CM_CLIP: Unveiling Code-Mixed Multimodal Learning with Cross-Lingual CLIP Adaptations

Gitanjali Kumari[1] **Arindam Chatterjee** [1,2] **Ashutosh Bajpai**[3]
Asif Ekbal[4] **Vinutha B N**[3]

[1]Department of Computer Science and Engineering,
[1]Indian Institute of Technology Patna, India, [2]Sahaj AI Software, Bangalore, India, [3]Wipro AI, Bangalore, India
[4]School of AI and Data Science, Indian Institute of Technology Jodhpur, India
[1]gitanjali_2021cs03@iitp.ac.in,[2]arindamc@sahaj.ai,
[3]{ashutosh.bajpai3,vinutha.narayanmurthy}@wipro.com,[4]asif@iitj.ac.in

Abstract

In this paper, we present **CMCLIP**, a **C**ode-**M**ixed **C**ontrastive **L**inked **I**mage **P**re-trained model, an innovative extension of the widely recognized CLIP model. Our work adapts the CLIP framework to the code-mixed environment through a novel cross-lingual teacher training methodology. Building on the strengths of CLIP, we introduce the first code-mixed pre-trained text-and-vision model, CM-CLIP, specifically designed for Hindi-English code-mixed multimodal language settings. The model is developed in two variants: CMCLIP-RB, based on ResNet, and CMCLIP-VX, based on ViT, both of which adapt the original CLIP model to suit code-mixed data. We also introduce a large, novel dataset called Parallel Hybrid Multimodal Code-mixed Hinglish (PHMCH), which forms the foundation for teacher training. The CMCLIP models are evaluated on various downstream tasks, including code-mixed Image-Text Retrieval (ITR) and classification tasks, such as humor and sarcasm detection, using a code-mixed meme dataset. Our experimental results demonstrate that CM-CLIP outperforms existing models, such as M3P and multilingual-CLIP, establishing state-of-the-art performance for code-mixed multimodal tasks. We would also like to assert that although our data, and frameworks are on Hindi-English code-mix, they can be extended to any other code-mixed language settings.

1 Introduction

In recent years, large pre-trained transformer-based language models (PLMs), such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), and GPT (Cohen and Gokaslan, 2020), have revolutionized the field of Natural Language Processing (NLP), achieving state-of-the-art performance across a wide range of tasks. This progress marks a significant shift towards a paradigm where general linguistic understanding is first learned from large-scale unsupervised data, followed by fine-tuning for task-specific applications (Min et al., 2023). These pre-training techniques, which derive representations directly from raw text, have fundamentally transformed the landscape of NLP, enabling models to perform exceptionally well on both traditional and emerging challenges in the field. Initially, monolingual pre-trained language models like BERT (Devlin et al., 2018), GPT-2 (Cohen and Gokaslan, 2020), and RoBERTa (Liu et al., 2019) laid the groundwork for subsequent expansions into multilingual models such as m-BERT (Devlin et al., 2018), Unicoder (Huang et al., 2019), and XLM/XLM-R (Lample and Conneau, 2019; Conneau et al., 2020).

In parallel, there has been a growing interest in pre-trained vision-and-language representations, as demonstrated by models like UNITER (Chen et al., 2020), ERNIE-ViL (Yu et al., 2020), Unicoder-VL (Li et al., 2019), Oscar (Li et al., 2020), VILLA (Gan et al., 2020), and M3P (Ni et al., 2020). These multimodal pre-trained models have made remarkable strides in aligning natural language and visual data, achieving state-of-the-art results in tasks such as image retrieval, text-to-image generation, image captioning, and visual quality assessment.

However, extending multimodal models to multilingual and code-mixed scenarios remains a challenge despite these advancements (Gupta et al., 2020; Maity et al., 2024). Cross-lingual alignment in such models often involves adapting the multimodal representations learned from English corpora to other languages. Recent studies have shown that multilingual textual representations often fail to learn equally high-quality representations across all languages, especially for low-resource languages and code-mixed languages (Hada et al., 2024; Watts et al., 2024). Code-mixed languages, in particular, have been significantly underrepresented in such models. While multimodal pre-trained models like CLIP (Radford et al., 2021) have demonstrated their ability to bridge vision and language tasks, their application in code-mixed settings has been limited. CLIP, trained using a contrastive learning objective, maximizes the similarity between corresponding image-text pairs. It measures the cosine similarity between image and text features extracted from their respective encoders, using this similarity to quantify the log odds of related pairs. While Multilingual-CLIP (Carlsson et al., 2022) attempts to address this issue by replacing the English text encoder with a pre-trained multilingual language model, such as m-BERT (Devlin et al., 2019), the handling of code-mixed languages still remains inadequate.

As social media platforms have gained widespread popularity, the prevalence of code-mixed language, particularly in informal communication, has surged (Sreelakshmi et al., 2020; Sengupta et al., 2022). Despite a large section of the global population using code-mixed languages, the development of pre-trained models for such data has been hindered by the scarcity of parallel data and the complexity of the language structures. Existing multimodal models are typically trained on high-resource languages, making them unsuitable for code-mixed languages, which involve more complex linguistic phenomena, including free word order and morphological diversity (Smith et al., 2017; Pacheco and Smith, 2015). Although some previous works have attempted to create synthetic datasets, they primarily rely on formal sources such as news articles, which rarely capture the code-mixing found in everyday speech and online communication.

To the best of our knowledge, no prior work has focused on creating a pre-trained vision-and-language model for code-mixed languages. This lack of attention can be attributed to several challenges, including the scarcity of annotated resources, the linguistic complexity of code-mixed languages, and the difficulty of aligning multimodal data in low-resource settings. Translating data from high-resource languages like English to code-mixed formats (e.g., Hindi-English) could mitigate the data scarcity problem. Leveraging transfer learning from state-of-the-art English models can enhance performance on various code-mixed tasks (Duh et al., 2011; Isbister et al., 2021). With the rapid advancements in Machine Translation (MT), there is now potential to bridge these gaps in low-resource, code-mixed language settings.

In this paper, we build upon the work of Carlsson et al. (2022), which demonstrated an effective way to incorporate new languages into pre-trained vision-and-language models such as CLIP. We extend this work to code-mixed languages, focusing on Hindi-English as a case study but laying the groundwork for broader application to other code-mixed languages.

We introduce CMCLIP, a vision-and-language model that explicitly aligns images with code-mixed text representations. Our model enforces the learning of universal representations, mapping both images and code-mixed text into a joint semantic space. The training process of our model is based on two key principles (c.f. Figure 1): (i) A translated code-mixed sentence should occupy the same vector space as the original English sentence, and (ii) A translated code-mixed sentence should occupy the same vector space as its corresponding image in cross-lingual scenarios.

In summary, the main contributions of this paper are as follows:

1. We present CMCLIP, a model designed to extend the CLIP model for better semantic representation of low-resource languages, specifically code-mixed languages such as Hindi-English.
2. Our approach consists of two phases: In the first, we create a large-scale parallel code-mixed Hindi-English corpus; in the second, we generate embeddings for multimodal code-mixed data by training the model using the teacher-student technique.
3. We demonstrate that CMCLIP outperforms state-of-the-art pre-trained models on two downstream tasks, namely image-text retrieval and multimodal classification.

2 Dataset

We create a parallel corpus of English and code-mixed [Hindi-English] sentences in order to train our model. Large part of this data also has parallel tagged images to exercise multimodal evaluation after training. For our downstream application evaluation, we also create another independent multimodal meme dataset.

2.1 Training Dataset

2.1.1 Publicly Available Corpus (PAC)

We used two publicly available parallel corpora in English and Hinglish (code-mixed Hindi-English).

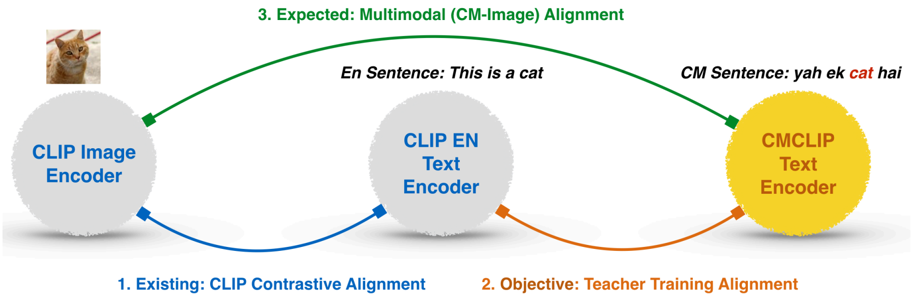

3. Expected: Multimodal (CM-Image) Alignment
En Sentence: This is a cat
CM Sentence: yah ek cat hai
CLIP Image Encoder
CLIP EN Text Encoder
CMCLIP Text Encoder
1. Existing: CLIP Contrastive Alignment
2. Objective: Teacher Training Alignment

Figure 1: Our proposed CMCLIP model building process via teacher training

(i). Srivastava et al. (Srivastava and Singh, 2020) has created one such dataset named PHINC, which was used in the scope of machine translation. This dataset consists of more than 13k tweets collected from Twitter and manually translated back into English. (ii). Another dataset is CMU Document Grounded Conversation (Zhou et al., 2018), which contains a set of conversations between users, with each conversation being grounded in a Wikipedia article about a particular movie. A subset of this dataset is translated into Hinglish (code-switched Hindi-English) and is available as the CMU Hinglish Document Grounded Conversations Dataset. It contains approximately 10K parallel sentences. In combination, both the publicly available datasets (PHINC and CMU Hinglish Document Grounded Conversations Dataset) have around 23K parallel English-Hinglish sentences and are biased towards specific domains. The scarcity of such publicly available parallel corpus led our work to extend beyond publicly available corpus to generate a large simulated machine-generated English-Hinglish parallel corpus.

2.1.2 SIMulated Corpus (SIM)

Significant progress has been made in code-mixed translation capabilities over the last few years. However, an efficient translation tool in the code-mixed domain is yet to be developed because of multiple known challenges discussed earlier (Refer to Section 1). Therefore, we reviewed multiple methods of generating code-mixed sentences from parallel language corpora. One such method, which is one of the simplest and fundamental, is presented in Appicharla et al. (2021). It requires a parallel corpus of primary and secondary languages in this case which is Hindi and English, respectively. A code-mixed sentence is generated from the alignment information between the sentences from two languages by replacement method. Fast Align (Dyer et al., 2013) is one of the most popular and well known tool available to generate alignment information between two sentences in different languages. It is also required to convert the Devanagari script (Hindi) into its Romanized form. For this purpose, we used ISO 15919 standard [1] (de Normalización, 2001) (c.f. Figure 2).

We used two publicly available datasets to generate parallel synthetic code-mixed sentences. One of those is multimodal, and other one is textual Hindi-English parallel dataset. First, (i). MSCOCO (Lin et al., 2014) is a popularly known dataset in multimodal domain. The subset of this dataset is also translated to Hindi language by google translate and publicly available[2]. This dataset contains more than half a million parallel captions in English and Hindi with image tags. (ii) Another dataset is the IITB Hindi-English parallel corpus (Kunchukuttan et al., 2018) manually created, which consists of around 1.6M parallel Hindi-English sentences.

2.1.3 Parallel Hybrid Multimodal Code-mixed Hinglish [PHMCH] Dataset Creation

This combined dataset (PAC and SIM) is henceforth referred to as Parallel Hybrid Multimodal Code-mixed Hinglish (PHMCH) dataset[3] (Refer to Table 1 for dataset statistics).

2.1.4 Dataset Statistics

Table 1 presents the detailed data statistics for both individual and combined PHMCH datasets. To characterize the code-mixed nature of the dataset, we utilize the Code-Mixing Index (CMI), a widely adopted metric for quantifying the extent of code-mixing in multilingual corpora (Gambäck and Das,

[1] https://www.iso.org/standard/28333.html

[2] https://github.com/nayeem8527/Chitra-VarNan

[3] We will make the dataset available after the acceptance of the paper.

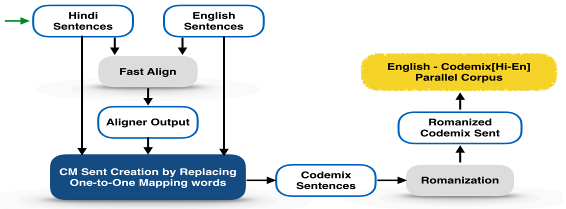


Figure 2: English & Code-mixed [Hi-En] Parallel Corpus Generation process

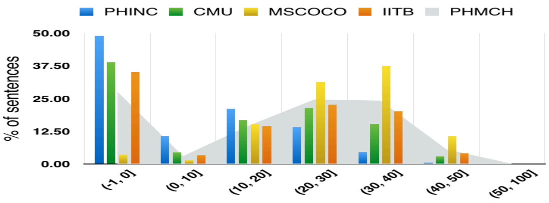


Figure 3: CMI bins for PHMCH dataset and its components

2016). The average CMI score for the PHMCH dataset is calculated to be 19.94, with the CMI for simulated data slightly exceeding that of publicly available corpora. The distribution of CMI across individual datasets, as well as the combined PHMCH dataset, is depicted in Figure 3. Additionally, Table 1 includes the average text length (in words) for each dataset, offering further insight into the linguistic characteristics of the data.

Source	Modality	# of Sentence	CM Type	Avg CMI	Avg Length (in words)
PHINC	Txt	13502	PAC	9.37	11.52
CMU-DoG	Txt	9934	PAC	15.20	11.69
IIT-B	Txt	1653930	SIM	17.26	13.97
MSCOCO	Txt-Img	526607	SIM	15.20	11.29
PHMCH	**Hybrid**	**2203973**	**Mixed**	**19.94**	**13.33**

Table 1: Data description and characteristics

2.2 Downstream Application Dataset

To evaluate the effectiveness of our proposed model on downstream tasks, we constructed a large-scale multimodal dataset specifically annotated for humor and sarcasm detection in code-mixed settings. Since no existing dataset addresses humor and sarcasm detection in code-mixed memes, we manually curated this dataset. Our data collection process was aligned with earlier works on meme analysis (Sharma et al., 2020; Kiela et al., 2020). We gathered memes from various domains, including politics, religion, terrorism, racism, sexism, and humor. After removing duplicates, the dataset comprised 5,647 unique memes. Each sample was annotated with two labels—humor (Yes/No) and sarcasm (Yes/No) —following the annotation guidelines provided by Kumari et al. (2024).

3 Methodology

Teacher learning (Cui et al., 2017; Reimers and Gurevych, 2020) is a method to train a student model under the supervision of a teacher model. In recent years, it has become a powerful process to distill knowledge from teacher model A to student model B with low resource data and with minimal infra requirement. CLIP has two different encoders for vision and text, which are aligned in the same vector space for a context in both modalities using contrastive learning. Another encoder for a different language could be aligned with the CLIP textual encoder using teacher learning. Alignment of the image encoder with the second language is expected if the alignment between two text encoders is achieved using teacher learning. In other words, we can say that for a given set of parallel (translated) sentences as well as its corresponding image $((s_1, t_1, i_1), ..., (s_n, t_n, i_n))$, where t_i is the code-mixed [Hi-En] translation of English sentence s_i and i_i is corresponding image-description of s_i and a teacher model T for the source language s, we employ mean squared loss to train a new student model M such that $M(t_i) \approx T(s_i)$ and $M(t_i) \approx T(i_i)$. This method is known as *knowledge transfer* since student M transmits teacher T knowledge (c.f. Figure 1). The loss function objective for cross-lingual teacher training is defined in Equation 1.

$$\mathcal{L} = \min \sum_{i=1}^{n} MSE(T(s_i) - M(t_i)) \qquad (1)$$

Where n is the total number of samples, s_i is an English sentence, t_i is a code-mixed translation of an English sentence s_i, $T(i)$ is a representation of i^{th} English sentence from the teacher model T and $M(i)$ is the representation of i^{th} code-mixed sentence from student model M.

When we train our teacher model to align the English and translated code-mixed [Hi-En] text encoders, we expect the same alignment between the translated code-mixed [Hi-En] text encoder and image encoder. We propose two code-mixed [Hi-En] CLIP encoders: (i). CMCLIP-RB (code-mixed CLIP with BERT and ResNet) and (ii). CMCLIP-VX (Code-mixed CLIP with XLM (Lample and Conneau, 2019) and Visual Transformer[ViT] (Dosovitskiy et al., 2020)). In the case of CMCLIP-RB, the RN50x4 variation of CLIP is used as a teacher model, and Multilingual BERT Uncased is the base student text encoder. Meanwhile, in the case of CMCLIP-VX, the Vit-L-14 variation of CLIP is used along with XLM-RoBERTa-Large as a student text encoder. We propose these two variations since ResNet is a base CLIP variation and ViT's (visual transformers) are advanced CLIP variations.

4 Experimental Details

4.1 Training Setup

We divide the PHMCH dataset into three parts for training (train), validation (Val), and evaluation(test) purposes. Our final training and validation(val_all) data have 20,40,418 and 1,13,557 sentences. The training dataset is a parallel dataset of English and code-mixed [Hi-En] sentences used for teacher training alignment between text encoders. val_all set is created with a split ratio of 0.05:0.3 from simulated and PAC datasets, respectively. We further subset the val[all] set into PAC and simulated category to track their individual and overall validation set performance. It contains 7,031 sentences of val[PAC] and 1,06,527 sentences of val[Sim], respectively. The test set is roughly 5% of an overall dataset containing 50K samples of triplets (parallel English and code-mixed sentence with images) largely from the MSCOCO subset to evaluate expected multimodal alignment. A maximum sentence length of 128 tokens is used throughout the training process. We have used mean square error[4](MSE) as a loss function for teacher training. Mean absolute error[5] (MAE) and cosine similarity[6] is used as accuracy measure metrices. It took 40 hours to train CMCLIP-RB for 170 epochs on three VX100-32GB GPUs and 120 hours to train CMCLIP-VX for 100 epochs on two VX100-32GB GPUs.

5 Results and Analysis

This section compares the performance of both proposed models on the training and validation dataset, analyses teacher training alignments, and discusses the evaluation results.

5.1 Training and Validation Results

This section explains the proposed models,*i.e.*, CMCLIP-RB and CMCLIP-VX behavior on training and validation datasets during model training.

5.1.1 Code-mixed CLIP ResNet and BERT Model (CMCLIP-RB)

Figure 4 shows training and validation results of teacher training for our proposed encoder, CMCLIP-RB. Multilingual BERT Uncased (Devlin et al., 2018) is used as the base text encoder (student) to align with the English text encoder of RN50x4 variations of CLIP (teacher). From Table 2, it can be observed that both val[PAC] and val[Sim], both MSE and MAE loss have decreased substantially, and cosine similarity increased as an outcome of teacher training. However, we also observed that val[PAC] loss is marginally higher than val[Sim] loss. A couple of inferences can be made out of this. First, the PAC dataset is way smaller than the simulated data. Therefore, PAC val loss is marginally higher. Second, the distribution and characteristics of generated simulated code-mixed data are very close to the original PAC data, leading to a very close loss of validation. Please refer to Section 5.4 for a detailed analysis of the impact of simulated data on the PAC dataset.

5.1.2 Code-mixed CLIP ViT and XLM Model (CMCLIP-VX)

Figure 5 shows training and validation results of teacher training for our second proposed encoder CMCLIP-VX. XLM-RoBERTa-Large (Conneau et al., 2020) is used as the base text encoder (student) to align with the English text encoder of ViT-L-14 variations of CLIP (teacher). The CMCLIP-RB model is consistent with the inferences that are taken from Table 2, which relates to the CMCLIP-VX model (as discussed in Section 5.1.1).

5.2 Teacher Training Alignment Analysis

It is expected that alignment between English and code-mixed [Hi-En] encoder shall be maximized after training. As a result, multimodal alignment between code-mixed [Hi-En] language and image encoder is also expected as the desired objective. We use a metric called margin to quantify the effectiveness of achieving the above desired objective, defined in Equation 2. Margin can be defined as the "relative difference between average dissimilar pair distance and the average similar pair distances, weighted by maximum magnitude to normalize the

[4] https://pytorch.org/docs/stable/generated/torch.nn.MSELoss.html

[5] https://pytorch.org/docs/stable/generated/torch.nn.L1Loss.html

[6] https://pytorch.org/docs/stable/generated/torch.nn.CosineSimilarity.html

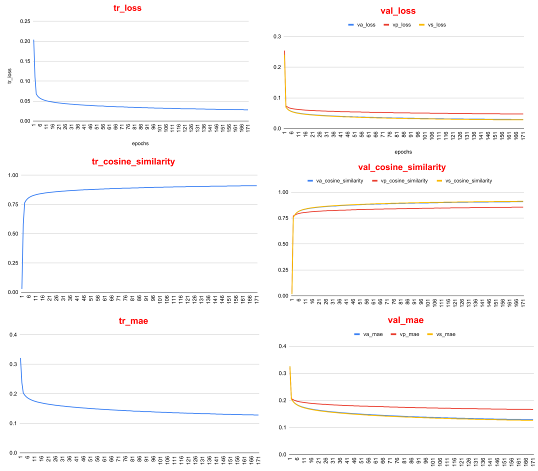


Figure 4: Training and validation results for CMCLIP-RB.

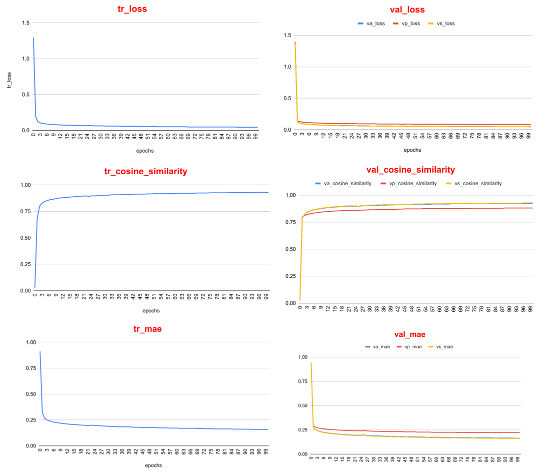


Figure 5: Training and validation results for CMCLIP-VX.

		At Start Epoch			At Saturation Epoch		
Model	**Data**	**MSE**	**MAE**	**Cosine Sim**	**MSE**	**MAE**	**Cosine Sim**
CMCLIP-RB	Train	0.204	0.321	0.028	0.028	0.128	0.911
	Val [All]	0.241	0.325	0.017	0.029	0.128	0.908
	Val [PAC]	0.254	0.319	0.035	0.047	0.166	0.855
	Val [Sim]	0.241	0.326	0.016	0.028	0.126	0.911
CMCLIP-VX	Train	1.29	0.913	0.025	0.043	0.158	0.931
	Val [All]	1.36	0.942	0.025	0.049	0.164	0.924
	Val [PAC]	1.39	0.936	0.027	0.084	0.22	0.881
	Val [Sim]	1.36	0.942	0.024	0.047	0.16	0.927

Table 2: Training and validation performance of proposed models

unit." In other words, the greater the margin better the model since it penalizes more on dissimilar pair distances by maximizing the distance and incentivizes similar pair distances by reducing distance among them. Therefore, the margin is expected to be similar after teacher training for text-to-text and text-to-image alignments.

$$margin = \frac{y - x}{max(\|x\|, \|y\|)}. \qquad (2)$$

where, x = Average similar pair distance and y =Average dissimilar pair distance.

5.2.1 Textual Alignment

In the Table 3, we have shown the Margin[Text-Text] (defined in Equation 2) between the baseline CLIP English encoder for English sentences and our two proposed CMCLIP-VX and CMCLIP-RB encoders for code-mixed [Hi-En] sentences. We use a 50K test dataset to generate a distance matrix in three different distance metrics. It is visible from the results that CMCLIP-VX, which is a combination of ViT and XLM is faring significantly better than CMCLIP-RB. CMCLIP-VX performance is statistically significant compared to CMCLIP-RB, with p-value < 0.05

5.2.2 Multimodal Alignment

This section discusses the results for expected multimodal alignment between the respective code-mixed encoder and image encoder for both our proposed models. For this, we compare the baseline CLIP multimodal alignment between the English encoder and image encoder with the alignment margin between the code-mixed [Hi-En] encoder and respective image encoder for both proposed models. We use 50K instances as the test set. We generate a distance matrix between the embedding of a sentence and its corresponding images to find all similar and dissimilar pair distances. For example, in the case of the baseline model, CLIP [RN50x4], the matrix is between embedding generated by an English encoder of CLIP for English text and embedding generated by an image encoder of CLIP for corresponding images for the ResNet version. The Margin[Image-Text] in Table 3 demonstrates that the multimodal alignment of the proposed encoders is very similar to the baseline CLIP encoders of the English language for respective versions. Therefore, it implies that the expected multimodal alignment between our proposed code-mixed [Hi-En] text and image encoder is achieved as closely as possible to the original multimodal alignment between CLIP-based English text and image encoder.

5.3 Performance on the Downstream Task

To show the robustness of our proposed model, we conduct experiments on two downstream tasks: (i) Image-Text Retrieval task and (ii) Multimodal classification task (humor and sarcasm detection in memes).

5.3.1 Performance on Image-Text Retrieval Task

The cross-modal image-text retrieval (ITR) task is to retrieve the relevant samples from one modality given the sample in another modality (Cao et al., 2022; Yuan et al., 2022), usually consisting of two sub-tasks: image-to-text (i2t) and text-to-image (t2i) retrieval. We compare mR(mean recall k =1,5,10) metrics for both i2t and t2i retrieval tasks. Reported numbers are averaged over three different random seeds on a 1K dataset randomly sampled from the 50K test instances. In Table 4, we reported mR for i2t and t2i retrieval tasks over English sentences and labeled images for the following models, mCLIP, M3P and two proposed CLIP variations. We observe that mCLIP outperforms M3P on English i2t and t2i retrieval tasks, whereas both the proposed models outperform CLIP and M3P by a substantive margin on both code-mixed i2t and t2i retrieval tasks. Moreover, in line with the previous findings, the CMCLIP-VX model outperforms CMCLIP-RB.

5.3.2 Multimodal Classification Task

This task aims to determine the correct label (*i.e.*, humor or sarcasm) of a given meme. We can define this task as follows: Given a meme sample S_i with image, text (V_i, T_i) pair where image V_i with the shape (224,224,3) in RGB pattern and meme-text $T_i = (t_{i1}, t_{i2}, ..., t_{ik})$ where t_{ik} is number of words in meme-text T_i, the task is to create classifier that predicts label $Y \subseteq \{sarcastic, non - sarcastic\}$ or $Y \subseteq \{humorous, non - humorous\}$ for S_i.

Results analysis using Automatic Metrics: The result analysis for humor and sarcasm detection tasks, as presented in Table 5, demonstrates the comparative performance between baseline classifiers (Refer to Appendix Section A for the detailed discussion of baseline models) and our proposed CMCLIP-based classifiers based on F1-score (F1) and accuracy (A). Among the models, CMCLIP-VX achieves the highest performance, with an F1-score of 69.36% and accuracy of 54.27% for humor detection, and 64.38% F1 and 58.24% accuracy for sarcasm detection. CMCLIP-RB also performs strongly, achieving 67.58% F1 for humor and 57.34% for sarcasm. In contrast, baseline models (Model 1, Model 2, and Model 3) perform significantly worse, particularly in humor detection, with F1-scores ranging from 19.96% to 23.37%. The pre-trained multimodal models MCLIP [ViT-L-14] and M3P show intermediate performance, with M3P scoring a notable 67.78% F1 for humor but falling short in sarcasm detection (52.54%). Overall, our proposed CMCLIP models, particularly CMCLIP-VX, significantly outperform other models, demonstrating their effectiveness in handling code-mixed multimodal data.

Results analysis using Detailed Discussion: To explain the feasibility of our proposed models, we perform a detailed qualitative and quantitative analysis of some samples from the test set. In Table 6, we show 3 examples with true labels of humor and sarcasm class. We show the results of our two proposed model setups by comparing their predicted and actual labels. We observe that our proposed model can properly understand code-mixed [Hi-En] and its alignment with corresponding images, resulting in the correct prediction of the associated class. Heatmaps of confusion matrices for both CMCLIP-RB and CMCLIP-VX setups are shown in Figure 6 in Appendix, which also shows the robustness of the proposed models.

5.3.3 Error Analysis

Despite the high performance mentioned for our CMCLIP models for both tasks, it still fails to anticipate the correct class in a few cases. Therefore, we thoroughly examine the reason for all the errors and categorize them into the following categories:

- *Long code-mixed sentences:* Our proposed model does not perform well for a few samples where sentence length is comparatively longer than average. The model is not able to generate a good multimodal representation for such examples (c.f. sample 1 in Table 7).
- *Lack of domain-related knowledge:* Since the training of our proposed pre-trained model is based on the idea of representing parallel code-mixed text and images in a common semantic space based on their semantic similarity. But in a few meme samples, text and image are not very strongly semantically aligned to

	Margin [Text-Text]		**Margin [Image-Text]**			
			ResNet based Model		ViT based Model	
Distance Metric	**CMCLIP-RB** (EN-CM)	**CMCLIP-VX** (EN-CM)	**CLIP [RN50x4]** (EN-IMG)	**CMCLIP-RB** (CM-IMG)	**CLIP [ViT-L-14]** (EN-IMG)	**CMCLIP-VX** (CM-IMG)
MSE	0.865	0.900	0.103	0.109	0.173	0.172
MAE	0.644	0.699	0.067	0.053	0.131	0.132
Cosine	0.866	0.902	0.196	0.201	0.190	0.189

Table 3: Margins [Textual and Multimodal]

Model	**t2i**	**i2t**	**Average**
English Baselines			
CLIP [RN50x4]	0.679	0.709	0.694
CLIP [ViT-L-14]	0.728	0.744	0.736
M3P	0.551	0.564	0.557
Code-mixed [Hi-En] comparison with multilingual code-mix/ code-switch baselines			
MCLIP [ViT-L-14]	0.624	0.649	0.636
M3P	0.380	0.398	0.389
CMCLIP-RB	**0.720**	**0.729**	**0.724**
CMCLIP-VX	**0.736**	**0.765**	**0.750**

Table 4: Code-mixed image-text retrieval results on test dataset.

Model	**Humor**		**Sarcasm**	
	F1	**A**	**F1**	**A**
Model 1	22.02	62.17	26.15	39.9
Model 2	23.37	61.47	32.09	55.39
Model 3	19.96	67.11	31.89	60.14
MCLIP [ViT-L-14]	61.55	39.61	54.98	58.69
M3P	67.78	43.28	52.54	52.59
CMCLIP-RB	**67.58**	**53.71**	**57.34**	**55.74**
CMCLIP-VX	**69.36**	**54.27**	**64.38**	**58.24**

Table 5: Results of our downstream tasks *i.e.* Humor and Sarcasm detection. Significance *t-test* p-values< 0.05

Input	Test meme 1	Test meme 2	Test meme 3
True Humor label	1	1	0
CMCLIP-RB	1	1	0
CMCLIP-VX	1	1	0
True sarcasm label	0	1	0
CMCLIP-RB	0	1	0
CMCLIP-VX	0	1	0

Table 6: Sample test examples with predicted humor and sarcasm labels for CMCLIP-RB and CMCLIP-VX models. Due to the space constraint, we placed the actual meme with translated text in the Table 8 in Appendix

	Test sample 1	Test sample 2	Test sample 3
True Humor label	1	1	1
CMCLIP-RB	0	1	0
CMCLIP-VX	1	0	0
True sarcasm label	1	1	0
CMCLIP-RB	0	0	0
CMCLIP-VX	0	1	0

Table 7: Examples of miss-classification by the proposed CMCLIP-RB and CMCLIP-VX models. Due to the space constraint, we placed the actual meme with translated text in the Table 9 in Appendix.

each other, unlike other multimodal tasks, *i.e.*, visual-common sense reasoning, image-text retrieval, image captioning, *etc.* The model needs strong contextual/domain knowledge to understand such memes. In such scenarios, our model makes an incorrect prediction (c.f. sample 2 and 3 in Table 7).

5.4 Ablation Study: Impact Analysis of Simulated Data

Simulated data enables substituting real datasets in domains where data scarcity is a known challenge (Yeomans et al., 2019) *i.e.*, code-mixing. It is also widely reported that large simulated data often helps to enhance the model's ability to learn new features, therefore leading to an improvement in performance and a reduction in variance. We demonstrate here that the extensive simulated data helps in enhancing the model performance compared to a model built only on a scarce real PAC dataset. We compare the performance of val[PAC] in two scenarios: (i) Training with full dataset, *i.e.*, PAC+Simulated, and (ii) Training with only the PAC dataset. This experiment was performed using CMCLIP-RB. From Table 10 in the Appendix, we can observe a significant reduction in val[PAC] numbers on all three distance metrics (in the case of training with simulated data). As a result, it can be inferred that simulated data helps the model learn new features and reduce variance further.

6 Conclusion

In this paper, we presented CMCLIP, a novel pre-trained model designed to learn joint visual-semantic embeddings through teacher-student training specifically for multimodal code-mixed languages, with experiments focusing on Hindi-English. We introduced a large-scale English-code-mixed parallel dataset, PHMCH, which played a critical role in training the model. Our approach generated more robust multimodal representations compared to baseline encoders for code-mixed data. Experimental results across various multilingual and code-mixed models demonstrated the superior performance of CMCLIP, especially in downstream tasks like image-text retrieval and multimodal classification. In the future, our framework could be extended to other code-mixed languages and additional vision-and-language tasks.

Limitations

In this paper, we discussed the vision-and-language model, which enforces the explicit alignments between images and language in a code-mixed setting and aims to learn universal representations to map images or code-mixed text into a joint semantic space. While this model includes a novel approach, which subsequently obtains state-of-the-art performance on our hybrid dataset, this work has some limitations. The proposed model could perform better for a few samples where sentence length is comparatively longer than average. We did not evaluate this model in real-world settings. This means we can not say how it will perform when an out-of-distribution dataset is used as input.

7 Ethics and Broader Impact

Individual Privacy To maintain the anonymity of any individual, we replaced the actual name with Person-XYZ throughout the paper. In addition, we also tried to anonymize the known faces presented in the visual part of the meme by masking them. We have masked these faces only to maintain the anonymity issues in the paper. During the implementation, we used the original image.

Biases Detecting and removing political and religious biases is an extensive research area. However, previous annotation studies show that we cannot correctly remove bias and subjectivity from the annotation process despite having some form of annotation scheme. However, any biases detected in our meme dataset are unintentional, and we have no intention of harming any individual or group. We ensure that our data collection is generated equally and comparably in order to answer any political and religious bias queries. Furthermore, we ensure that the topic includes various issues relevant in the Indian context over the last seven years by using a keyword-based data-gathering technique. Moreover, we made sure that the terms included were inclusive of all the conceivable politicians, political organizations, young politicians, extreme groups, and religions and were not prejudiced against any one group. Based on previous work done by (Davidson et al., 2019) to remove biases from the dataset during annotation, in our dataset, annotators were strictly instructed not to make decisions based on what they believe but on what the social media user wants to transmit through that meme.

Misuse Potential We suggest that researchers be aware that our meme dataset might be abused to filter the memes based on prejudices that may or may not be connected to demographics or other textual information. To prevent this from happening, human intervention with moderation would be essential.

Intended Use Our dataset is presented to encourage research into studying code-mixed Hindi-English representation. We believe that it represents a valuable resource when used appropriately.

Acknowledgements

The research reported in this paper is an outcome of the project “**HELIOS: Hate, Hyperpartisan, and Hyperpluralism Elicitation and Observer System,**" sponsored by Wipro AI Labs, India.

References

Ramakrishna Appicharla, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2021. Iitp-mt at calcs2021: English to hinglish neural machine translation using unsupervised synthetic code-mixed parallel corpus. In *CALCS*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Min Cao, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang. 2022. Image-text retrieval: A survey on recent research and development.

Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. 2022. Cross-lingual and multilingual clip. In *Proceedings of the Language Resources and Evaluation Conference*, pages 6848–6854, Marseille, France. European Language Resources Association.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *ECCV*.

Vanya Cohen and Aaron Gokaslan. 2020. Opengpt-2: Open language models and implications of generated text. *XRDS*, 27(1):26–30.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jia Cui, Brian Kingsbury, Bhuvana Ramabhadran, George Saon, Tom Sercu, Kartik Audhkhasi, Abhinav Sethy, Markus Nussbaum-Thom, and Andrew Rosenberg. 2017. Knowledge distillation across ensembles of multilingual models for low-resource languages. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4825–4829.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Organización Internacional de Normalización. 2001. *ISO 15919: Information and Documentation : Transliteration of Devanagari and Related Indic Scripts Into Latin Characters*. ISO.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale.

Kevin Duh, Akinori Fujino, and Masaaki Nagata. 2011. Is machine translation ripe for cross-lingual sentiment classification? In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, page 429–433, USA. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *NAACL*.

Björn Gambäck and Amitava Das. 2016. Comparing the level of code-switching in corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1850–1855, Portorož, Slovenia. European Language Resources Association (ELRA).

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. *CoRR*, abs/2006.06195.

Deepak Gupta, Pabitra Lenka, Asif Ekbal, and Pushpak Bhattacharyya. 2020. A unified framework for multilingual and code-mixed visual question answering. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 900–913, Suzhou, China. Association for Computational Linguistics.

Rishav Hada, Varun Gumma, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024. Metal: Towards multilingual meta-evaluation.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. *CoRR*, abs/1909.00964.

Tim Isbister, Fredrik Carlsson, and Magnus Sahlgren. 2021. Should we stop training more monolingual models, and simply use machine translation instead? *CoRR*, abs/2104.10441.

Satyajit Kamble and Aditya Joshi. 2018. Hate speech detection from code-mixed hindi-english tweets using deep learning models. *ArXiv*, abs/1811.05145.

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. *CoRR*, abs/2005.04790.

Gitanjali Kumari, Chandranath Adak, and Asif Ekbal. 2024. Mu2sts: A multitask multimodal sarcasm-humor-differential teacher-student model for sarcastic meme detection. In *Advances in Information Retrieval*, pages 19–37, Cham. Springer Nature Switzerland.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. 2019. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *CoRR*, abs/1908.06066.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. *CoRR*, abs/2004.06165.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. Microsoft coco: Common objects in context.

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. *CoRR*, abs/1907.11692.

Krishanu Maity, A.s. Poornash, Sriparna Saha, and Pushpak Bhattacharyya. 2024. ToxVidLM: A multimodal framework for toxicity detection in code-mixed videos. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11130–11142, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Jianfeng Gao, Dongdong Zhang, and Nan Duan. 2020. M3P: learning universal representations via multitask multilingual multimodal pre-training. *CoRR*, abs/2006.02635.

Mark B. Pacheco and Blaine E. Smith. 2015. Across languages, modes, and identities: Bilingual adolescents' multimodal codemeshing in the literacy classroom. *Bilingual Research Journal*, 38(3):292–312.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation.

T. Y.S.S. Santosh and K. V.S. Aravind. 2019. Hate speech detection in hindi-english code-mixed social media text. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, CoDS-COMAD '19, page 310–313, New York, NY, USA. Association for Computing Machinery.

Ayan Sengupta, Sourabh Kumar Bhattacharjee, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022. Does aggression lead to hate? detecting and reasoning offensive traits in hinglish code-mixed texts. *Neurocomputing*, 488:598–617.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. Semeval-2020 task 8: Memotion analysis - the visuolingual metaphor! *CoRR*, abs/2008.03781.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Blaine E. Smith, Mark B. Pacheco, and Carolina Rossato de Almeida. 2017. Multimodal codemeshing: Bilingual adolescents' processes composing across modes and languages. *Journal of Second Language Writing*, 36:6–22.

K Sreelakshmi, B Premjith, and K.P. Soman. 2020. Detection of hate speech text in hindi-english code-mixed data. *Procedia Computer Science*, 171:737–744. Third International Conference on Computing and Network Communications (CoCoNet'19).

Vivek Srivastava and Mayank Singh. 2020. PHINC: A parallel Hinglish social media code-mixed corpus for machine translation. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 41–49, Online. Association for Computational Linguistics.

Ishaan Watts, Varun Gumma, Aditya Yadavalli, Vivek Seshadri, Manohar Swaminathan, and Sunayana Sitaram. 2024. Pariksha : A large-scale investigation of human-llm evaluator agreement on multilingual and multi-cultural data.

Jordan Yeomans, Simon Thwaites, William S. P. Robertson, David Booth, Brian Ng, and Dominic Thewlis. 2019. Simulating time-series data for improved deep neural network performance. *IEEE Access*, 7:131248–131255.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *CoRR*, abs/2006.16934.

Zhiqiang Yuan, Wenkai Zhang, Changyuan Tian, Xuee Rong, Zhengyuan Zhang, Hongqi Wang, Kun Fu, and Xian Sun. 2022. Remote sensing cross-modal text-image retrieval based on global and local information. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations.

A Baseline Models

Baseline models: For the evaluation, we have created these baseline classifiers:

- Model 1: For this baseline, we use FastText (Bojanowski et al., 2017) word embedding for the text representation. We obtained the region-specific features for the visual feature using a pre-trained VGG19 (Simonyan and Zisserman, 2015) framework. These feature vectors are concatenated through one softmax layer for the final prediction.
- Model 2: For this baseline, we followed the approach mentioned in (Kamble and Joshi, 2018) to represent the text. For the visual part, we used the same architecture as mentioned in model 1.
- Model 3: We followed a prevalent approach to deal with code-mixed sentences, *i.e.*, character-level encoding (Sengupta et al., 2022; Santosh and Aravind, 2019). For the visual part, we use the same architecture as mentioned in model 1.
- MCLIP [ViT-L-14]: In this model, we used mCLIP for learning textual and visual representations of a given meme. We concatenate those features and use a softmax layer at the end for classification. Only pre-trained weights are used in this stage.
- M3P: This is an m3p-based baseline model. After getting the joint representation with the help of the M3P model, we forwarded this representation to one softmax layer for the final classification.

Proposed models: We used our proposed CMCLIP models as feature extractors in this setup.

- CMCLIP-RB: This is the first proposed model explained in Section 5.1.1.

- CMCLIP-VX: This is the second proposed model explained in Section 5.1.2.

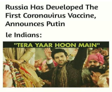
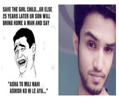
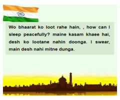
Input image	 Russia Has Developed The First Coronavirus Vaccine, Announces Putin le Indians: "TERA YAAR HOON MAIN"	 SAVE THE GIRL CHILD...OR ELSE 25 YEARS LATER UR SON WILL BRING HOME A MAN AND SAY "ASHA TO MILI NAHI ASHISH KO HI LE AYA..."	 Wo bhaarat ko loot rahe hain, , how can I sleep peacefully? maine kasam khaee hai, desh ko lootane nahin doonga. I swear, main desh nahi mitne dunga.
English Translation	Russia has developed the first coronavirus vaccine, Announces Putin. Le Indians: I am your friend	Save the girls child.. or else 25 years later your son will bring home a man and say : I didnot get Asha, so I married Ashish"	They were looting India. How do I sleep peacefully I have sworn I will not let the country be looted. I have taken an oath, I will not let the country perish.
Name	test meme 1	test meme 2	test meme 3

Table 8: Sample test examples for CMCLIP-RB and CMCLIP-VX models mentioned in Table 6

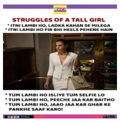
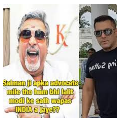
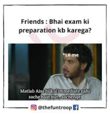
Input image	 STRUGGLES OF A TALL GIRL * ITNI LAMBI HO, LADKA KAHAN SE MILEGA * ITNI LAMBI HO FIR BHI HEELS PEHENE HAIN * TUM LAMBI HO ISLIYE TUM SELFIE LO * TUM LAMBI HO, PEECHE JAA KAR BAITHO * TUM LAMBI HO, JAAO JAA KAR GHAR KE PANKHE SAAF KARO!	 Salman ji apka advocate mile tho hum bhi lalit modi ke sath wapas INDIA a jaye??	 Friends : Bhai exam ki preparation kb karega? *Le me Matlab Aisa bilkul immediate nahi soche hain but.. sochenge @thefuntroop
English Translation	Struggles of a Tall Girl,you are so tall, How did you find a boy. So tall still wearing heels. You are tall that's why you take selfie . You're tall, go back and sit you are tall, go home and clean the fans	Salman ji, if I get your lawyer, can we also come to India with Lalit Modi?	Brother, when will he prepare for the exam? Means they have not thought immediately but will think
Name	Test sample 1	Test sample 2	Test sample 3

Table 9: Examples of miss-classification by the proposed CMCLIP-RB and CMCLIP-VX models mentioned in Table 7.

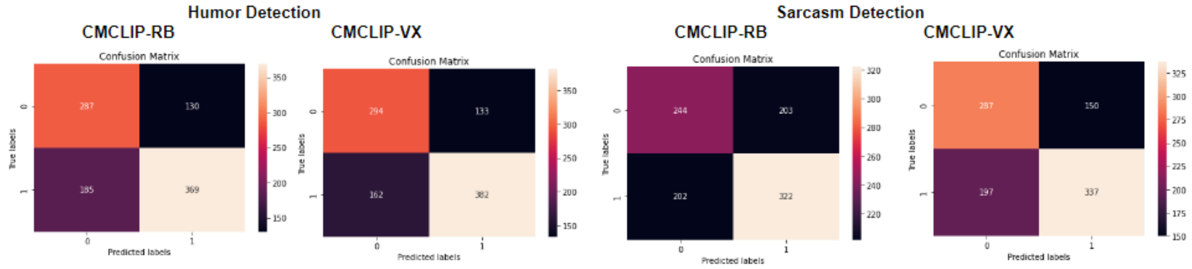


Figure 6: Heatmaps of the confusion matrix for humor and sarcasm detection task for both proposed model setups.

	Full Dataset [PAC + Sim]		Only PAC Dataset	
Distance Metric	**Train**	**Val [PAC]**	**Train**	**Val [PAC]**
MSE	0.028	0.047	0.042	0.053
MAE	0.128	0.166	0.160	0.175
Cosine	0.089	0.145	0.133	0.163

Table 10: Impact of using simulated data on performance of real data