Penta ML at EXIST 2024: Tagging Sexism in Online Multimodal Content With Attention-enhanced Modal Context

Notebook for the EXIST Lab at CLEF 2024

Deeparghya Dutta Barua[1,*,†], Md Sakib Ul Rahman Sourove[1,*,†], Fabiha Haider[1], Fariha Tanjim Shifat[1], Md Farhan Ishmam[1,3], Md Fahim[1,2,*] and Farhad Alam Bhuiyan[1,*]

[1]Research and Development, Penta Global Limited, Bangladesh
[2]CCDS Lab, IUB, Bangladesh
[3]Islamic University of Technology, Bangladesh

Abstract

Content moderation at scale warrants automated systems that are capable of understanding nuance from the text and images being posted online. Transformer-based models have been shown to perform with these preconditions in mind, but the additional complexities originating from multimodality and multilinguality mandate the need for better tuned systems that can capture more enriched representations of the context. This can essentially translate to downstream tasks such as sexism identification in online content, which is the forefront of the EXIST 2024 shared tasks. This paper, as part of the EXIST challenge at CLEF 2024, investigates an attention-based approach to improve performance over baseline multimodal models by assigning separate importance to the textual and visual representations. The proposal is evaluated against CLIP and ViLT, two established multimodal models, while achieving state-of-the-art performance in multi-label classification tasks in the hard-hard evaluation context. The study is further augmented by the inclusion of different forms of ablations, involving confusion metrics for the applicable tasks.

Keywords

Hateful Memes Detection, Multimodal Fusion, Vision Language Modeling

1. Introduction

The proliferation of internet users throughout the globe has caused an upsurge in the amount of content being generated and consumed on a daily basis. This huge volume of content includes a wide range of information, personal opinions and entertainment, which reflect the diverse perspective of its global audience. Naturally, given the volume, a large portion of the content on the internet enforces harmful and problematic behaviors [1]. One of the most concerning issues among them is sexism.

Memes have recently entered the cultural zeitgeist where a piece of visual or audio-visual media may be accompanied by a humorous piece of text [2]. However, as with any form of media, memes too can be weaponized to diffuse problematic views and ideologies. Consequently, a good portion of memes on the internet enforce discriminatory behavior, such as sexism. The influence of such content is multifaceted, ranging from creating further polarization and radicalization of impressionable people, to causing psychological discomfort to its victims.

In order to combat such problematic trends and create a safer online experience regardless of gender, it has become quite important for automated systems to screen and flag for potentially inimical content at scale. The simple rule-based approaches may work to a certain extent for filtering textual content, but these approaches tend to fail when multiple modalities, such as images along with text, are involved.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
*Corresponding authors.
†These authors contributed equally.
deeparghyadutta-2018425323@cs.du.ac.bd (D. D. Barua); souroveskb@gmail.com (M. S. U. R. Sourove); fabihahaider4@gmail.com (F. Haider); fariha.tanjim.shifat@gmail.com (F. T. Shifat); farhanishmam@iut-dhaka.edu (M. F. Ishmam); fahimcse381@gmail.com (M. Fahim); pdcsedu@gmail.com (F. A. Bhuiyan)

This is due to the fact that nuance and context cannot be inferred easily from a set of rules. To take this into account, transformer-based architectures have recently gained prevalence in hateful content detection due to their ability to better understand context.
The role of a comprehensive dataset is quintessential in order to tackle the challenge of training these transformer-based models for the purpose of employing them to downstream tasks. The Hateful Memes Challenge dataset [3] is a popular pick, containing over 10,000 multimodal samples of binary-labeled data. However, datasets that cover hateful behaviour at a broad scale fail to capture the specifics of sexism found in online content. Moreover, while not a low-resource language, there is a lack of Spanish datasets for sexism detection that are adequately labeled. The SemEval-2019 dataset [4] does address sexism in English and Spanish textual content specifically, but the binary labels lack resolution when it comes to understanding intent, and it does not address sexism found in images. Similarly, the Automatic Misogyny Identification (AMI) text dataset [5] offers labeled data for sexist text classification with some granularity, but the classifications are exclusive, with each sample being assigned to a single category only. In order to address multiple dimensions at once, the EXIST 2024 [6, 7] dataset is the only multimodal dataset for both Spanish and English content, where the labels are assigned with varying levels of granularity, ranging from binary classification, to multi-class and even multi-label classification where multiple labels can co-exist for the same sample.
The multimodal nature of the EXIST 2024 memes dataset imposes a different set of challenges as opposed to its text-only counterpart. The multilinguality also adds to the complexity since most pretrained models tend to be trained on datasets that are predominantly in English. To address all these issues, we have proposed an attention-enhanced approach that uses both the textual and the visual context to get a more enriched representation of the sample and then use it for the downstream task of sexism classification, the architecture of which is discussed in finer detail within this paper. We have also evaluated our approach empirically over existing multimodal models, such as CLIP and ViLT, along with the error analysis that delineates the shortcomings of our system. As per the experiments, our approach yields superior performance in multi-class and multi-label classification problems when considering the hard-hard evaluation context (ICM-Hard), improving the performance by 7%-12% and 6%-13% respectively.

2. Background

2.1. Dataset and Tasks

The EXIST 2024 memes dataset includes a total of 5,044 image-text pairs in English in Spanish. The training split contains 4,044 memes, each labeled by 6 different annotators for their respective tasks. The test split provides 1,053 unlabeled samples. Since the memes dataset does not provide a separate split for validation, an 80-20 split has been performed on the original training split, resulting in 3,235 samples for training and 809 samples for validation. A detailed language-wise breakdown of the dataset splits can be visualized in Table 1.
The dataset also provides additional features pertaining to the annotators, such as their genders, ages, ethnicities, education level, and countries.

The tasks for the memes dataset seek to classify sexism in memes to varying degrees. A more detailed overview of the hard label distribution of each of these tasks can be found in Table 2. The soft labels for each of the tasks assign a numeric value to each category in the Learning With Disagreement (LeWiDi) format. The soft values for tasks 4 and 5 add up to 1, as the labels are unique for each sample. For task 6, they do not necessarily add up to 1, as it is a multi-label problem.

- **Task 4:** Given an image and its caption, the subtask is to classify whether meme in question contains any references to sexism or not. This includes any form of sexist content, describing situations involving discrimination towards women, and even contexts where sexism is criticised. It is essentially a binary classification problem.

Table 1
EXIST dataset distribution by the train-validation-test splits. The train-validation split is 80%-20% of the initial training samples since the memes dataset does not provide a separate validation set.

Split	Language	Samples
Train	English	1588
	Español	1647
	All	**3235**
Validation	English	422
	Español	387
	All	**809**
Test	English	513
	Español	540
	All	**1053**
	All	**5044**

Table 2
Class/Label distribution in the dataset based on hard labels. For task 4, the labels are binary. Task 5 is uses multi-class labeling where each sample belongs to a single unique class. For task 6, multiple labels can coexist for the same sample. Both tasks 5 and 6 are hierarchical, meaning that the labels other than "NO" are only assignable when the meme in question is sexist.

Task	Class/Labels	Train	Validation
Task 4	YES	1604	434
	NO	1111	271
Task 5	DIRECT	1078	277
	JUDGEMENTAL	351	109
	NO	1111	271
Task 6	IDEOLOGICAL-INEQUALITY	672	171
	STEREOTYPING-DOMINANCE	807	216
	OBJECTIFICATION	736	208
	SEXUAL-VIOLENCE	372	111
	MISOGYNY-NON-SEXUAL-VIOLENCE	307	90
	NO	1111	271

- **Task 5:** This subtask is a multi-class classification problem where the intention behind the meme needs to be identified. Given that a meme is sexist, this can either be classified as "DIRECT" or "JUDGEMENTAL". If it is not sexist to begin with, then the identification should also indicate that.
  - **DIRECT:** The meme itself enforces sexist ideology without any ironic or satirical aspect to it.
  - **JUDGEMENTAL:** The meme condemns sexist behavior, either directly or satirically.
- **Task 6:** The image can be classified into multiple labels — "IDEOLOGICAL-INEQUALITY", "STEREOTYPING-DOMINANCE", "OBJECTIFICATION", "SEXUAL-VIOLENCE" or "MISOGYNY-NON-SEXUAL-VIOLENCE". Multiple labels can coexist for a single image at the same time and similar to the previous tasks, there should be a separate label if the meme is not sexist.
  - **IDEOLOGICAL-INEQUALITY:** Ideological discrediting refers to all memes that discredit the feminist movement with the intention of devaluing, belittling and defaming the plight of women. On the other hand, "Inequality" refers to the memes that establish a narrative that no gender discrimination exists in the current society, or the flipped narrative where men are presented as the victims.

- **STEREOTYPING-DOMINANCE:** Memes that impose the idea of specific roles being better suited for women fall under stereotyping. Dominance is characterized by the positioning of men above women in various standings.
- **OBJECTIFICATION:** Memes dehumanizing or treating women as objects or commodities count under this label. These may also include the exertion of beauty or societal standards.
- **SEXUAL-VIOLENCE:** These memes in question call for sexual suggestions, sexual favors or sexual abuse.
- **MISOGYNY-NON-SEXUAL-VIOLENCE:** Expressions of physical violence and hatred towards women fall under this label.

3. Model Architecture

In this section, we will provide a brief overview of the methodology employed to address the tasks at hand. Specifically, we were provided with memes sourced from the internet, structured in image-caption pairs denoted as (V, T). For solving the tasks, the model architecture we have used has five different components i) Pretrained Vision-Language Model, ii) Semantics from Pooled Representations, iii) Attention Enhanced Context Vector for each Modality, iv) Modality Fusion, and v) Classification Head

3.1. Pretrained Vision-Language Model

Our experiment uses a pre-trained ViLT model. Each input meme sample $M = (V, T)$, comprising the image content V and its caption T, is processed individually. The text processor tokenizes the caption into its constituent tokens $T = t_1, t_2, \ldots, t_n$. Meanwhile, the image processor divides the input image $V \in \mathbb{R}^{C\times H\times W}$ into patches, which are then flattened to $v \in \mathbb{R}^{N\times(P^2\cdot C)}$, where (P, P) represents the patch resolution, and $N = \frac{HW}{P^2}$. Specifically, ViLT employs the BERTTokenizer as the text processor and the ViT (Vision Transformer) processor as the image processor.

The tokens and the image patches are then fed into the pre-trained ViLT model to enhance the comprehensive understanding of image-text pairs through multimodal fusion. From the model, we get image-aware text representations $H_T = \{h_{t_1}, h_{t_2}, \ldots, h_{t_n}\}$ and text-aware image representations $H_V = \{h_{p_1}, h_{p_2}, \ldots, h_{p_n}\}$ where h_{t_i} is the last layer hidden representation for $i-$th token and h_{p_i} is last layer hidden representation for $i-$th image-patch

3.2. Semantics from Pooled Representations

The ViLT model also gives a pooled representation of the whole multimodal input. ViLT incorporates a special [CLS] token at the beginning of the multimodal input. We also extract the last layer representation of this token. A linear projection is applied to that representation to extract the pooled representation

$$h_{pool} = W_{pool} \cdot h_{[\text{CLS}]} + b_{pool}$$

3.3. Attention Enhanced Context Vector for each Modality

Each token and patch holds its unique representation, with some being more crucial for prediction than others. To effectively combine these representations based on their significance, we utilize an additional attention network within each modality, ultimately determining a context vector.

- **Text Context Vector:** The text representations $H_T = \{h_{t_1}, h_{t_2}, \ldots, h_{t_n}\}$ obtained from ViLT are passed into an additional attention layer to compute learnable attention scores α_{t_i} for each token t_i in H. The calculation is as follows:

$$\alpha_{t_i} = \textit{softmax}(W_T \cdot h_{t_i} + b_T),$$

$$i = 1, 2, \ldots, n$$

After finding attention scores for each token, we find the context vector for the text modality by multiplying the h_{t_i} of token t_i with its attention score α_{t_i}.

$$c_T = \sum_{i=1}^{n} \alpha_{t_i} \cdot h_{t_i}$$

- **Vision Context Vector:**
  Similarly, the vision representations $H_V = \{h_{p_1}, h_{p_2}, \ldots, h_{p_n}\}$ are also fed into another additional attention layer to get the attention weights:

$$\alpha_{p_i} = \textit{softmax}(W_V \cdot h_{p_i} + b_V), \quad i = 1, 2, \ldots, n$$

Having calculated the attention scores for each patch, we determine the context vector for the vision modality by multiplying the representation h_{p_i} of each patch p_i with its corresponding attention score α_{p_i}.

$$c_V = \sum_{i=1}^{n} \alpha_{p_i} \cdot h_{p_i}$$

3.4. Modality Fusion

The context vectors c_T and c_V are summed to the pooled representation h_{pool} to get more enhanced representations $c_T^{'}$ and $c_V{}'$ where $c_T{}' = h_{pool} + c_T{}'$ and $c_V{}' = h_{pool} + c_V{}'$. Finally, we fuse both vision and text modality concatenating $c_T^{'}$ and $c_V{}'$ and pass them into a Multi Layer Perceptron (MLP) to get the modality fused feature.

$$c = \text{concat}[c_T, c_V] \tag{1}$$
$$z = \text{MLP}(c) \tag{2}$$

3.5. Classification Head

After finding the modality fused feature representation z, it is fed into a classification layer. The representation is the logits z is employed for the classification process by the following:

$$z^{'} = W \cdot z + b \tag{3}$$

Finally, we calculate the Cross-Entropy (CE) loss based on $z^{'}$ with the ground truth.

4. Experimental Setup

4.1. Settings

All experiments were conducted using Python (version 3.12) and PyTorch, leveraging the free NVIDIA Tesla P100 GPU provided by Kaggle. For the pretrained vision language models, we use `HuggingFace` transformers library (version 4.40.1). The models used for tasks 4 and 5 have been trained for 15 epochs, and the models for task 6 have been trained for 50 epochs from their pretrained checkpoints. We have

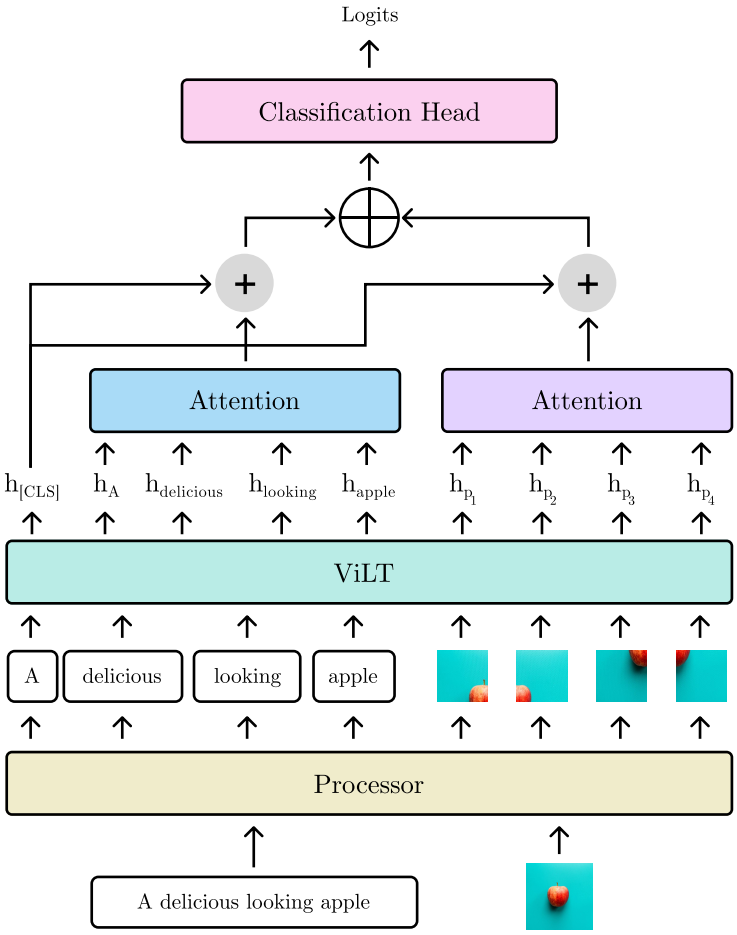


Figure 1: Overview of the proposed architecture. The text is tokenized and the images are divided into fixed-size patches using the processor. These are then passed through a ViLT model to get the embeddings, along with extra [CLS] tokens that capture the context. This contextual representation is further enhanced by passing and weighting the text and image embeddings through separate attention networks, and then adding them to the CLS tokens. Finally, this enhanced representation is used for the downstream classification task.

used the schedule-free *AdamW* optimizer [8] which requires no explicit hyperparameters for the optimization stopping step T due to not using scheduling altogether. The initial learning rate used is 2×10^{-4}. The epsilon value of the optimizer is 10^{-6}, and the β coefficients for computing the running average of the gradient and its square are 0.9 and 0.999 respectively. The seed or random state value used across all operations is 42. The batch size used in all the instances is 64.

For the classification head, the dimension of the hidden layer is 256 and the dropout layer sets 10% of the input values to zero. To do better modeling and include different explainability results, we adopt different training and explainable experiment settings from EDAL, ITPT, HateXplain [9, 10, 11] papers.

4.2. Evaluation Metrics

Our evaluation metrics for validation follow the same scoring system selected for the EXIST test rankings. "ICM-Hard" is the primary metric in the hard-hard evaluation context and "ICM-Soft" serves the same purpose in the soft-soft evaluation context. Additionally, the "Macro F1" values have also been listed as a more traditional metric that does not consider hierarchical classifications. The PyEvALL library [12] has been used to calculate the scores throughout all the experiments.

- **ICM-Hard:** The Information Contrast Model-Hard (ICM-Hard) metric [13] is designed to evaluate unbalanced hierarchical multi-label classification problems by incorporating hierarchical

relationships and category specificity. ICM-Hard operates by considering the information content (IC) of categories and their intersections. The IC of a category represents the probability of items to appear in the category or any of its descendants, providing a measure of category specificity. Given two feature sets A and B, the metric is defined as:

$$ICM(A,B) = \alpha_1 IC(A) + \alpha_2 IC(B) - \beta IC(A \cup B)$$
$$IC(A) = -log(P(A))$$

The values selected for α_1, α_2 and β for evaluating these experiments are 2, 2 and 3 respectively.

- **ICM-Soft:** The Information Contrast Model-Soft (ICM-Soft) metric is an extension of ICM-Hard, designed to handle hierarchical multi-label classification problems in a learning with disagreement (LeWiDi) scenario. The metric is defined similarly using information content but it accommodates for soft ground truth assignments and soft system outputs. Given a category c with an agreement v to a given item, the IC is defined as —

$$IC(\{\langle c, v\rangle\}) = -\log_2(P(\{d \in D : g_c(d) \geq v\}))$$

A recursive function is applied to calculate the ICM-Soft over a set of assignments:

$$IC\left(\bigcup_{i=1}^{n}\{\langle c_i, v_i\rangle\}\right) = IC(\langle c_1, v_1\rangle) + IC\left(\bigcup_{i=2}^{n}\{\langle c_i, v_i\rangle\}\right) - IC\left(\bigcup_{i=2}^{n}\{\langle lca(c_1, c_i), \min(v_1, v_i)\rangle\}\right)$$
$$IC(\{\langle c, v\rangle\}) = -\log_2(P(\{d \in D : g_c(d) \geq v\}))$$

Where $lca(a, b)$ is defined as the lowest of common ancestor of categories a and b.

- **Macro F1:** The F1 score is the harmonic mean of precision and recall for a class. The macro F1 metric is the average of the F1 scores of each class. All classes are given equal weight in this metric, which is desirable for underrepresented classes in highly imbalanced datasets.

$$\text{Macro-F1} = \frac{1}{N}\sum_{i=1}^{N} F_i$$
$$F_i = \frac{2 \cdot P_i \cdot R_i}{P_i + R_i}$$

5. Experimental Evaluation

For our experiments, we have chosen CLIP [14] and ViLT [15] as our baseline models due to their multimodal capabilities in image-text pairs. For standard classification using CLIP, we concatenated the representations obtained from both vision and language inputs and passed them into an MLP for classification. For ViLT, we extracted the representations of the CLS token and fed them into an MLP for standard classification. As ViLT outperformed CLIP in most scenarios, we selected ViLT as our final model for experiments with the proposed architecture.

Table 3
Model performances on the validation dataset. Our approach is shown to have a superior macro F1 score in tasks 4 and 6 over the baseline models CLIP and ViLT.

		ICM-Hard	ICM-Soft	Macro F1
Task 4	CLIP	-0.2196	-1.9589	0.5592
	ViLT	**-0.1414**	**-1.5918**	0.5755
	Our Approach	-0.1553	-2.0167	**0.5815**
Task 5	CLIP	-1.1371	**-7.0817**	0.3526
	ViLT	**-1.0804**	-7.7936	**0.3822**
	Our Approach	-1.1196	-7.2194	0.3344
Task 6	CLIP	-	-14.8292	0.2812
	ViLT	-	**-13.9726**	0.2723
	Our Approach	-	-17.2587	**0.2921**

Table 4
Task 4 performance on the test dataset. The ViLT model without any context-enhancing consistently outperforms other models in all splits across all metrics.

		ICM-Hard	ICM-Soft	Macro F1
All	CLIP	-0.1745	-1.5664	0.6524
	ViLT	**-0.1308**	**-1.2910**	**0.6742**
	Our Approach	-0.2049	-1.7425	0.6101
Español	CLIP	-0.2666	-1.9498	0.6354
	ViLT	**-0.2201**	**-1.7547**	**0.6657**
	Our Approach	-0.3016	-2.2800	0.5957
English	CLIP	-0.0824	-1.2165	0.6713
	ViLT	**-0.0420**	**-0.8683**	**0.6837**
	Our Approach	-0.1083	-1.2548	0.6260

5.1. Results of the models on Validation Dataset

Table 3 presents the performance of the models on the validation dataset. For Task 4, the CLIP model achieves a macro F1 score of 55.92%, with ICM-Hard and ICM-Soft scores of -0.2196 and -1.9589, respectively. When using the ViLT pretrained vision-language model instead of CLIP, we observe an improvement of approximately 2% in the macro F1 score. The ICM-Hard and ICM-Soft scores also improve with ViLT. The performance of the ViLT model is further enhanced when incorporated with our model design, achieving a 1% gain and reaching a macro F1 score of 58.55%.

For Task 5, ViLT again outperforms CLIP, showing a 3% improvement in the macro F1 score. However, when integrating our approach with ViLT, the performance decreases. ViLT with the [CLS] token-based classification head performs best on the validation data for this task. For Task 6, CLIP performs better than ViLT, with CLIP achieving a macro F1 score of 28% compared to ViLT's 27%. Our approach surpasses these baselines, improving by 1% over CLIP and 2% over ViLT. Nevertheless, ViLT yields the best ICM-Soft score for Task 6. The ICM-Hard scores for Task 6 could not be calculated as there is no public-facing implementation for multi-label scenarios.

5.2. Results on Test Dataset of Task 4

Table 4 displays the performance comparison of three models — CLIP, ViLT, and our approach on the task 4 test dataset. Overall, ViLT outperforms both CLIP and our approach, exhibiting the lowest error rates and achieving the highest macro F1 score of 0.6742. CLIP performs moderately well, with a macro F1 score of 0.6524, while our approach shows the least favorable performance with a score of 0.6101. This trend is consistent across subsets, with ViLT consistently outperforming the other models.

Table 5
Task 5 performance on the test dataset. Our approach outperforms other models in all splits in the hard-hard evaluation metric, as indicated by the ICM-Hard scores.

		ICM-Hard	ICM-Soft	Macro F1
All	CLIP	-0.6546	-5.3096	**0.3856**
	ViLT	-0.7089	-5.9832	0.3841
	Our Approach	**-0.6123**	**-5.2668**	0.3841
Español	CLIP	-0.7533	-5.7513	0.3694
	ViLT	-0.7131	-5.7680	**0.3733**
	Our Approach	**-0.6128**	**-5.4847**	0.3636
English	CLIP	-0.5554	**-4.9616**	**0.4019**
	ViLT	-0.7043	-6.2139	0.3835
	Our Approach	**-0.6144**	-5.1208	0.3949

Particularly, in the English subset, ViLT achieves the highest macro F1 score of 68.37%, followed by CLIP at 67.13%, emphasizing ViLT's effectiveness in this context. However, in the Spanish subset, although ViLT still leads with a score of 66.57%, the performance gap between ViLT and CLIP is narrower. Unfortunately, our proposal demonstrates the highest error rates across all subsets, suggesting potential areas for improvement.

5.3. Results on Test Dataset of Task 5

Table 5 illustrates the performance of the models for Task 5 test dataset. In this task, CLIP achieves a macro F1 score of 38.56%, with values of -0.6546 for ICM-Hard and -5.3096 for ICM-Soft instances. ViLT performs marginally worse than CLIP with a macro F1 score of 38.41%, having -0.7089 for ICM-Hard and the value of -5.9832 for ICM-Soft. Our approach shows the best performance in both the ICM metrics with the scores being -0.6123 and -5.2668 respectively. The macro F1 performance of 38.41% is comparable to ViLT.

On the Spanish tweet subset, CLIP performs similarly to its overall performance, with a macro F1 score of 36.94%, while ViLT slightly outperforms CLIP with a score of 37.33% and the lowest error rate for ICM-Soft instances at -5.7680. Our approach achieves the most promising results on the Spanish subset, with the highest values for both ICM-Hard and ICM-Soft instances being -0.6144 and -5.4547 respectively. For the English subset, CLIP performs slightly better with a macro F1 score of 40.19%, whereas ViLT shows a slight decrease with a score of 38.35%. Our approach remains competitive with the highest score for ICM-Hard instances and a macro F1 score of 39.49%.

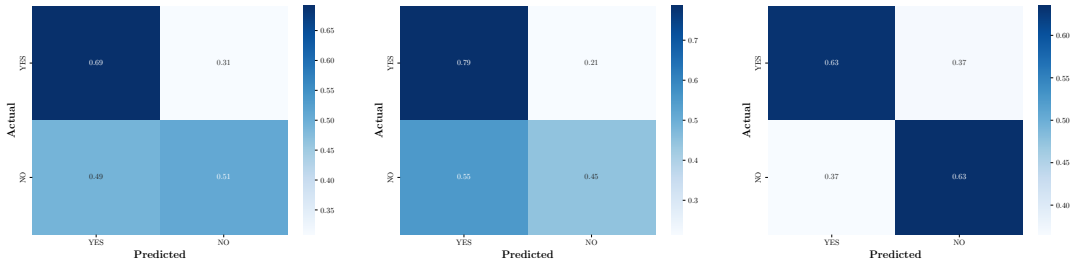
Key observations indicate that performance varies across different subsets and models, with CLIP and ViLT performing consistently and our approach obtaining the highest scores in the ICM-Hard metric and largely having better performance in the ICM-Soft metric as well, while maintaining competitive macro F1 scores.

5.4. Results on Test Dataset of Task 6

In Table 6, we reported the performance of models on Task 6 in test dataset. For the overall dataset, our approach outperforms both CLIP and ViLT, achieving the best ICM-Hard score of -1.3631 and the highest macro F1 score of 33.56%, though it shows a slightly lower ICM-Soft score of -13.2556 compared to ViLT's best score of -11.2593. In the Spanish subset, Our approach again demonstrates superior performance with the highest macro F1 score of 31.92% and the best ICM-Hard score of -1.6610, although it has a worse ICM-Soft score of -15.5081 compared to CLIP's -12.4430. For the English

Table 6
Task 6 performance on the test dataset. Our approach outperforms other models in all splits in the hard-hard evaluation metric, as indicated by the ICM-Hard scores. It also achieves the best performance in the macro F1 metric.

		ICM-Hard	ICM-Soft	Macro F1
All	CLIP	-1.5499	-11.8047	0.3053
	ViLT	-1.4684	**-11.2593**	0.3093
	Our Approach	**-1.3631**	-13.2556	**0.3356**
Español	CLIP	-1.7780	**-12.4430**	0.2873
	ViLT	-1.7534	-12.8432	0.2987
	Our Approach	**-1.6610**	-15.5081	**0.3192**
English	CLIP	-1.3484	-11.7789	0.3207
	ViLT	-1.2112	**-10.3388**	0.3170
	Our Approach	**-1.1027**	-11.9271	**0.3487**



(a) CLIP (b) ViLT (c) Our Approach

Figure 2: Confusion matrices on the predictions using CLIP, ViLT and our approach respectively for task 4. Our approach offers the best estimation for true positives and true negatives.

subset, our approach achieves the highest overall performance, leading with a macro F1 score of 34.87% and the highest ICM-Hard score of -1.1027, while ViLT shows the best ICM-Soft score of -10.3388.

Across all subsets, our approach excels consistently in macro F1 and ICM-Hard metrics, indicating its robustness in handling hard instances for multi-label tasks, while ViLT often provides the best scores for ICM-Soft instances, demonstrating its effectiveness in handling soft instances. Overall, this table highlights the strengths of each model in different aspects of Task 6, with our approach showing the most balanced and highest performance in key metrics across various subsets.

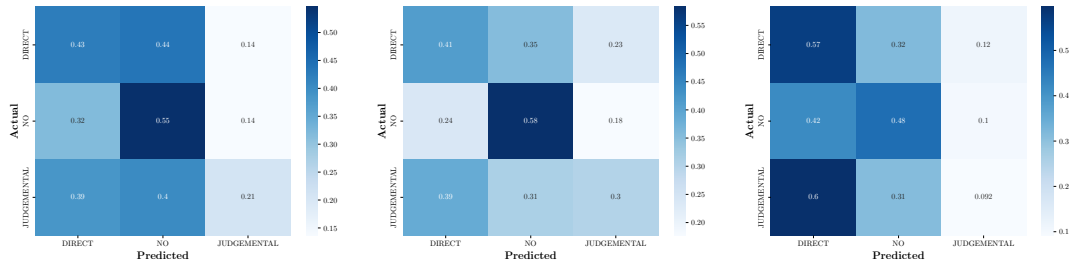
6. Error Analysis

6.1. Confusion Metrics

6.1.1. Confusion Metrics Analysis for Task 4

From the confusion matrices of task 4 in figure 2 we can see that the CLIP model excels in identifying the 300 true positive cases, which is 69% of the total. But it has high Type-I (49% false positives) and Type-II (31% false negatives) error, resulting in significant misclassifications.

ViLT performs better than CLIP yielding 341 true positive cases, which accounts for 79% of the total. But it has a higher Type-I (55% false positives) and moderate Type-II (21% false negatives) error, resulting lower accuracy to detect the true negative cases which only account for 45% of the total. Our approach, on the other hand, performs consistently on both the true positives and the true negatives, with both covering 63% of the total. To be more specific, the true negative cases have been handled



(a) CLIP (b) ViLT (c) Our Approach

Figure 3: Confusion matrices on the predictions using CLIP, ViLT and our approach respectively for task 5. While CLIP and ViLT offer better estimates for the true negatives ("NO"), our proposal handles the true "DIRECT" values better than the other two.

better than both CLIP and ViLT. It also has consistent rates for Type-I and Type-II errors, both clocking at 37%.

6.1.2. Confusion Metrics Analysis for Task 5

Using the confusion matrices for task 5 from figure 3, we observe that the CLIP model performs well on the 'NO' class with 148 correct predictions but suffers from high misclassification rates for 'DIRECT' and 'JUDGEMENTAL' classes. The ViLT model shows a similar trend. It performs better than CLIP on both the 'NO' and 'DIRECT' classes with 58% and 30% correct predictions respectively. But it still shows a tendency towards wrongly considering other classes to be 'JUDGEMENTAL', despite it having very low representation in the dataset. Our approach, on the other hand, performs really well in predicting the true 'DIRECT' samples with the highest amount of 57%. It has comparable performance for the 'NO' class, but struggles heavily with 'JUDGEMENTAL' samples, showing signs of being affected by the dataset distribution.

7. Conclusion

In this study, we addressed the pressing issue of sexism in online content, focusing on the complexity posed by multimodal and multilingual (English and Spanish) data. Our proposed attention-enhanced approach effectively integrates textual and visual contexts, providing enriched representations for sexism classification. By leveraging the EXIST 2024 dataset, which offers a thoroughly outlined look into the variants of sexism present in online content, we demonstrated improvements over existing models such as CLIP and ViLT. Our empirical evaluations and subsequent error analysis highlighted both the strengths and areas for improvement in our system. The findings underscore the importance of comprehensive datasets and advanced model architectures in tackling online discrimination. Future work will explore further enhancements in multimodal learning and expand the approach to other forms of harmful content beyond sexism, contributing to a safer and more inclusive online environment.

Acknowledgements

This project has been sponsored by Penta Global Limited, Bangladesh. We would like to express our deepest gratitude to Penta Global for for their financial support.

References

[1] J. B. Walther, Social media and online hate, Current Opinion in Psychology 45 (2022) 101298.

[2] B. Kostadinovska-Stojchevska, E. Shalevska, Internet memes and their socio-linguistic features, English Language and Linguistics 2 (2018). doi:10.5281/zenodo.1460989.

[3] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The hateful memes challenge: Detecting hate speech in multimodal memes, 2021. arXiv:2005.04790.

[4] H. R. Kirk, W. Yin, B. Vidgen, P. Röttger, SemEval-2023 Task 10: Explainable Detection of Online Sexism, in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, 2023. URL: http://arxiv.org/abs/2303.04222. doi:10.48550/arXiv.2303.04222.

[5] E. Fersini, D. Nozza, P. Rosso, Overview of the evalita 2018 task on automatic misogyny identification (ami), in: EVALITA@CLiC-it, 2018.

[6] L. Plaza, J. Carrillo-de-Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of EXIST 2024 – Learning with Disagreement for Sexism Identification and Characterization in Social Networks and Memes, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), 2024.

[7] L. Plaza, J. Carrillo-de-Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of EXIST 2024 – Learning with Disagreement for Sexism Identification and Characterization in Social Networks and Memes (Extended Overview), in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[8] A. Defazio, Xingyu, Yang, H. Mehta, K. Mishchenko, A. Khaled, A. Cutkosky, The road less scheduled, 2024. arXiv:2405.15682.

[9] M. Fahim, M. S. Shahriar, M. R. Amin, Hatexplain space model: Fusing robustness with explainability in hate speech analysis (2023).

[10] M. Fahim, Aambela at blp-2023 task 2: Enhancing banglabert performance for bangla sentiment analysis task with in task pretraining and adversarial weight perturbation, in: Proceedings of the First Workshop on Bangla Language Processing (BLP-2023), 2023, pp. 317–323.

[11] M. Fahim, A. A. Ali, M. A. Amin, A. M. Rahman, Edal: Entropy based dynamic attention loss for hatespeech classification, in: Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation, 2023, pp. 775–785.

[12] UNEDLENAR, Pyevall, https://github.com/UNEDLENAR/PyEvALL, 2024.

[13] E. Amigo, A. Delgado, Evaluating extreme hierarchical multi-label classification, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 5809–5819. URL: https://aclanthology.org/2022.acl-long.399. doi:10.18653/v1/2022.acl-long.399.

[14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[15] W. Kim, B. Son, I. Kim, Vilt: Vision-and-language transformer without convolution or region supervision, in: International conference on machine learning, PMLR, 2021, pp. 5583–5594.