DIMEMEX-2024: Detection of Inappropriate Memes from Mexico

Mario García-Hidalgo, Mario García-Rodríguez, Jorge Payno, María Fernanda Salerno, and Isabel Segura-Bedmar

Computer Science Department
Universidad Carlos III de Madrid, 28911 Leganés, Madrid, Spain

Abstract. The DIMEMEX 2024 competition proposes the development of multimodal computational models to detect abusive memes in Mexican Spanish, focusing on hate speech, offensive language, and vulgar content. This paper presents our approach to the two subtasks defined by the competition: a three-way classification distinguishing hate speech, inappropriate content, and neutral content, and a finer-grained classification that categorizes hate speech into specific types such as classism, sexism, and racism. Our methodology uses dataset expansion techniques, enriching the dataset by sourcing new memes and employing data augmentation methods to tackle class imbalances and increase the overall volume of data. We gathered memes from diverse sources, with a focus on underrepresented classes, resulting in a more balanced dataset. To further enhance the dataset, we leveraged state-of-the-art multimodal models such as Google Gemini 1.5 Pro for text extraction and Meta's LLAMA 3 for text augmentation. This augmentation strategy increased the dataset size, providing a more robust training set. For the categorization of the memes, initially we used the BETO model for text representation and Vision Transformers (ViTs) for image features. We then experimented with multimodal models, such as CLIP, Multi-CLIP, and SIGLIP, to map features into a common feature space, fusing them and performing the classification with a MLP.

Keywords: Natural Language Processing, Multimodal Analysis, Memes, Spanish, Inappropriate Content, Hate Speech

1 Introduction

The multimodal classification of online content, which involve the integration of textual and visual information, represents a growing area of research in the field of Natural Language Processing (NLP) [1] [2]. Until recently, efforts dedicated to such multimodal tasks were sparse. However, advancements in computational power and the development of transformer-based models have stimulated significant interest and progress in this domain.

According to a comprehensive survey by [3] the evolution of text classification techniques, particularly with the advent of Large Language Models (LLMs), has

expanded the scope from unimodal (text-only) inputs to more complex multimodal applications. This survey highlights the importance of transformer-based models in capturing complex contextual relationships and semantic nuances, facilitating the processing of text data that has historically been challenging and expensive to analyze.

In this context, the DIMEMEX 2024 competition focuses on the detection of abusive memes in Mexican Spanish [4]. This competition aims to advance research on identifying hate speech, offensive language, and vulgar content within memes; a challenging problem due to their multimodal nature. Memes typically combine text and image to convey humor or irony, and the removal of either component can significantly alter the intended message. Addressing this challenge, DIMEMEX 2024 includes two key subtasks: (a) a three-way classification to distinguish between hate speech, inappropriate content, and neutral content, and (b) a finer-grained classification to categorize hate speech into specific types such as classism, sexism, and racism.

Although this is the first time that DIMEMEX is being held, there is precedent in the form of the DA-VINCIS competition at IberLEF 2023 [5]. The DA-VINCIS 2023 shared task aimed to develop automatic solutions for detecting violent events in social networks, involving both binary classification to identify violent incidents and multi-label multi-class classification to categorize types of violence. This competition, which used a multimodal corpus of tweets and associated images, achieved competitive results and highlighted the potential of multimodal approaches for content detection tasks.

These initiatives are very beneficial and significant because of the growing role of social networks in communication and information dissemination. The ability to accurately detect and mitigate abusive content has important social and economic implications. Effective multimodal detection systems can help foster safer online environments, reduce the spread of harmful messages, and protect vulnerable populations from abuse and discrimination. By participating in DIMEMEX 2024, we contribute to this effort and explore the potential of what multimodal analysis can achieve in the context of abusive content detection.

In our approach, we have extended the dataset by retrieving new memes and applying data augmentation techniques to address the issue of class imbalance and increase the overall data volume. Initially, new memes were gathered from diverse sources, including Image Downloader, Telegram channels dedicated to dark humor, and Reddit, focusing on underrepresented classes like "inappropriate content" and "hate speech." This process resulted in a more balanced dataset. The text from each meme was extracted using the Google Gemini 1.5 Pro multimodal model, which effectively labeled the memes based on the content.

Additionally, to further enhance the dataset, we employed Meta's LLAMA 3 large language model for text augmentation, generating nine variants of each meme's text. This increased the dataset size by tenfold, providing a robust set of data for training. For image augmentation, various transformations such as random rotation, affine transformation, and perspective distortion were applied to introduce diversity. These augmentations not only expanded the dataset but also

improved the model's generalization ability, reducing the risk of overfitting and enhancing performance on unseen data. These efforts collectively contributed to a more effective and comprehensive multimodal hate speech detection system.

2 State of the art

2.1 Multimodal classification to detect hate speech in memes

Multimodal classification involves the integration and analysis of data from multiple modalities, such as text and images. This approach uses the complementary information present in different types of data to improve the performance and robustness of classification models.

The detection of hate speech in memes presents an application of multimodal classification due to the combination of visual and textual elements that convey meaning. This task requires models to effectively interpret both modalities to accurately identify and classify harmful content.

Kumar and Nandakumar [6] developed the Hate-CLIPper architecture for multimodal hateful meme classification, which explicitly models cross-modal interactions between image and text representations using Contrastive Language-Image Pre-training (CLIP) encoders. Their approach employs a feature interaction matrix (FIM) to capture the correlations between image and text features, achieving state-of-the-art performance on the Hateful Memes Challenge (HMC) dataset with an AUROC of 85.8, surpassing human performance. The Hate-CLIPper architecture effectively combines multimodal pretraining with intermediate fusion, demonstrating its generalizability across different meme datasets and highlighting the importance of modeling cross-modal interactions for robust classification

Similarly, Sabat et al. [7] explore a multimodal approach to hate speech detection involving vision and language (text), specifically in the context of memes. They gathered meme data from various sources to create a hate memes dataset, which was used to train and evaluate statistical models based on state-of-the-art neural networks. Their approach included fine-tuning pretrained descriptors for the specific task. The implementation showcased the integration of BERT and VGG-16 features for robust performance.

3 Methodology

This section contains the information related to the methodology and the approaches used during the experimentation phase.

3.1 Dataset

The dataset used was provided by the competition organizers. This dataset includes both the image of each meme and the text extracted from the image using an OCR tool. The organizers supplied three data sets:

– **Training:** 2,263 memes.
– **Validation:** 323 memes.
– **Test:** 649 memes.

The first subtask has three classes: hate speech, inappropriate content, and none. The second subtask, being a fine-grained task for hate speech, divides this category into four more classes, therefore we have: classicism, racism, sexism, other (hate speech), inappropriate content, and none.

In the training data, besides the image and text of the meme, the corresponding labels for both tasks are included. However, for the validation and test data, labels were not provided as validation was conducted through the competition's website without disclosing the exact labels of these sets. Therefore, the initial work focused solely on the training dataset.

The first step was to better understand this dataset, which involved studying the class distribution for both tasks:

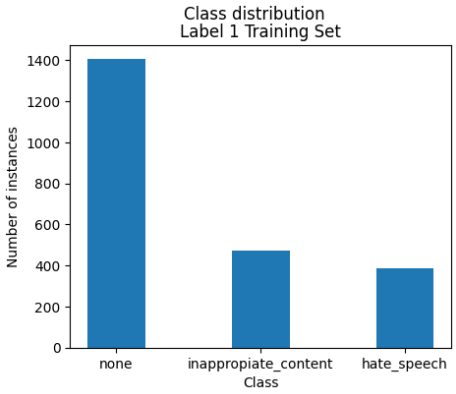

Fig. 1. Class distribution for task 1

Upon analyzing this dataset, two main problems were identified:

– **Insufficient Data:** 2,263 examples are not enough to train a robust classifier [8].
– **Class Imbalance:** The class distribution is disproportionate. In the first task, there are many more examples in the "none" class than in the "inappropriate_content" and "hate_speech" classes. In the second task, this imbalance is even more pronounced, as the "hate_speech" class is divided into four subclasses, making the difference more noticeable. This can cause the classifier to favor the more numerous classes, which is problematic for classification.

For these reasons, we have made two solution: extended the dataset by retrieving new memes and using data augmentation.

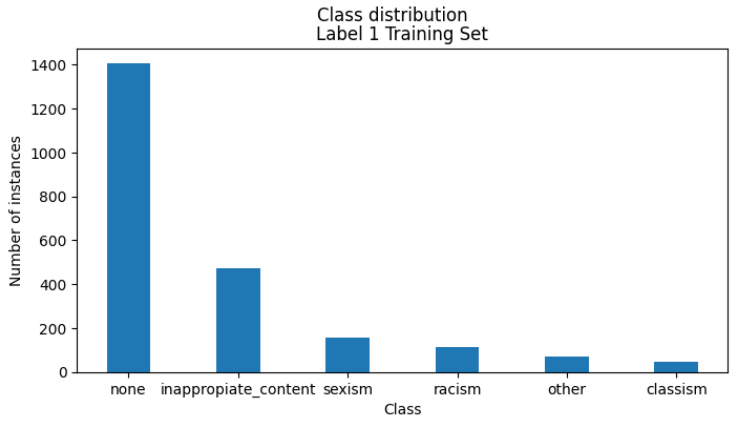

Fig. 2. Class distribution for task 2

To balance and extend the dataset, new data was added to the training dataset. The first step involved searching for new memes, focusing on classes with fewer examples, such as "inappropriate_content" and "hate_speech" (and their subclasses for the second task), to match the majority class "none". Three sources were used for searching new memes:

- **Image Downloader [9]:** This tool, developed in Python, allows crawling and downloading images using search engines like Google, Bing, and Baidu. Keywords such as "dark humor meme", "offensive meme", "racist meme", and "inappropriate meme" were used to search for memes.
- **Telegram:** Observing that the results from the Image Downloader were not optimal, memes were searched in Telegram channels dedicated to dark humor. These memes were of much better quality, and most of the selected ones came from this source.
- **Reddit:** Although the previous two sources helped balance the classes, some subclasses in the second task, such as "classism", still had few examples. Therefore, Reddit was used to obtain more memes.

Once new data was collected, they were labeled and the text from each meme was extracted. The Google Gemini 1.5 Pro multimodal model was used for this purpose [10]. This model is one of Google's most advanced and multimodal models, and although there are superior models and other alternatives like GPT-4, using this model with the API is free in a limited way (2 requests per minute). The API was used with a specific prompt to extract the text from the image and reasonably label the two proposed tasks for each meme. Then, we used Gemini for text extraction, instead of another OCR model, labeling the memes on this way.

The dataset now consists of 3,641 memes, with a much more homogeneous and balanced class distribution:

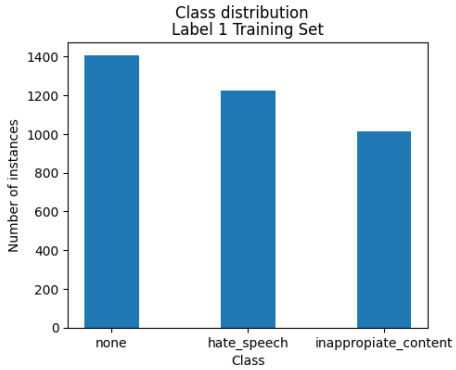

Fig. 3. Class distribution for task 1 with new data

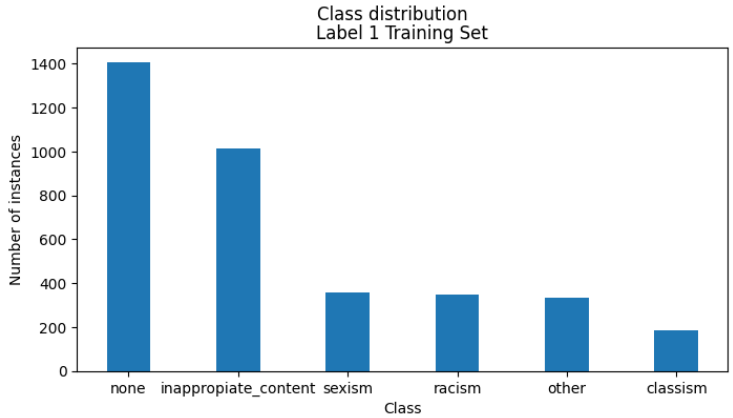

Fig. 4. Class distribution for task 2 with new data

For the second task, although logically it is more difficult to balance due to the division of "hate_speech" into the other 4 categories, it is seen that these categories that comprise "hate_speech" have been balanced.

Even with a more balanced dataset, the overall amount of data was still limited, so data augmentation was performed on both texts and images. This not only increases the dataset size but also improves the model's generalization to new and unseen data by exposing it to diverse modified versions of the images and texts, reducing the risk of overfitting [11].

For text data augmentation in the training dataset, Meta's LLAMA 3 large language model (LLM) [12] was used. This model is open-source, which facilitated its use. Several data augmentation techniques were tested, such as synonym

replacement (library NLPAug[13]) and back translation (library textaugment [14]), but none matched the quality obtained with LLAMA 3. Therefore, a specific prompt was used to generate 9 variants of each meme's text The approach was to use this model to generate sentences with similar meanings. The prompt consisted of a text from a meme taken from the training set, and the model was asked to generate 9 similar texts. Each variant or similar text is then linked to the same image associated with the original input text. This increasing is used in two ways in the experimentation. The first way by using the usual training size and iterate from the 10 different text inputs on each epoch which was called "**Image-like augmentation**". And second way by increasing the dataset size by 10 times, which will create an instance of each meme and the same photo for each text data augmentation. With this second way, we have a total of 30,641 instances. We have named this second way "**Multi-instance augmentation**".

For images, the training set size was not increased directly; instead, various random transformations were applied to each image to increase data diversity. Specifically, the following transformations were applied sequentially to each image:

- **Random Rotation:** Rotates the image randomly up to a maximum of 40 degrees in any direction.
- **Random Affine Transformation:** Applies a random affine transformation, allowing translation (up to 40% in both directions) and scaling (between 0.7x and 1.3x of the original size).
- **Random Perspective:** Applies a random perspective transformation with a distortion scale of 0.5.

So the two ways of data augmentation used the transformations on images and different ways to select the text associated to them.

3.2 Data preprocessing

To ensure an optimal evaluation process of the developed models, the provided labeled dataset was randomly divided into training, validation, and test splits. The splits were performed in a stratified way, maintaining the original proportion of the instances on each of the splits. From the original data, 70 % of the instances were used for training, 15% for validation and the remaining 15% for test.

Before providing the input texts and images of the memes to the models, a series of preprocessing operations were applied. Regarding the text, the function provided by the organizers on their baseline models was used. This function converts the text to lowercase, normalizing usernames and URLs, separating special characters, and reducing multiple spaces to a single one. For the text tokenization, the tokenizers were inherited from the architectures used.

With respect to the image inputs, the image processors were inherited as well from the architectures, these modules are in charge of performing the necessary operations on the images such as resizing or normalization.

Additionally, several techniques of data augmentation were applied to the images and texts for artificially expanding the training split, as explained in Section 3.1.

3.3 Arquitecture

Having specified the preprocessing operations, the next step is to establish the different architectures used during the experimentation phase. It is important to remark that the main objective was to use multimodal architectures for a better understanding of the meme's image and text combinations.

The architectures used can be divided into three different modules: a feature extractor module, a cross-modal fusion module, and a Multi-layer Perceptron (MLP) classifier. Figure 5 provides an visual representation of the general architecture used.

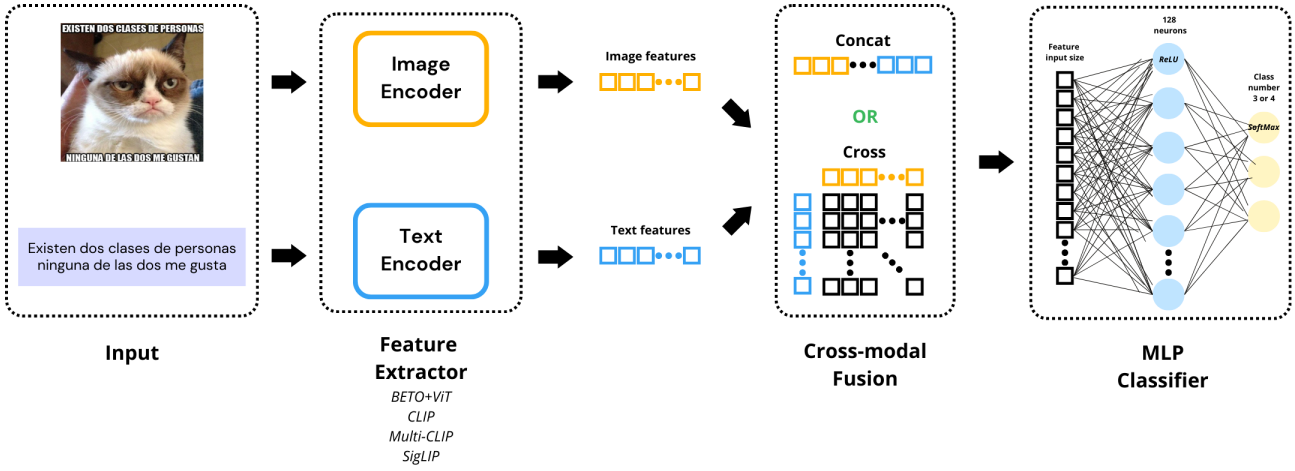

Fig. 5. General architecture used

3.3.1 Feature Extractor

This module will be in charge of receiving the preprocessed image and text inputs and encoding them, obtaining their corresponding embeddings.

For this task, we started using a baseline approach similar to the one provided by the competition organizers. It consisted of using two independent models for the feature extraction process.

- To represent texts, we used the BETO[15] model, which is based on BERT[16] (Bidirectional Encoder Representations from Transformers) and trained on a big Spanish corpus.
- Features from the images were extracted using Vision Transformers (ViTs)[17], one of the techniques used for computer vision nowadays. The ViT model processes images by dividing them into fixed-size patches, embedding these patches into a sequence of tokens, and then applying standard transformer layers to these tokens.

In addition to this baseline approach, we decided to experiment with pure multimodal models. Their main difference is that, instead of creating independent embeddings for each modality, their resulting features will be mapped into a common feature space.

- **CLIP[18] (Contrastive Language-Image Pre-training)**: aligns images and text embeddings by learning from a large dataset of image-text pairs. It maps both modalities into a shared embedding space to understand and associate textual descriptions with visual content effectively.
- **MULTI-CLIP[19]**: extends the CLIP model by incorporating additional modalities or using multiple views of the same data. This improves robustness and accuracy in understanding complex multimodal content.
- **SIGLIP[20] (Structured Image-Graph Language Interaction Pre-training)**: focuses on capturing the relationships between image regions and their corresponding textual descriptions, to understand and generate detailed and contextually accurate descriptions.

3.3.2 Cross-modal Fusion

This module will be in charge of performing the combinations between the image and text embeddings. Two different types of fusion were used during the experimentation:

- **Concat fusion**: directly joins both features into a single feature vector by concatenating them.
- **Cross fusion**: combines text and image embeddings by crossing the information from the vectors into a unified matrix to enhance the integration of multimodal information, resulting in more contextually rich and aligned features.

3.3.3 MLP Classifier

After the feature extraction, these embeddings will pass to a fixed neural network. It will include an initial layer with the unified feature size which will depend on the type of fusion used. After this layer, there will be a hidden layer with 128 neurons and with ReLU as activation function. The final layer will contain as many neurons as classes, using SoftMax as activation function. Also, after each one of the first two layers a fixed dropout rate of 25% will be used for regularization.

3.4 Model training

For the training of the models, the following techniques and parameters were established:

- **Loss function**: the loss function that will always be minimized during the training process is cross-entropy. This measurement is widely used for classification tasks as it quantifies the difference between the predicted probability distribution and the actual distribution.

– **Optimizer**: the Adam (Adaptive Moment Estimation) optimization algorithm is used, which is an extension of the Stochastic Gradient Descent (SGD) algorithm. It works by maintaining an adaptive learning rate for each of the network's weights, increasing the efficiency of the learning process.
– **Early stopping/checkpoint mechanism**: is a mechanism used during the training process, it is in charge of monitoring the validation loss on each epoch and storing the model when this error decreases. It has a 'patience' variable, that indicates the maximum number of epochs the training will continue if the validation error does not decrease.
– **Hyperparameters**: the training parameters used during the experimentation phase were the number of epochs, batch size, and learning rate. The batch size was fixed to 32 for all the experiments. For the number of epochs and the learning rate, their values were established empirically to guarantee the models convergence in a reasonable number of epochs.

4 Results and discussion

In this section, we present and analyze the results of experiments with different configurations. We discuss how each setup performed and contributed to solving the challenges of classifying memes using both text and image inputs. We will compare each one of the models generated in the experimentation phase taking a look at the classification reports generated over our test split and finally in section 4.3 we present the results from the competition.

4.1 Task 1: Detection of Hate Speech, Inappropriate, and Harmless memes

To evaluate whether extending the training dataset by adding new memes from different sources and annotating them with Google Gemini 1.5 (as described in Section 3.1) was a good approach to improve the results, we trained a baseline approach combining BETO and ViT models. The model trained only with the original training dataset obtained a macro F1 of 47% on our test split, while the use of the extended training dataset provided an improvement of 5 points in macro F1. Therefore, as expected, the inclusion of new memes helps to improve the results. Thus, we always use the extended training dataset in the following experiments.

Table 1 shows the results for our approaches, including our baseline model BETO + ViT, three different multimodal models (CLIP, SigLIP, and Multi-CLIP) with different fusion types, and the use of image-like augmentation.

ID	Architecture	Fusion type	Image-like augmentation	Precision	Recall	F1-score
B1	BETO + ViT	concat	no	0.53	0.52	0.52
1			yes	0.55	0.49	0.51
2		cross	yes	0.46	0.47	0.47
3	CLIP	concat	no	0.50	0.47	0.48
4			yes	0.54	0.52	0.51
5		cross	yes	0.56	0.51	**0.52**
6	SigLIP	concat	yes	0.47	0.47	0.46
7		cross	no	0.50	0.51	0.46
8			yes	0.51	0.53	**0.52**
9	Multi-CLIP	cross	no	0.48	0.49	0.48
10			yes	0.53	0.53	**0.52**

Table 1. Results on the test dataset. Best scores are in bold.

As seen in the **Table 1**, the macro F1-score ranges from 46% (SigLip model) to 52% (CLIP/SigLIP/Multi-CLIP + concat + image-like augmentation). However, the differences between the models are not significant enough to determine which one is better. We chose the multimodal models **5**, **8** and **9** because they share a common embeddding space for images and texts. Moreover, we decided to choose cross fusion and the use of image-like augmentation because they tend to provide slightly better results in previous experiments.

For the third and final phase, the experimental variable used was the multi-instance augmentation (explained in Section 3.1) over the previously selected configurations.**Table 2** presents the performance results on our test split. The underlined models are the ones that persist from the previous phase.

ID	Architecture	Multi-instance augmentation	Precision	Recall	F1-score
5	CLIP	no	0.56	0.51	0.52
5.1		yes	0.52	0.53	0.52
8	SigLIP	no	0.51	0.53	0.52
8.1		yes	0.52	0.52	**0.53**
10	Multi-CLIP	no	0.53	0.53	0.52
10.1		yes	0.53	0.53	**0.53**

Table 2. Comparison between the selected configurations with multi-instance augmentation

As it is seen on the **Table 2**, using multi-instance augmentation does not in fact make much of a difference in the models. Resulting in a slightly better performance in SigLIP and Multi-CLIP using this method.

The best configuration as seen in the previous tables, is Multi-CLIP with multi-instance augmentation, cross fusion and data augmentation. In the training evolution the validation loss quickly goes to minimum but gets stuck there, after that it starts overfitting as seen in the 6. It is important to note that with

different configurations of hyperparameters as learning rate and dropout rates, the validation loss remains stuck in a similar value.

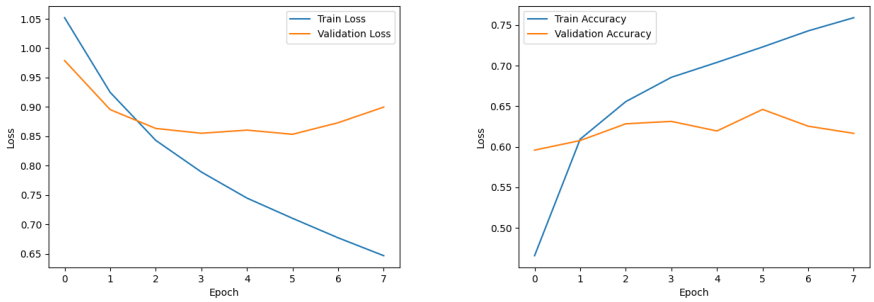

Fig. 6. *Loss* (left) and *accuracy* (right) evolution for the model 10.1

Finally, the confusion matrix of this configuration is the one presented in the 7. The conclusions about them are that there are a lot of none values and it is a common confusion to classify mostly inappropriate content memes into none.

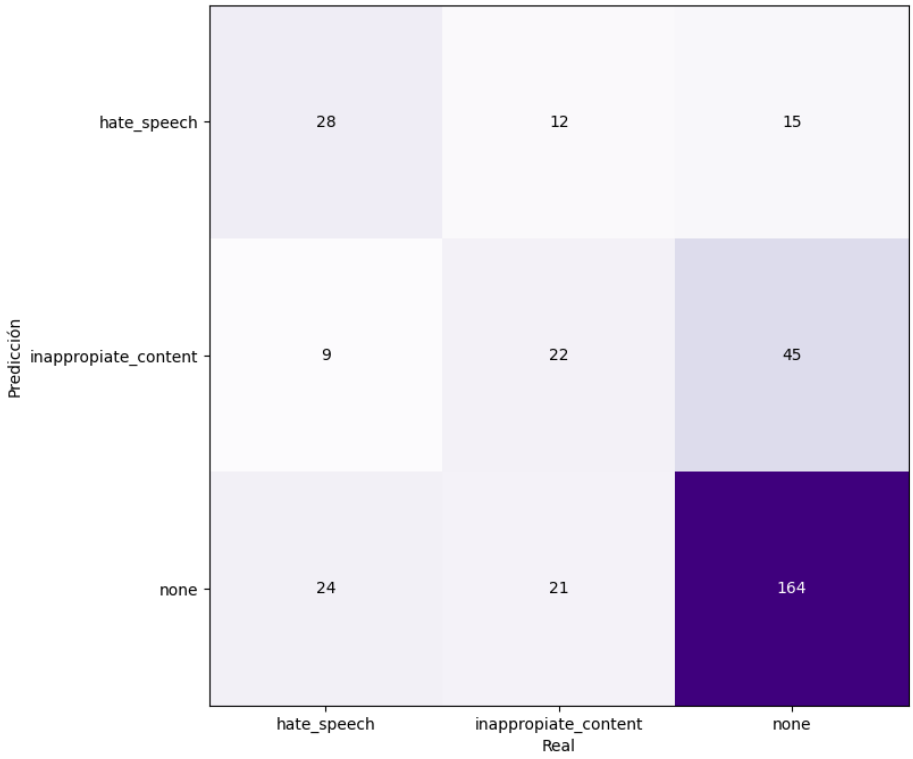

Fig. 7. Confusion matrix of the model 10.1

4.2 Task 2: Finer-grained detection of Hate Speech in Memes

It is important to note that the methodologies for classifying and training the two tasks were independent. First, we created an architecture to classify the classes from the first task. The final configurations selected as the best from the first task will be used to determine if a meme is hateful. The chosen configuration for the detailed classification of the hateful category will then be used in the second task.

As shown in **Table 3**the best models from task 1 SigLIP and Multi-CLIP both with "Multi-instance augmentation". Multi-CLIP, along with the SigLIP, is now used to predict specific hateful categories (racism, classism, sexism, or other). The Multi-CLIP model is significantly better than the SigLIP, so when the final models detect hate speech in the first task, the Multi-CLIP model will predict one of the four specific hateful categories.

ID	Architecture	Precision	Recall	F1-score
8.1_t2	SigLIP	0.59	0.59	0.58
10.1_t2	Multi-CLIP	0.76	0.71	**0.77**

Table 3. Results from the best 2 models on task 2 in our test split.

It is important to analyze the evolution on the training and validation splits. It is presented in the 9, and as it is seen, the evolution of the validation loss remains going down until the 10-12 epochs.

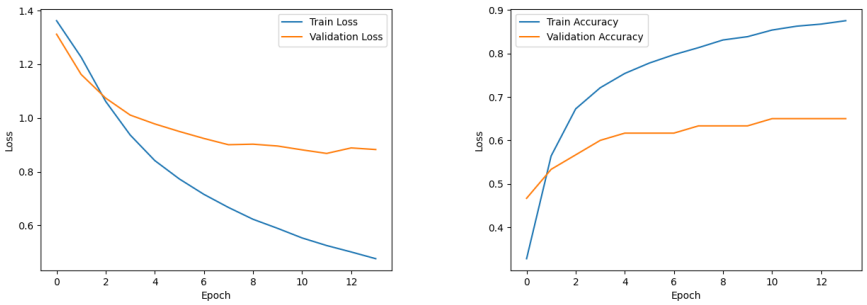

Fig. 8. *Loss* (left) and *accuracy* (right) evolution for the model 10.1

The confusion matrix in the 9 shows a good distribution in the main diagonal which results in better accuracy and performances in the finer-grained classification.

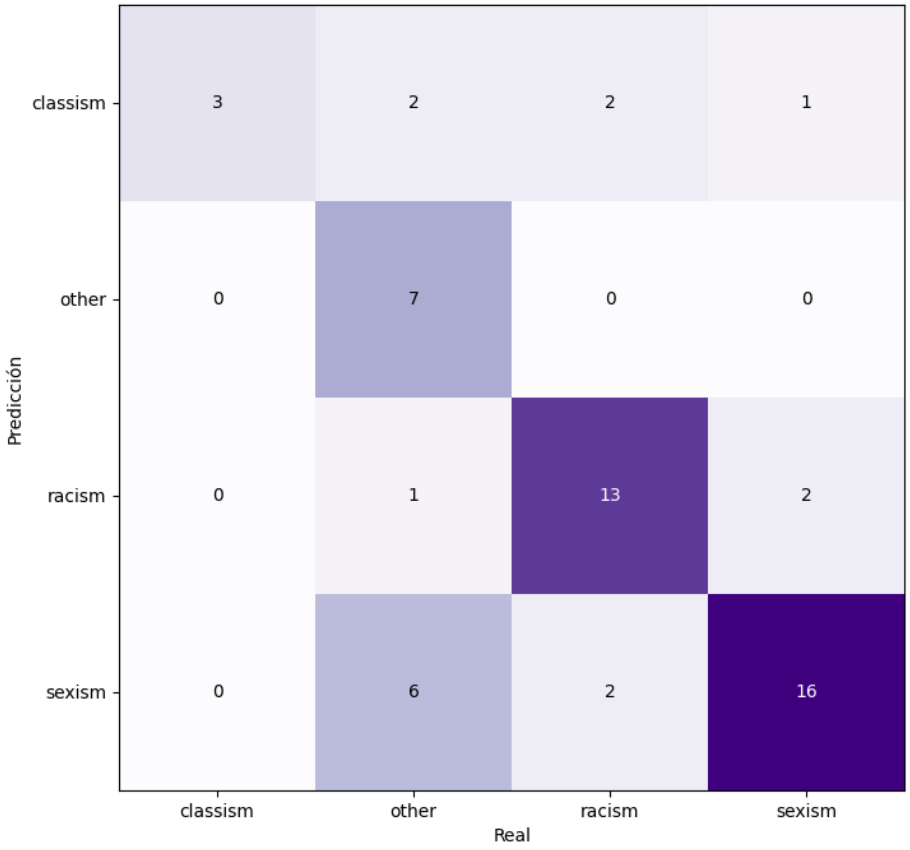

Fig. 9. Confusion matrix of the model 10.1

4.3 Results in the competition

For the first task, we submitted the approaches with the architectures CLIP, SigLip, and Multi-CLIP explained, all with multi-instance augmentation. However, the architecture with the best result in this competition was the same as in our tests: Multi-CLIP. This approach gave us the following results:

- **Precision**: 0.36
- **Recall**: 0.36
- **F1-score**: 0.36

For the second task, we used the approach of using the Multi-CLIP architecture in both parts of the task as we explained, since it gave us the best results. In the competition, when classifying between the six classes, the results were:

- **Precision**: 0.20

– **Recall**: 0.20

– **F1-score**: 0.20

4.4 Discussion

Looking at the results obtained in the competition, we can conclude that labeling the memes used for dataset extension with Gemini 1.5 Pro is not the best approach. Despite being a large multimodal model with many parameters, Gemini, although it provides correct reasoning, is not as accurate as initially thought. To verify this, Gemini was tested on predicting labels for a total of 220 images from the original labeled training set. When comparing the predicted labels with the actual ones, the results showed that in subtask 1, it had a macro-average F1-score of 0.53, and in subtask 2, a macro-average F1-score of 0.48. These results are not particularly good and much worse than expected.

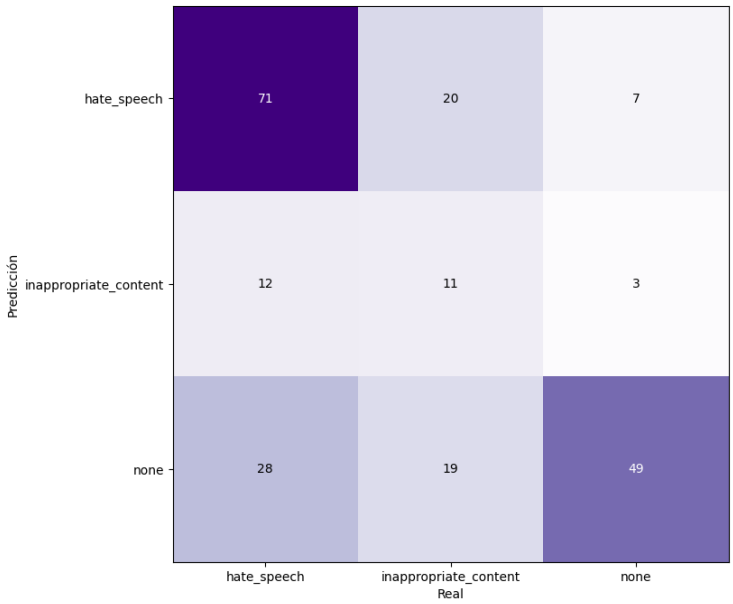

Fig. 10. Confussion matrix with Gemini predictions in subtask 1

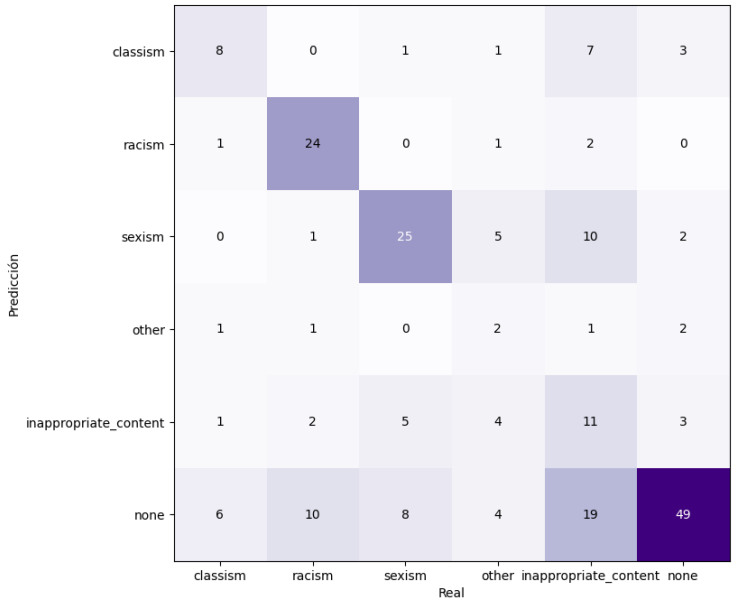

Fig. 11. Confussion matrix with Gemini predictions in subtask 2

Additionally, there is clear confusion between "inappropriate_content" and "hate_speech" since, even for a human, some cases are very difficult to distinguish. It may have been better to label these new data manually, despite the considerable effort involved.

5 Conclusions

In this study, we focused on detecting inappropriate memes in Mexican Spanish by enhancing our dataset and employing advanced multimodal classification techniques. Our approach involved adding new instances to the dataset using Google Gemini and utilizing data augmentation techniques for both text and images.

On the one hand, we added new meme instances from sources like Image Downloader, Telegram channels, and Reddit. This effort balanced the dataset by increasing the representation of underrepresented classes such as "inappropriate content" and "hate speech". Text extraction and initial labeling were performed using the Google Gemini 1.5 Pro model.

On the other hand, Meta's LLAMA 3 was used to generate variations of the text data, effectively increasing the text dataset size tenfold. Also to the images we applied transformations such as random rotation, affine transformation, and perspective distortion to increase the dataset's size and diversity.

The final architectures were chosen by the performances on our test split by changing features like concatenation or cross fusion, using the extended dataset or using our "Multi-instance augmentation". The models that produced the best performances were multimodal and, more specifically MultiCLIP as the feature extractor.

In the section 4.3 we can see that our group obtains way worse performance results than expected when we submitted our predictions using our models.

In the competition, we had worse results than expected, with F1: 0.36 in the first task and 0.2 in the second. For the first task, the best performance was observed in identifying "None" memes, and the worst was in identifying inappropriate content. In the second task, as in the first one, the "None" class obtained the best results, with the worst performance noted in detecting sexism.

Adding new memes helps balance the data, which reduces overfitting. Although text augmentation and image augmentation techniques provide improvements, they have not had a significant impact. We performed poorly compared to other contestants, finishing 7th out of 8 in the first task and 3rd out of 4 in the second. Although we achieved good results in our tests, this was not the case in the competition. We believe this is mainly due to the use of Gemini, which performed worse than we expected. The main reason that our extended dataset did not work well is because the criteria used to classify the memes in the original dataset is not the same as the one Gemini has to classify between the 3 or 6 classes of this competition. Especially when distinguishing between inappropriate content and hate speech.

Although we improved over the baseline of BETO-ViT, our techniques and the inclusion of more robust models, such as LLAMA 3, could help overcome these deficiencies.

By using different approaches with the newest and more innovative models nowadays, we developed some prediction models that can help to detect these types of offensive memes.

Overall, our work demonstrates the importance of dataset enhancement and multimodal techniques in improving the detection of inappropriate content, contributing to safer online environments.

Appendix

GitHub Repository

The dataset and the models can be found in the following GitHub repository: https://github.com/mario01gh/DIMEMEX-CyT-TEAM.

References

[1] W. C. Sleeman IV, R. Kapoor, and P. Ghosh, "Multimodal classification: Current landscape, taxonomy and future directions," *ACM Computing Surveys*, vol. 55, no. 7, pp. 1–31, 2022.

[2] A. Gandhi, K. Adhvaryu, and V. Khanduja, "Multimodal sentiment analysis: Review, application domains and future directions," in *2021 IEEE Pune section international conference (PuneCon)*, 2021, pp. 1–5.

[3] J. Fields, K. Chovanec, and P. Madiraju, "A survey of text classification with transformers: How wide? how large? how long? how accurate? how expensive? how safe?" *IEEE Access*, vol. 12, pp. 6518–6531, 2024. DOI: 10.1109/ACCESS.2024.3349952.

[4] O. of DiMex2024, *Dimex 2024 home*, https://sites.google.com/inaoe.mx/dimemex-2024/home, Accessed: 2024-06-04, 2024.

[5] H. Jarquín-Vásquez *et al.*, *Overview of da-vincis at iberlef 2023: Detection of aggressive and violent incidents from social media in spanish*, 2023. DOI: 10.26342/2023-71-27.

[6] G. K. Kumar and K. Nandakumar, "Hate-clipper: Multimodal hateful meme classification based on cross-modal interaction of clip features," 2022. DOI: 10.18653/v1/2022.nlp4pi-1.20.

[7] B. O. Sábat and X. G.-i.-N. C. Canto, "Multimodal hate speech detection in memes," Degree's Thesis, Universitat Politécnica de Catalunya, 2019.

[8] J. B. Simon, D. Karkada, N. Ghosh, and M. Belkin, "More is better in modern machine learning: When infinite overparameterization is optimal and overfitting is obligatory," Nov. 2023. [Online]. Available: https://arxiv.org/abs/2311.14646v4.

[9] *Qianyantech/image-downloader: Download images from google, bing, baidu.* . [Online]. Available: https://github.com/QianyanTech/Image-Downloader.

[10] G. Team *et al.*, *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context*, 2024. arXiv: 2403.05530 [cs.CL].

[11] J. Wang and L. Perez, "The effectiveness of data augmentation in image classification using deep learning,"

[12] H. Touvron *et al.*, *Llama: Open and efficient foundation language models*, 2023. arXiv: 2302.13971 [cs.CL].

[13] E. Ma, *Nlp augmentation*, https://github.com/makcedward/nlpaug, 2019.

[14] V. Marivate and T. Sefara, "Improving short text classification through global augmentation methods," in *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, Springer, 2020, pp. 385–399.

[15] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez, "Spanish pre-trained bert model and evaluation data," in *PML4DC at ICLR 2020*, 2020.

[16] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. arXiv: 1810.04805. [Online]. Available: http://arxiv.org/abs/1810.04805.

[17] A. Dosovitskiy *et al.*, *An image is worth 16x16 words: Transformers for image recognition at scale*, 2021. arXiv: 2010.11929 [cs.CV].

[18] A. Radford *et al.*, *Learning transferable visual models from natural language supervision*, 2021. arXiv: 2103.00020 [cs.CV].

[19] F. Carlsson, P. Eisen, F. Rekathati, and M. Sahlgren, "Cross-lingual and multilingual clip," in *Proceedings of the Language Resources and Evaluation Conference*, Marseille, France: European Language Resources Association, Jun. 2022, pp. 6848–6854. [Online]. Available: https://aclanthology.org/2022.lrec-1.739.

[20] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, *Sigmoid loss for language image pre-training*, 2023. arXiv: 2303.15343 [cs.CV].