

Overview of DIMEMEX at IberLEF 2024: Detection of Inappropriate Memes from Mexico

Resumen de la Tarea DIMEMEX en IberLEF 2024: Detección de Memes Inapropiados de México

Horacio Jarquín-Vásquez,¹ Itzel Tlelo-Coyotecatl,¹ Marco Casavantes,¹
Delia Irazú Hernández-Farías,¹ Hugo Jair Escalante,¹ Luis Villaseñor-Pineda,^{1,2}
Manuel Montes-y-Gómez¹

¹Laboratorio de Tecnologías del Lenguaje (INAOE), Mexico

²Centre de Recherche GRAMMATICA (EA 4521), Université d'Artois, France
{horacio.jarquin, itlelo, mcasavantes, dirazuhf, hugojair, villasen, mmontesg}@inaoep.mx

Abstract: In this paper, we present the overview of the DIMEMEX shared task which was organized at IberLEF 2024 and co-located in the framework of the 40th International Conference of the Spanish Society for Natural Language Processing (SEPLN 2024). The main aim of this task is to promote the research on developing automatic solutions for detecting inappropriate content in memes. Two subtasks were considered: (i) A three-way classification task aimed to determine if a meme contains hate speech, inappropriate content, or neither; and (ii) A fine-grained classification task in which a meme may be categorized into specific hate speech categories. A multimodal manual annotated corpus comprising both images and text associated with each meme was provided to the participants. A total of 5 systems were submitted for the final evaluation phase. Competitive results were reported for both subtasks being Subtask 1 the one with higher results. The data and results are available at the shared task website at <https://codalab.lisn.upsaclay.fr/competitions/18118>.

Keywords: DIMEMEX, hate speech detection, meme classification.

Resumen: En este documento presentamos el resumen de la tarea DIMEMEX organizada como parte del IberLEF 2024 junto con la 40^a Conferencia Internacional de la Sociedad Española para el Procesamiento de Lenguaje Natural (SEPLN 2024). El principal objetivo de la tarea es promover la investigación en el desarrollo de soluciones automáticas para la detección de contenido inapropiado en memes. Se consideraron dos tareas: (i) Clasificación ternaria cuyo objetivo es determinar si un meme contiene contenido relacionado con discurso de odio, contenido inapropiado o ninguno; y (ii) Clasificación de grano fino en la que el meme puede ser clasificado en una sub categoría de discurso de odio. Se les proporcionó a los participantes un conjunto de datos multimodal, anotado manualmente, que contiene tanto imágenes como texto relacionado con cada meme. Para la fase final de evaluación se recibieron en total 5 sistemas. Se reportaron resultados competitivos para ambas subtareas, siendo los mejores aquellos asociados a la tarea de clasificación ternaria. El conjunto de datos y los resultados detallados pueden consultarse en el sitio web de la tarea: <https://codalab.lisn.upsaclay.fr/competitions/18118>.

Palabras clave: DIMEMEX, detección de discurso de odio, clasificación de memes.

1 Introduction

Social media have become a powerful way of expression for people. Numerous platforms allow individuals the chance to express their thoughts openly, sometimes under the veil of anonymity. While freedom of expression is a human right, utilizing it to promote hostility towards others represents an abuse of this privilege (MacAvaney et al., 2019). Hate speech has the potential to cause various harms to individuals or groups, such as convincing others to adopt harmful stereotypes, emotional distress, and degrading human dignity (Gelber and McNamara, 2015). This behavior is considered an impactful issue of global concern for many countries and organizations (Nascimento, Cavalcanti, and Costa-Abreu, 2023). Despite social media platforms’ efforts to establish policies to regulate hateful behaviors through reporting tools, options to flag content, and content moderators, these measures are labor-intensive, time-consuming, and therefore not scalable or sustainable in the long term (Cao, Lee, and Hoang, 2021).

Computational approaches have been introduced as a way to facilitate the detection and monitoring of content through social media platforms. Considering that content sharing includes not only text but also images (e.g., memes) or video, studies have been conducted concerning the identification of hate speech in multimodal resources, with efforts focused on analyzing textual and visual content through binary classification approaches on multimodal datasets in English. Particularly (Suryawanshi et al., 2020) leveraged internet memes associated with the 2016 U.S. presidential election sourced from platforms including Reddit, Facebook, Twitter, and Instagram. Their efforts led to the creation of a multimodal meme dataset designated for offensive content detection known as Multi-OFF. It consists of 743 memes, annotated into either an *offensive* or *non-offensive* category.

The *hateful memes challenge* directs attention towards the detection of hate speech within multimodal memes (Kiela et al., 2020). They incorporated benign contrasting instances involving different images or captions for each hateful meme. After a series of filtering and annotation stages, their dataset culminated in 10k memes exactly.

The research conducted in SemEval-2022

Task 5 delves into *Multimedia Automatic Misogyny Identification (MAMI)*, with a specific focus on the identification of misogynous memes (Fersini et al., 2022). This task was divided into two sub-tasks: one centered on the identification of whether or not a meme exhibits misogyny, and another one dedicated to the identification of various forms of misogyny. For MAMI, approximately 11k memes were gathered from various social media platforms, and subsequently annotated by human annotators.

Unlike the previously mentioned related works, our task and dataset were designed with the intention of advancing the research and development of multimodal computational models, specifically to distinguish between inappropriate content and different types of hate speech in Mexican Spanish memes. The DIMEMEX task introduced a homonymous dataset, it is a collection of around 3k memes. These have been compiled from public Facebook groups rooted in Mexico and manually annotated on the presence of hate speech, inappropriate content, and harmful content.

The DIMEMEX shared task comprises two subtasks: i) A three-way classification task to distinguish instances containing hate speech, inappropriate content, or neither; and ii) A finer-grained classification task to categorize instances of hate speech into specific categories like classism, sexism, and racism. Seven teams submitted results for Subtask 1, with four of them also participating in Subtask 2. The evaluation encompassed a stimulating array of proposals. Notably, Transformer-based approaches exhibited a dominant presence and outperformed traditional machine learning methods; the participating teams employed a wide variety of pre-trained language models, vision models, and multimodal models. The obtained results in the evaluation demonstrate the difficulty of both subtasks. These results motivate further research in this task and the creation of better pre-trained multimodal models in Spanish capable of aligning the information present in the images and texts of memes on social networks.

The remainder of this paper is organized as follows. Section 2 describes the dataset developed and related details of the shared task. Section 3 presents a summary of the approaches proposed by the participating sys-

tems as well as the obtained results. Finally, our conclusions and future work are exposed in Section 4.

2 Task description

2.1 Dataset

The DIMEMEX dataset consists of a set of memes manually annotated on the presence of abusive content compiled from public Facebook groups rooted in Mexico that are dedicated to the distribution of this kind of content. Our study specifically focuses on the identification of inappropriate content, various types of hate speech, and non-abusive content (indicated by the ‘*neither*’ label); we regard all the categories as mutually exclusive to maintain clear distinctions within our analysis. Specifically, a meme is classified as containing hate speech if it presents any kind of communication in speech, writing, or behavior, that attacks or uses pejorative or discriminatory language with reference to a person or a group based on their identity factors (Davidson et al., 2017). Additionally, a meme is deemed to contain inappropriate content if it exhibits any kind of manifestation of offensive, vulgar (profane, obscene, sexually charged), and/or morbid humor content. Regarding the different types of hate speech, we focus on the following categories and definitions:

- *Classism*. Any manifestation that promotes an attitude or tendency to discriminate or minimize someone based on the difference of **social status**.
- *Racism*. Any manifestation that promotes an attitude or tendency to discriminate or minimize someone based on **ethnic characteristics** or that promotes the **superiority of a group**.
- *Sexism*. Any manifestation that promotes an attitude or tendency to discriminate or minimize someone based on **gender characteristics**. This includes misogyny, misandrist, and LGBTQ+ related content.
- *Other*. Any manifestation that promotes an attitude or tendency to discriminate or minimize someone based on characteristics that do not belong to the previously defined ones.

Concerning the data collection approach, we gathered all images using the *facebook-scrapers*¹ library. More than twelve thousand images were collected and passed through a manual filtering process to ensure their quality: images with low resolution, with non-Spanish text, and images having only text with a monochromatic background or without text, were discarded. Additionally, all repeated images were removed. This filtering process resulted in the discarding of over nine thousand images. The remaining 3,300 instances served as the starting point to develop the dataset.

Manual Annotation

Once the filtering process was completed, a pilot task on a subset of 300 images divided into two partitions was performed. Two groups of annotators were asked to label each partition of samples on the presence of abusive content. They were not provided with any annotation guidelines, instead, we asked them to follow their own definition of the considered categories. Inter-annotator agreement rates were low, reflecting the complexity of the task. According to our annotators’ feedback discussions, annotation guidelines were defined by including explicit definitions for each category which were used for labeling the final dataset.

After analyzing the results of the pilot task, we moved to the final labeling stage. In this phase, our team comprised 12 annotators, including 5 men and 7 women, all native Spanish speakers from Mexico. To ensure a balanced distribution, the dataset was divided into four partitions, each containing 825 instances. The distribution of annotators across these partitions was meticulously planned to achieve as equitable a balance as possible between male and female annotators. Consequently, one partition was assigned to 2 men and 1 woman, while the other three partitions were each handled by 1 man and 2 women. The final label for each meme was determined by a majority vote; out of the total 3,300 instances, there were 259 instances where the annotators initially assigned different categories. These instances underwent a final re-labeling stage where all annotators convened and voted to decide on the final label. During this process,

¹<https://pypi.org/project/facebook-scrapers/>

instances with split votes and those causing uncertainty among annotators were excluded, resulting in a refined final dataset of 3,235 memes.

Table 1 displays the average Fleiss’ kappa values achieved by the four groups of annotators, considering the seven classes of our DIMEMEX dataset. Moderate agreement was reached in 5 of the 7 classes, while the ‘*other*’ and ‘*neither*’ classes showed fair agreement. The final distribution of the classes is as follows: *classism*: 63, *racism*: 163, *sexism*: 223, *other*: 103, *inappropriate content*: 675, and *neither*: 2008.

Class	Kappa value
Classism	0.4301
Racism	0.5148
Sexism	0.4764
Other	0.3445
Inappropriate content	0.4552
Neither	0.3830
Hate speech	0.4343

Table 1: Average Fleiss’ kappa values obtained from the four groups of annotators, considering the seven classes defined for the DIMEMEX dataset.

2.2 Subtasks

DIMEMEX encompasses two subtasks: i) A three-way classification task, which involves distinguishing instances containing hate speech, inappropriate content, or neither; and ii) A finer-grained classification task, which involves categorizing instances of hate speech into specific categories such as classism, sexism, racism, and others.

Both subtasks were evaluated using the DIMEMEX dataset. The CodaLab platform (Pavao et al., 2022) was used to run the challenge. The shared task was divided into the following two phases:

- **Development phase.** Both labeled training data and unlabeled validation data were available to participants. Participants were able to submit to the CodaLab website their predictions for the validation set during this phase receiving a quick evaluation.
- **Final phase.** Unlabeled test results were made available to participants. During the contest, they could upload up to ten submissions. Participants were ranked based on their performance on the test set.

For evaluation purposes, we considered the macro average recall, precision, and f_1 score for both subtasks. Being the latter score, the leading evaluation measure in both subtasks.

2.3 Baselines

As baseline methods three established approaches known for their strong performance in tasks involving image and text classification were chosen. The first baseline approach focuses solely on the text modality and involves the fine-tuning of the pre-trained BETO² model. Using only the visual modality, a second baseline was defined. It entails fine-tuning the pre-trained Vision Transformer³ (ViT) model. Lastly, our third baseline method integrates both modalities (visual and textual) by employing an early fusion technique that concatenates the classification vectors obtained from BETO and ViT models. For the sake of readability, the obtained results of these baselines were included and are referred to as Baseline (TXT), Baseline (IMG), and Baseline (TXT + IMG).

3 Overview of Participating Systems

The subsequent subsections provide an overview of the primary concepts explored by the participating systems, followed by an overall evaluation of their findings.

3.1 Systems’ Descriptions

A total of 7 teams submitted results for Subtask 1, with 4 of them also participating in Subtask 2. However, only 5 teams reported descriptions of their solutions, Table 2 summarizes the information of the systems submitted by the participants. As can be observed, most teams used both modalities to address the task of detecting hate speech and inappropriate content in Mexican Spanish memes. All teams utilized different pre-trained language models, vision models, and multimodal models like CLIP, employing different fusion approaches ranging from early fusion to cross-modal fusion. Some approaches explored different strategies for data augmentation and the correction of the text in the memes provided by the organizers.

²<https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased>

³https://huggingface.co/docs/transformers/model_doc/vit

Team	Pre-processing		Pre-trained models		Fusion	Model	Extra Details
	Text	Image	Text	Image			
CyT (García-Hidalgo et al., 2024)	x	x	BETO	ViT CLIP Multi-CLIP SigLIP	Concat Cross-modal	MLP	Augmented dataset
UCM (Parveen, 2024)	x	x	BERT	ViT	Concat	LR RF MLP	
fariha32 (Maqbool and Fersini, 2024)	x	x	BLIP Image Text Encoder		Concat		Text was translated to English
ITC (Cabada et al., 2024)	x	x	BETO	ViT	Concat	GBC	Corrected the provided text
CLTL (Wang and Markov, 2024a)			XLM-T Multilingual-E5 RoBERTa-base-BNE BETO	Swin Transformer V2	Concat	MLP	
<i>Baseline</i> (TXT)	x		BETO				
<i>Baseline</i> (IMG)		x		ViT			
<i>Baseline</i> (TXT + IMG)	x	x	BETO		Concat		

Table 2: Summary of participant systems’ descriptions were pre-processing for text and image modalities are indicated with an **x**, used pre-trained models and models (MLP-Multi Layer Perceptron, LR-Linear Regression, RF-Random Forest, GBC-Gradient Boosting Classifier) are listed, fusion methods are specified (**concatenation**), and extra details of the approaches are described.

The detailed descriptions of the participants’ systems are presented below.

Team CyT explored a combination of features fusion through concatenation and cross-modal fusion into multimodal models (BETO+ViT, CLIP, MULTI-CLIP, and SIGLIP). An augmentation approach to the originally provided dataset was proposed to tackle the data imbalance. This augmentation included adding up memes labeled using Gemini 1.5 and sourced from Telegram, Reddit, and the Image Downloader tool. Additionally, they suggested augmenting data by rephrasing the text extracted from the memes and applying various transformations to the images (García-Hidalgo et al., 2024).

Team UCM applied a variety of classification models (Logistic Regression, Random Forest, Multi Layer Perceptron) adjusting the hyperparameters to find the best combination, as well as the starting kit model provided by the organizers. All models were fed with previously preprocessed text and image features (Parveen, 2024).

Team fariha32 proposed a vision-language based pre-trained model named *BLIP* to extract combined image text embeddings, and then used a *Gradient Boosting Classifier* to detect the class of each meme. As preprocessing for text, they decided to translate it to English (Maqbool and Fersini, 2024).

Team ITC adopted a BETO model classification approach that was enhanced by a data preprocessing scheme. The preprocessing included cleaning and correction of erroneous text (provided by the organizers) by

using LLAMA, OCR method, and the integration of memes captions obtained from ViT (Cabada et al., 2024).

Team CLTL explored the combination of four state-of-the-art language models (*XLM-T*, *Multilingual-E5*, *RoBERTa-base-BNE*, *BETO*) with the *Swin* Transformer-based visual model to create a multimodal system with a Multilayer Perceptron fusion module for classification, achieving the highest results for both tasks (Wang and Markov, 2024a).

3.2 Evaluation campaign results

Table 3 presents the obtained results by the participating teams in Subtask 1, which involves distinguishing memes containing hate speech, inappropriate content, or neither. The teams are ranked in descending of macro average f_1 -score. Macro average Precision and Recall values are also provided to facilitate a comprehensive interpretation of these findings. Notably, the results highlighted in gray correspond to the teams that submitted working notes containing descriptions of their systems. Additionally, we have included the results obtained from our three baseline approaches.⁴

The CLTL team achieved the highest performance in Subtask 1 (Wang and Markov, 2024a), followed by the CUFE and UCM (Parveen, 2024) teams. As can be observed, the difference between the top places is very small; the difference between the first and second places was only 0.02 in the macro

⁴In this paper, we report the median performance over 3 runs for each baseline, as it offers a more reliable estimation of their performance.

Subtask 1:			
Team	Precision	Recall	F1-Score
CLTL (Wang and Markov, 2024a)	0.61	0.56	0.58
CUFE	0.63	0.53	0.56
UCM (Parveen, 2024)	0.49	0.49	0.49
Baseline (TXT + IMG)	0.49	0.48	0.48
ITC (Cabada et al., 2024)	0.48	0.47	0.48
fariha32 (Maqbool and Fersini, 2024)	0.52	0.50	0.47
Baseline (TXT)	0.51	0.48	0.46
Baseline (IMG)	0.45	0.42	0.43
mashd3v	0.48	0.42	0.42
CyT (García-Hidalgo et al., 2024)	0.36	0.36	0.36

Table 3: Results of the participant teams in Subtasks 1. Bold numbers correspond to the best results of each metric.

average f_1 -score, while the difference between the second and third places was 0.07. The approaches that secured the top places shared the commonality of utilizing both modalities by fusing features extracted with state-of-the-art pre-trained transformer models. Specifically, the CLTL team used models that had previously achieved state-of-the-art results in detecting hate speech related to the Russia-Ukraine conflict (Wang and Markov, 2024b), which may have contributed to their strong performance in both subtasks. The use of both modalities demonstrates the advantage of leveraging the complementarity provided by textual and visual information for detecting hate speech and inappropriate content in memes.

In order to further analyze the differences in performance we carried out a statistical analysis using the tool by (Nava-Muñoz, Graff-Guerrero, and Escalante, 2023). Figure 1 shows confidence intervals to the mean with 95% confidence for the macro average f_1 scores, 1000 bootstrap samples were considered (details can be found in (Nava-Muñoz, Graff-Guerrero, and Escalante, 2023)). From this Figure it can be confirmed that differences between the top-2 ranked teams in terms of macro average f_1 scores are not statistically significant. These results align with those presented in Table 3, where the difference between the first and second place is only 0.02.

The results obtained in Subtask 2, the finer-grained classification task involving the categorization of instances of hate speech into specific categories, are shown in Table 4. The highest performance in this subtask is attributed to the CLTL team, followed by the CUFE and CyT (García-Hidalgo et al., 2024) teams. These teams primarily focused on employing pre-trained transformer models and implementing data fusion techniques in both

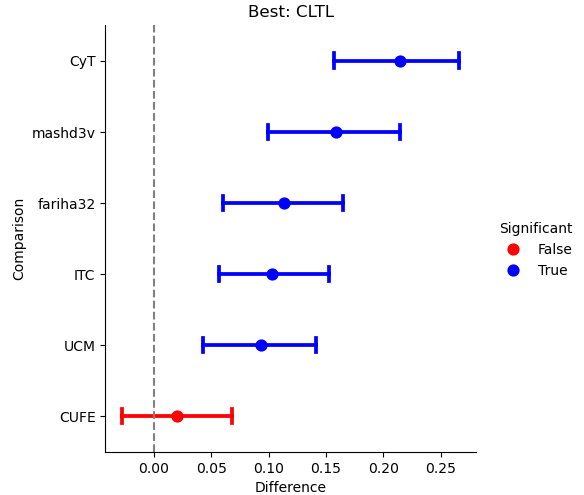


Figure 1: 95% confidence intervals for subtask 1 using bootstrapping.

image and text modalities. The CyT team secured the third-best performance by enhancing the quality of the text provided by the task organizers and integrating image captions into the BETO model. The results obtained in this subtask were relatively low, with a macro average f_1 score of 0.44 being the best result achieved by the participating teams. This reflects the complexity of detecting different types of hate speech and highlights the need for developing multimodal models capable of accurately identifying nuanced instances of hate speech. Notably, the CLTL team also obtained the first position in Subtask 1, indicating the effectiveness and robustness of their approach to the task of identifying hate speech and inappropriate content in social media memes.

Subtask 2:			
Team	Precision	Recall	F1-Score
CLTL (Wang and Markov, 2024a)	0.52	0.42	0.44
CUFE	0.52	0.33	0.37
Baseline (TXT + IMG)	0.35	0.29	0.29
Baseline (IMG)	0.31	0.26	0.28
Baseline (TXT)	0.31	0.26	0.27
CyT (García-Hidalgo et al., 2024)	0.20	0.20	0.20
mashd3v	0.29	0.20	0.18

Table 4: Results of the participant teams in Subtask 2. Results in bold correspond to the best results of each measure.

Figure 2 shows confidence intervals for Subtask 2. Unlike Figure 1, where no significant difference was found between the first and second places, this figure confirms a significant difference between the first and second places. This aligns with the larger difference observed in the results of Subtask 2,

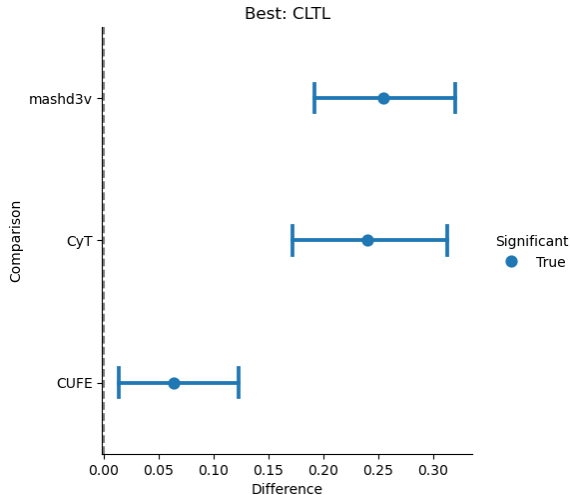


Figure 2: 95% confidence intervals for sub-task 2 using bootstrapping.

with a difference of 0.07 in the macro average f_1 score. Additionally, it can be observed that there is a significant difference between the second place and the third and fourth places. This is attributed to the substantial performance gap between the second and third place, with a difference of 0.17 in the macro average f_1 score.

3.3 Analysis

In order to conduct a more comprehensive analysis of the obtained results by the participants, we first focused on examining the complementarity and diversity observed in their predictions. To quantify the level of complementarity, we employed the *Maximum Possible Accuracy (MPA)* metric, which is defined as the ratio of correctly classified instances to the total number of test instances. In this case, an instance is considered correctly classified if at least one of the participating teams correctly classified it. For measuring diversity, we utilized the *Coincident Failure Diversity (CFD)* metric (Tang, Suganthan, and Yao, 2006), which focuses on calculating the error diversity among the participant’s predictions. The minimum value of this metric is 0, indicating that all teams simultaneously predict a pattern correctly or wrongly, while the maximum value is 1, representing unique misclassifications.

The results of applying the MPA and CFD metrics to assess the performance of all participating teams in the task of identifying hate speech and inappropriate content are presented in Table 5. Regarding the results of

	Class	BPA	MPA	CFD	
1	all classes	0.779	0.916	0.808	7
1	hate	0.491	0.767	0.510	7
1	inappropriate	0.548	0.818	0.451	7
1	neither	0.937	0.997	0.591	7
2	all classes	0.670	0.827	0.649	4
2	classism	0.153	0.231	0.237	4
2	racism	0.515	0.545	0.149	4
2	sexism	0.644	0.711	0.162	4
2	other	0.047	0.126	0.104	4

Table 5: Comparison of BPA, MPA, and CFD results between the different general approaches. The first column refers to the sub-task number while the last one refers to the number of systems’ results involved in the calculation.

Subtask 1 (rows 2-5), these results reveal that the MPA values achieved by all teams are remarkably higher than the Best Performance Accuracy (BPA) achieved by the top two performing teams. This suggests that the systems and approaches employed by all participants exhibit a high degree of complementarity. In contrast, the results obtained in Subtask 2 (rows 6-10) show a smaller increase in MPA values compared to the BPA achieved by the top two performing teams. This may be due to the smaller number of teams participating in Subtask 2 and the difficulty in distinguishing between different types of hate speech. From this analysis, it can be concluded that the most challenging classes to distinguish in Subtask 2 were *classism* and *other*. Additionally, Table 5 also compares the diversity of errors among the participating teams. Greater diversity is observed in Subtask 1 compared to Subtask 2 (rows 2-5 vs. rows 6-10). This finding aligns with the improvement observed in the MPA metric for both subtasks.

Qualitative Analysis

NOTE: This subsection contains samples that may be offensive to some readers, these do not represent the perspectives of the authors.

To further analyze the outcomes of the participating systems, we decided to take advantage of the obtained results in terms of MPA with the objective of identifying those instances that were misclassified by all the proposed approaches. A manual qualitative analysis was performed over a subset of these

memes. In the following paragraphs, we briefly describe the main observed features as well as some samples of these instances.






Category: Hate	Category: Hate
	
Translation: When you see that your parents' address is on Insurgentes Sur.	Translation: F**k it life goes on.
Category: Inappropriate	Category: Inappropriate
La ciencia ha llegado a los nuevos interruptores. 	
Translation: Science has reached the new switches.	Translation: Life is like a priest, you never know what you're going to get. ⁵
Category: Neither	
	
Translation: - Every time we talk, I end up wet. - Do I turn you on? - No, you spit when you talk.	

Table 6: Samples of memes that were incorrectly classified by all participating teams in Subtask 1.

Table 6 shows examples of memes misclassified by all participating teams in Subtask 1. A common denominator found in most of these memes is the need for extralinguistic context for correct interpretation. For instance, the second meme in the upper right corner adds the suffix “*tl*” to certain words, it is commonly used to mock people who speak indigenous languages in Mexico. Another fundamental aspect for detecting inappropriate content and hate speech in

⁵The verb “get” in Spanish has various meanings. In the context of this sentence, it can be interpreted as “touch”.

memes is the correct interpretation of both image and text. This is exemplified by the third meme in the middle left, where the textual content “*Science has reached the new switches*” seems harmless, but the addition of visual content referencing intimate body parts turns it into vulgar content with sexual connotations.

We also identified some instances that even labeled incorrectly by all systems maybe not considered as mistakes at all. An example is the last meme, which belongs to the “Neither” category but was classified as inappropriate content by most teams, likely due to the inclusion of an initial conversation with sexual overtones.

Category: Classism	Category: Racisms
cuando estas con tu celular en la calle y un niño te pregunta si tienes free fire: 	Cuando te dice “Mi chocolatito” pero cuando esta enojada te grita: Cállate Coca cola con ojos. 
When you're on your phone in the street and a kid asks if you have Free Fire - let me guess, public school?	Translation: When she calls you ‘my little chocolate’ — when she’s mad she says: shut up, Coke bottle with eyes.
Category: Sexism	Category: Other
cómo que no 	*La morra que pesa más que un tanque soviético dice que le gustan los chicos de color* El chico de color del curso: 
Translation: What do you mean you didn't cook anything?	Translation: The girl who weighs more than a Soviet tank says she likes black guys. The black guy in the class:

Table 7: Samples of memes that were incorrectly classified by all participating teams in Subtask 2.

Finally, Table 7 shows examples of memes misclassified by all participating teams in Subtask 2. Again, the need for extralinguistic context is evident for correctly classifying the different types of hate speech in memes. For example, the fourth meme in the lower right corner uses the expression “*Soviet tank*” to mock an overweight person. The correct interpretation of both image and text is also crucial, as seen in the

third meme in the lower left corner, where a drawing of a man practicing various strikes, combined with the textual content, promotes violence against women. These characteristics identified in misclassified memes reveal the complexity of this task, as well as the low performance achieved by participating teams in Subtask 2. They also highlight the need for new multimodal models and resources in Spanish.

4 Conclusions

We presented an overview of the DIMEMEX shared task organized within the framework of IberLEF. DIMEMEX promotes research into the identification of hate speech and inappropriate content in Mexican Spanish memes, a task of substantial societal significance due to the increasing rise of this type of content on social networks. In this shared task, we presented a new dataset consisting of 3,235 memes collected from public Facebook groups. This shared task featured two subtasks: Subtask 1 consisted of a three-way classification task, which involved distinguishing instances containing hate speech, inappropriate content, or neither; while Subtask 2 consisted of a finer-grained classification task, which involved categorizing instances of hate speech into specific categories such as classism, sexism, racism, and others. This evaluation campaign facilitated the assessment of a wide array of approaches, enabling a comparative analysis of their effectiveness. Various models, features, and techniques were presented within the proposed approaches, thereby contributing to advancements in this field.

The evaluation encompassed a stimulating array of proposals; all the approaches proposed by the participating teams utilized text and image modalities for the detection of hate speech and inappropriate content. Notably, Transformer-based approaches exhibited a dominant presence and outperformed traditional machine learning methods; the participating teams employed a wide variety of pre-trained language models, vision models, as well as multimodal models like CLIP and Multi-CLIP. Different teams introduced notable innovations such as data augmentation with other memes from social networks labeled using Gemini 1.5, the introduction of image captioning into the task, augmenting the contextual information available, the

improvement of the text extracted from the memes using LLaMA, and the use of different fusion approaches to combine the information from the text and image modalities.

The obtained results in the evaluation demonstrate the difficulty of both subtasks. As anticipated, the finer-grained Subtask 2 presented greater difficulty compared to the three-way classification involved in Subtask 1, due to the significant class imbalance in the different types of hate speech, the scarce number of pre-trained vision and language models in Spanish, and the difficulty both tasks presented in requiring extralinguistic context for the correct interpretation of the memes. These results motivate further research in this task and the creation of better models capable of aligning the information present in the images and texts of memes on social networks.

Building upon the significant areas of opportunity in this shared task, future work is proposed to extend the current dataset, reducing the class imbalance, and creating a new multi-class, multi-label subtask. An additional challenge to be considered within this task involves extending the scope of hate speech and inappropriate content detection to encompass videos disseminated on social networks.

Acknowledgements

We thank CONAHCyT-Mexico for partially supporting this work under scholarships 925996, 972915 and 883688.

References

- Cabada, R. Z., M. L. B. Estrada, R. A. C. Sapien, V. M. B. Beltrán, N. L. López, and M. A. S. Rivas. 2024. DIMEMEX at IberLEF 2024: When hate goes Viral: Detection of Hate Speech in Mexican Memes Using Transformers. In S. M. Jiménez-Zafra, L. Chiruzzo, F. Rangel, U. B. Corrêa, A. B. Jover, H. Gómez-Adorno, J. Á. G. Barba, D. I. H. Farías, A. M. Ráez, P. Moral, C. R. Abellán, M. E. V. Rodríguez, M. Taulé, and R. Valencia-García, editors, *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2024)*, Valladolid, Spain, September, 2024, CEUR Workshop Proceedings, Valladolid, Spain. CEUR-WS.org.

- Cao, R., R. K.-W. Lee, and T. Hoang. 2021. DeepHate: Hate speech detection via multi-faceted text representations. In *Proceedings of the 12th ACM Conference on Web Science*, 03.
- Davidson, T., D. Warmusley, M. W. Macy, and I. Weber. 2017. Automated hate speech detection and the problem of offensive language. In *International Conference on Web and Social Media*.
- Fersini, E., F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, and J. Sorensen. 2022. SemEval-2022 task 5: Multimedia automatic misogyny identification. In G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, and S. Ratan, editors, *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States, July. Association for Computational Linguistics.
- García-Hidalgo, M., M. García-Rodríguez, J. Payno, M. F. Salerno, and I. Segura-Bedmar. 2024. DIMEMEX-2024: CyT at DIMEMEX: Leveraging Data Augmentation for the Detection of Hate Speech in Memes. In S. M. Jiménez-Zafra, L. Chiruzzo, F. Rangel, U. B. Corrêa, A. B. Jover, H. Gómez-Adorno, J. Á. G. Barba, D. I. H. Farías, A. M. Ráez, P. Moral, C. R. Abellán, M. E. V. Rodríguez, M. Taulé, and R. Valencia-García, editors, *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2024)*, Valladolid, Spain, September, 2024, CEUR Workshop Proceedings, Valladolid, Spain. CEUR-WS.org.
- Gelber, K. and L. McNamara. 2015. Evidencing the harms of hate speech. *Social Identities*, 22:1–18, 12.
- Kiela, D., H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624.
- MacAvaney, S., H.-R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.
- Maqbool, F. and E. Fersini. 2024. Multimodal Hate Speech Detection in Memes from Mexico using BLIP. In S. M. Jiménez-Zafra, L. Chiruzzo, F. Rangel, U. B. Corrêa, A. B. Jover, H. Gómez-Adorno, J. Á. G. Barba, D. I. H. Farías, A. M. Ráez, P. Moral, C. R. Abellán, M. E. V. Rodríguez, M. Taulé, and R. Valencia-García, editors, *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2024)*, Valladolid, Spain, September, 2024, CEUR Workshop Proceedings, Valladolid, Spain. CEUR-WS.org.
- Nascimento, F. R. S., G. D. C. Cavalcanti, and M. D. Costa-Abreu. 2023. Exploring automatic hate speech detection on social media: A focus on content-based analysis. *Sage Open*, 13(2):21582440231181311.
- Nava-Muñoz, S., M. Graff-Guerrero, and H. J. Escalante. 2023. Comparison of classifiers in challenge scheme. In *Pattern Recognition - 15th Mexican Conference, MCPR 2023, Tepic, Mexico, June 21-24, 2023, Proceedings*, volume 13902 of *Lecture Notes in Computer Science*, pages 89–98. Springer.
- Parveen, A. A. 2024. UCM’s Participation to the 2024 DIMEMEX Task: Automatic Detection of Inappropriate Memes in Mexico. In S. M. Jiménez-Zafra, L. Chiruzzo, F. Rangel, U. B. Corrêa, A. B. Jover, H. Gómez-Adorno, J. Á. G. Barba, D. I. H. Farías, A. M. Ráez, P. Moral, C. R. Abellán, M. E. V. Rodríguez, M. Taulé, and R. Valencia-García, editors, *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2024)*, Valladolid, Spain, September, 2024, CEUR Workshop Proceedings, Valladolid, Spain. CEUR-WS.org.
- Pavao, A., I. Guyon, A.-C. Letournel, X. Baró, H. Escalante, S. Escalera, T. Thomas, and Z. Xu. 2022. CodaLab Competitions: An open source platform

- to organize scientific challenges. Technical report, Université Paris-Saclay, FRA., April.
- Suryawanshi, S., B. R. Chakravarthi, M. Arcan, and P. Buitelaar. 2020. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In R. Kumar, A. K. Ojha, B. Lahiri, M. Zampieri, S. Malmasi, V. Murdock, and D. Kadar, editors, *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France, May. European Language Resources Association (ELRA).
- Tang, E. K., P. N. Suganthan, and X. Yao. 2006. An analysis of diversity measures. *Mach. Learn.*, 65(1):247–271.
- Wang, Y. and I. Markov. 2024a. CLTL at DIMEMEX Shared Task: Fine-Grained Detection of Hate Speech in Memes. In S. M. Jiménez-Zafra, L. Chiruzzo, F. Rangel, U. B. Corrêa, A. B. Jover, H. Gómez-Adorno, J. Á. G. Barba, D. I. H. Farías, A. M. Ráez, P. Moral, C. R. Abellán, M. E. V. Rodríguez, M. Taulé, and R. Valencia-García, editors, *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024) co-located with the Conference of the Spanish Society for Natural Language Processing (SE-PLN 2024)*, Valladolid, Spain, September, 2024, CEUR Workshop Proceedings, Valladolid, Spain. CEUR-WS.org.
- Wang, Y. and I. Markov. 2024b. CLTL@multimodal hate speech event detection 2024: The winning approach to detecting multimodal hate speech and its targets. In A. Hürriyetöglu, H. Tanev, S. Thapa, and G. Uludoğan, editors, *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 73–78, St. Julians, Malta, March. Association for Computational Linguistics.