nakov. 2023. Detecting Propaganda Techniques in Code-Switched Social Media Text. *arXiv preprint arXiv:2305.14534* (2023).

[27] Sunayana Sitaram, Khyathi Raghavi Chandu, Sai Krishna Rallabandi, and Alan W Black. 2019. A survey of code-switched speech and language processing. *arXiv preprint arXiv:1904.00784* (2019).

[28] Surendrabikram Thapa, Kritesh Rauniyar, Shuvam Shiwakoti, Sweta Poudel, Usman Naseem, and Mehwish Nasim. 2023. NEHATE: Large-Scale Annotated Data Shedding Light on Hate Speech in Nepali Local Election Discourse. In *ECAI 2023*. IOS Press, 2346–2353.

[29] Pham Thai Hoang Tung, Nguyen Tan Viet, Ngo Tien Anh, and Phan Duy Hung. 2023. SemiMemes: A Semi-supervised Learning Approach for Multimodal Memes Analysis. *arXiv preprint arXiv:2304.00020* (2023).

[30] Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, and Francesco Fuso Nerini. 2020. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature communications* 11, 1 (2020), 233.

[31] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079* (2023).

[32] Chunpu Xu and Jing Li. 2022. Borrowing Human Senses: Comment-Aware Self-Training for Social Media Multimodal Classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 5644–5656. https://doi.org/10.18653/v1/2022.emnlp-main.381

A Appendix

A.1 Dataset Collection Schema

The overall schema of our data collection process is given in Fig. 6.

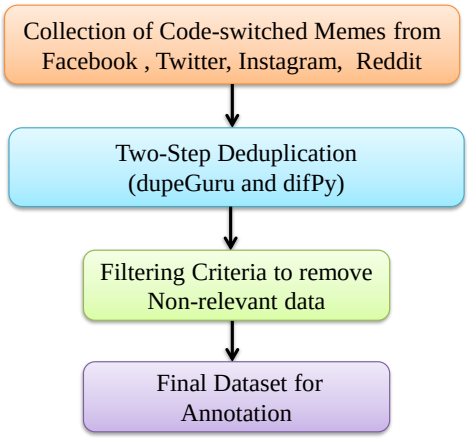

Figure 6: Flow diagram of data collection and preparation process

A.2 Broader Impact

This research carries a profound broader impact, aligning with the United Nations Sustainable Development Goals (SDGs) [12, 30], with a focus on the empowerment of low-resource languages [28]. By developing multimodal analysis tools for code-switched content, particularly in languages like Nepali, often marginalized in the realm of natural language processing, this work directly contributes to SDG #10 (Reduced Inequalities) by ensuring linguistic and cultural inclusivity, bridging the digital divide, and empowering underrepresented communities. its relevance to SDG #16 (Peace, Justice, and Strong Institutions) by addressing the socio-political implications of code-switched content, promoting responsible digital discourse, and contributing to the development of strong institutions that can effectively moderate and govern digital spaces. In doing so, it exemplifies how technological advancements can drive social and cultural change, aligning with multiple SDGs simultaneously.

A.3 Ethical Considerations

Ethical considerations underpinning our research in multimodal analysis, particularly in sentiment analysis and hate speech detection within code-switched content, encompass a commitment to safeguarding data privacy and securing informed consent, vigilance against bias and fairness in data and algorithms, and the responsible handling of harmful content. In our data collection, we ensured the anonymization and de-identification of user data while adhering to platform policies and ethical guidelines. Mitigating bias was a top priority, achieved through diversification of data sources, annotators, and meticulous quantitative analysis of inter-annotator agreements. To ensure the emotional well-being of human annotators exposed to hate speech and negative sentiment content, we provided warnings and access to psychiatrists when needed, emphasizing the importance of their mental health throughout the research process. The annotators were paid fair wages consistent with local rates.

A.4 Negative Impacts and Mitigation

Sharing hate speech examples in the dataset, even for research purposes, might inadvertently amplify harmful content and further expose individuals and communities to offensive and hateful material. Similarly, there is a risk that the dataset could be misused by individuals or groups to study hate speech for malicious purposes, such as refining their own hate speech tactics. For this, we need to implement strict access controls and ethical guidelines for dataset usage. Researchers and institutions should adhere to ethical research practices and ensure that the dataset is not used for harmful or malicious purposes. Additionally, an overreliance on automated hate speech detection models trained on this dataset might lead to content moderation systems that overly restrict free speech or disproportionately target specific groups. Thus, we should encourage human-in-the-loop approaches, where human reviewers work alongside automated systems to ensure more nuanced decision-making in content moderation. We should continuously fine-tune and evaluate the performance of detection models to reduce false positives and negatives.

A.5 Limitations

Despite the promising results and contributions, our study has several limitations. The dataset, primarily sourced from a few social media platforms, may introduce platform-specific biases and might not fully capture the nuances of sentiment and hate speech in general. Annotation subjectivity and moderate inter-annotator agreement could impact the reliability of labels. While MM-RAD outperforms baselines, its generalizability to other multimodal content or different code-switched languages is yet to be explored. Ethical considerations around data privacy and potential misuse are critical, necessitating stringent access controls. Additionally, significant computational resources required for training complex models like MM-RAD may limit accessibility for researchers with fewer resources.