Procesamiento del Lenguaje Natural, Revista nº 75, septiembre de 2025, pp. 401-412 recibido 03-07-2025 revisado 18-07-2025 aceptado 01-08-2025

Overview of DIMEMEX at IberLEF 2025: Detection of Inappropriate Memes from Mexico

Resumen de la Tarea DIMEMEX en IberLEF 2025: Detección de Memes Inapropiados de México

Horacio Jarquín-Vásquez,[2] Itzel Tlelo-Coyotecatl,[1] Delia Irazú Hernández-Farías,[1] Hugo Jair Escalante,[1] Luis Villaseñor-Pineda,[1,3] Manuel Montes-y-Gómez[1]

[1]Laboratorio de Tecnologías del Lenguaje (INAOE), Mexico
[2]Dipartimento di Informatica, Universita degli Studi di Torino, Italy
[3]Centre de Recherche GRAMMATICA (EA 4521), Université d'Artois, France
{itlelo, dirazuhf, hugojair, villasen, mmontesg}@inaoep.mx
horaciojesus.jarquinvasquez@unito.it

Abstract: This paper presents the overview of the DIMEMEX shared task, organized at IberLEF 2025 and co-located with the 41^{th} International Conference of the Spanish Society for Natural Language Processing (SEPLN 2025). The aim of this task is to promote research on automatic solutions for detecting inappropriate content in memes with a particular focus on Mexican Spanish. Three subtasks were considered: (i) A three-way classification task aimed at determining whether a meme contains hate speech, inappropriate content, or neither; (ii) A fine-grained classification task in which a meme may be categorized into specific hate speech categories; and (iii) A three-way classification, as in (i), restricting participants to exclusively focus on leveraging the use of Large Language Models (LLMs). Participants were provided with a multimodal manual annotated corpus comprising both images and text associated with each meme. As a result, a total of 6 teams out of a total of 10 submissions reported their system descriptions for the final evaluation phase. Results show competitive performance for all subtasks being subtask 1 the one with higher reported results. The data and results are available at `https://codalab.lisn.upsaclay.fr/competitions/22012`.
Keywords: DIMEMEX, hate speech detection, meme classification.

Resumen: En este documento presenta el resumen de la tarea DIMEMEX organizada como parte del IberLEF 2025 junto con la 41^{a} Conferencia Internacional de la Sociedad Española para el Procesamiento de Lenguaje Natural (SEPLN 2025). El principal objetivo de la tarea es promover la investigación en el desarrollo de soluciones automáticas para la detección de contenido inapropiado en memes enfocados principalmente al español mexicano. Se consideraron tres tareas: (i) Clasificación ternaria cuyo objetivo es determinar si un meme contiene contenido relacionado con discurso de odio, contenido inapropiado o ninguno; (ii) Clasificación de grano fino en la que el meme puede ser clasificado en una sub categoría de discurso de odio; y (iii) Clasificación ternaria, como en (i), restringiedo a los participantes a utilizar enfoques orientados modelos de lenguaje de gran tamaño. Se les proporcionó a los participantes un conjunto de datos multimodal anotado manualmente, que contiene tanto imágenes como texto relacionado con cada meme. Como resultado, 6 equipos de un total de 10 sistemas recibidos reportaron las descripciones de sus soluciones en la fase final. Los resultados muestran un desempeño competitivo para todas las subtareas, siendo los mejores aquiellos asociados a la tarea 1. El conjunto de datos y los resultados se encuentran disponibles en `https://codalab.lisn.upsaclay.fr/competitions/22012`.
Palabras clave: DIMEMEX, detección de discurso de odio, clasificación de memes.

ISSN 1135-5948 DOI 10.26342/2025-75-29

1 Introduction

Social media has become a powerful way of expression for people. Numerous platforms allow individuals to express their thoughts openly, sometimes under the veil of anonymity. Although 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 MultiOFF, consisting 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 in its previous DIMEMEX@IberLEF2024 edition (Jarquín-Vásquez et al., 2024). For this edition, an enhanced version is provided.

The DIMEMEX shared task comprises three subtasks: i) A three-way classification task to distinguish instances containing hate speech, inappropriate content, or neither; ii) A finer-grained classification task to categorize instances of hate speech into specific categories like classism, sexism, and racism; and iii) A three-way classification, as in i), solely focusing on the use of Large Language Models (LLMs). For task 1, a total of 10 teams submitted results. Additionally, 3 of them also participated in Subtask 2. For task 3, 2 out of the 3 participating teams also submitted solutions for tasks 1 and 2. The evaluation exhibited a dominant presence of Transformer-based approaches and a bias towards working with the text modality. Despite the competitive results obtained in the evaluation, it is still notable the difficulty of all 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 provided dataset and related details of the subtasks. Section 3 provides a summary of the proposed approaches by the participating teams as well as the reported results. Finally, our conclusions and future work are presented in

Section 4.

2 Task description

2.1 Dataset

The DIMEMEX 2025 dataset is a curated[1] version of the one released for the previous edition DIMEMEX@IberLEF2024 (Jarquín-Vásquez et al., 2024). The dataset consists of a set around 3,000 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 work is focused on the identification of inappropriate content, various types of hate speech, and non-abusive content (indicated by the '*neither*' label); regarding all the categories as mutually exclusive to maintain clear distinctions within our analysis. A meme is considered as *inappropriate content* if it exhibits any kind of manifestation of offensive, vulgar (profane, obscene, sexually charged), and/or morbid humor content. Alternatively, if the meme 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), it is classified as *hate speech*. Regarding types of hate speech, the following categories and definitions are considered:

- *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, misandry, and hateful content related to LGBTQ+.
- *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.

[1] In this context, *curated* means that the quality of the text extracted from the memes was improved.

Manual Annotation

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 without any annotation guideline reference. We asked them to follow their own definition of the considered phenomena. After analyzing the inter-annotator agreement rates and feedback discussions, annotation guidelines were defined and used for the final labeling stage where the annotation team comprised 12 annotators (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. Then, the distribution of annotators across these partitions was meticulously planned to achieve as equitable a balance as possible between male and female annotators. The final label for each meme was determined by a majority vote and the instances where the agreement between annotators was unclear underwent to a final re-labeling stage where all annotators convened and voted to decide on the final label. 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**
Hate speech	0.4343
Inappropriate content	0.4552
Neither	0.3830
Classism	0.4301
Racism	0.5148
Sexism	0.4764
Other	0.3445

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

Instances Examples

For each meme in the DIMEMEX dataset we provide: (1) the ID of the meme associated to its corresponding image; (2) the extracted and curated text from the meme, which was obtained using GPT-4o and carefully crafted

Category: Neither	**Category:** Hate Speech	**Category:** Inappropriate Content
		
Translated OCR text: Me seeing it's 1:59 AM /////And suddenly it changes to 3:00 AM	**Translated OCR text:** Scissors for women	**Translated OCR text:** Wody, that's what condoms are for to prevent idiots like this one from being born
Translated Image Caption: The image is a meme that consists of two panels taken from a scene of a television show. The character in the first panel has a relaxed and smiling expression, while in the second panel their face shows extreme surprise, with wide-open eyes.	**Translated Image Caption:** The image shows a package of kitchen scissors from the brand "EKCO". In the top right corner of the packaging, the text "For us, the women" appears. The packaging design includes a photo of a red bell pepper on a decorative plate, accompanied by a brief description of the scissors: "Scissors, multifunctional for the kitchen, good, strong, and resistant".	**Translated Image Caption:** The image shows a scene featuring the characters Woody and Buzz Lightyear from the movie Toy Story, placed in a humorous context. Buzz Lightyear is pointing at something off-screen with a serious expression, while Woody looks at him with a mix of confusion and resignation.

Table 2: Samples of memes from the DIMEMEX 2025 dataset.

prompts to ensure high-quality text extraction; and (3) a textual description of what is depicted in the image using a state-of-the-art language and vision model. Table 2 depicts some sample memes from our provided DIMEMEX dataset.

2.2 Subtasks

As previously mentioned DIMEMEX 2025 encompasses three subtasks: i) Three-way classification, which involves distinguishing instances containing hate speech, inappropriate content, or neither; ii) Finer-grained classification, which involves categorizing instances of hate speech into specific categories such as classism, sexism, racism, and others; and iii) Detecting the same three categories as the i) subtask, specifically using LLMs for the solution.

All 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. During this phase, they were able to submit their predictions for the validation set through the CodaLab website, receiving instant feedback on the performance of their submission.
- **Final phase.** Unlabeled test data was available to participants. They could submit up to ten submissions during the contest. Then, teams 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 all subtasks. Being the latter score, the leading evaluation measure in all subtasks.

2.3 Baselines

As baseline methods, we selected four well-established approaches recognized for their strong performance in tasks involving image and text classification. The first three baselines were adapted for evaluation on the first two subtasks, while the fourth baseline was specifically designed for subtask 3. The first baseline relies exclusively on the text modality and involves fine-tuning the pre-trained BETO model[2]. Complementing the text-based approach, the second baseline relies solely on the visual modality and entails fine-tuning the pre-trained Vision Transformer (ViT) model[3].

The third baseline integrates both visual and textual modalities through an early fusion approach, which concatenates the classification vectors obtained from the BETO and ViT models. Finally, the fourth baseline adopts a zero-shot (ZS) in-context learning approach, leveraging the Llama-3.2-11B-Vision-Instruct model[4]. In this setting, the model receives a prompt containing the meme image to be classified along with the extracted text from the image, which is provided as additional context. For the sake of readability, the results obtained from these baselines are referred to as Baseline (TXT), Baseline (IMG), Baseline (TXT + IMG), and Baseline (ZS-LLM). The proposed baselines, along with the development kit provided to participants, are available in the following repository[5].

[2] https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased

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

[4] https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct

[5] https://github.com/MasterHoracio/DIMEMEX-2025-Baselines

Team	Pre-processing		Pre-trained models			Fusion	Classifier	Extra Details
	Text	Image	Text	Image	LLM			
UC-UCO-CICESE (Martínez-López et al., 2025)			BERT BETO	ViT	Qwen-7B Nous-Hermes Mistral Tiiulo	concat		
HARGP-BETO (Jin and Zhou, 2025)			BETO			gated unit		Local and global attention plus a hierarchical approach
MICHAEL (Ibrahim, 2025)			ModernBERT					Hierarchical multi-task framework
UAEMemex (Neri-Mendoza et al., 2025)	x		BERT				LR KNN	ASCII vectorization
ITC (Cabada et al., 2025)	x		BERT BETO RoBERTa					Data augmentation with generative AI and class balancing
INFOTEC+CentroGeo (Moctezuma et al., 2025)	x	x	EvoMSA	CLIP ViT-B/32			SVM RF	
Baseline (TXT)	x		BETO					
Baseline (IMG)		x		ViT				
Baseline (TXT + IMG)	x	x	BETO	ViT		concat		
Baseline (ZS-LLM)	x	x			Llama-3.2 11B-Vision Instruct			

Table 3: Summary of participant systems' descriptions were pre-processing for text and image modalities are indicated with an **x**, used pre-trained models and classifiers (LR-Logistic Regression, KNN- K-Nearest Neighbors, SVM-Support Vector Machine, RF-Random Forest) are listed, fusion methods are specified (**concat**enation), and extra details of the approaches are described.

3 Overview of Participating Systems

The following subsections present a summary of the main approaches adopted by the participating systems, followed by a comprehensive evaluation of their respective results.

3.1 Systems' Descriptions

A total of 10 teams submitted results for Subtask 1. Additionally, 3 teams participated in Subtask 2, all of whom also submitted results for Subtask 1. Another 3 teams submitted results for Subtask 3, with 2 of them also participating in Subtask 1. However, only 6 teams provided detailed descriptions of their proposed systems. Table 3 presents a summary of the solutions submitted by the participating teams.

Team UC-UCO-CICESE participated in all subtasks. For subtasks 1 and 2 the team considered text-only and text-image with BERT, BETO+ViT models as their submitted solutions, and for task 3 they used DeepSeek-R1-Distill-Qwen 7B after trying a variety of LLMs.

Team HARGP-BETO adopted a gated unit approach to fuse local and global attention from the provided OCR and textual description where both embeddings were obtained by applying a BETO encoder. Then, the fused output was fed into a hierarchical mechanism to later train dense layers for classification.

Team MICHAEL proposed an adaptation of ModernBERT for hierarchical multi-task classification. The model was enhanced with rotary positional embeddings (RoPE) and flash attention allowing a better handling of extended context for text modality.

Team UAEMemex explored the use of lexical and semantic features from different text vectorization methods (One-hot-encoding, TF-IDF, Doc2Vec, BERT, and ASCII) considering pre-processing or not the OCR text. These representations were fed into a variety of machine learning standard models (Naive Bayes, Logistic Regression, KNN, MLP, and SVM). The team reported BERT vectorization on OCR texts fed into a Logistic Regression classifier as their final submission.

Team ITC adopted a BETO model classification approach after evaluating other transformer based models like BERT and RoBERTa. To enhance the model the team augmented the text data by generating descriptions with Humarin paraphrase generative AI model and balancing the number of instances for each one of the evaluated classes.

Team INFOTEC+CentroGeo applied a variety of standard classifiers (SVM, Random Forest) fed with representations obtained from pre-trained models like CLIP, and EvoMSA by using text, image, and the combination of both modalities to enhance the classification. As their final submission CLIP embeddings for images only plus SVM was reported.

3.2 Evaluation campaign results

This section reports the performance achieved by the participating teams across the three proposed subtasks. Table 4 presents the results for Subtask 1, which involves distinguishing memes containing hate speech, inappropriate content, or neither. Teams are ordered in descending macro-averaged F_1 score, and we also provide the corresponding macro-averaged Precision and Recall to enable a comprehensive interpretation of the findings. Rows shaded in gray denote the teams that submitted working notes describing their systems. The table additionally includes the scores of our first three baseline approaches; for each baseline we report the median over three independent runs, as it offers a more reliable estimation of their performance.

Subtask 1:			
Team	**Precision**	**Recall**	**F1-Score**
HARGP-BETO	**0.58**	**0.58**	**0.58**
Onarion	0.58	0.56	0.57
UC-UCO-CICESE	0.57	0.55	0.55
ITC	0.54	0.51	0.52
Baseline (TXT + IMG)	0.52	0.49	0.50
Baseline (TXT)	0.47	0.50	0.48
Baseline (IMG)	0.47	0.44	0.45
MICHAEL	0.46	0.43	0.44
UAEMemex	0.45	0.42	0.43
INFOTEC+CentroGeo	0.42	0.42	0.42
VeronicaNeriMendoza	0.42	0.42	0.42
csuazob	0.42	0.36	0.34
AngelBaron	0.35	0.34	0.33

Table 4: Results of the participant teams in Subtasks 1. Bold numbers correspond to the best results of each metric.

The HARGP-BETO team achieved the best results in Subtask 1 (Jin and Zhou, 2025), followed by the Onarion and UC-UCO-CICESE (Martínez-López et al., 2025) teams. The differences among the top three systems were minimal: the gap between first and second place was only 0.01 macro-averaged F_1 points, while the gap between second and third place was 0.02. All three leading approaches adopted multimodal fusion strategies that combine features extracted with state-of-the-art pre-trained Transformer models. Specifically, HARGP-BETO employed a gated fusion mechanism that models the interaction between the OCR text and the image captions, which, as reported, contributed to its strong performance. In contrast, the UC-UCO-CICESE team employed an early fusion approach that combined features extracted from both text and image inputs. Overall, the integration of both modalities highlights the benefit of exploiting the complementary information offered by textual and visual modalities, as well as by multiple information sources (e.g., meme text and image captions), for detecting hate speech and inappropriate content in memes.

To further analyze the performance differences among the participating systems, we conducted a statistical analysis using the tool proposed by (Nava-Muñoz, Graff-Guerrero, and Escalante, 2023). Figure 1 presents a comparison between the top-ranked team and the remaining participants, based on 95% confidence intervals of the macro-average F_1 scores. For this purpose, 1000 bootstrap samples were generated (further details can be found in (Nava-Muñoz, Graff-Guerrero, and Escalante, 2023)). As shown in the figure, the differences in macro-average F_1 scores among the top three teams are not statistically significant. These findings are consistent with the results reported in Table 4, where the score gap between the first and third place is only 0.03.

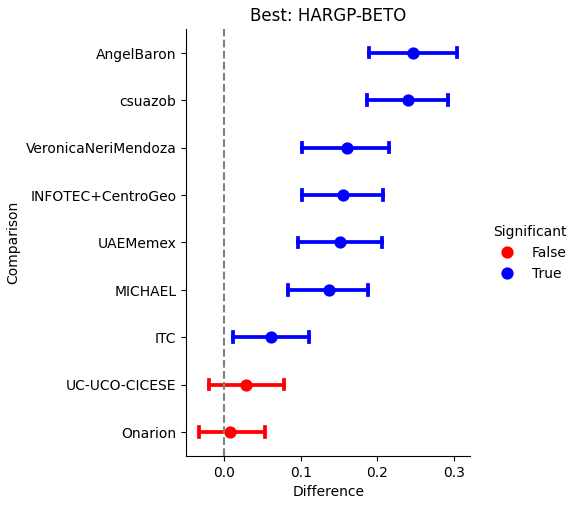

Figure 1: 95% confidence intervals for subtask 1 using bootstrapping.

The results obtained in Subtask 2, the finer-grained classification task involving the categorization of instances of hate speech into specific categories, are presented in Table 5. The best performance in this subtask was achieved by the UC-UCO-CICESE team, followed by the csuazob and MICHAEL teams. Among the proposed approaches, the UC-UCO-CICESE team employed a multimodal

strategy combining both text and image modalities, whereas the MICHAEL team proposed approach relied exclusively on textual information. The results obtained by these teams highlight the advantage of using multimodal inputs, a trend further corroborated by the proposed baselines, where the best-performing baseline used both modalities.

Overall, the scores for this subtask were relatively low, with the highest macro-average F_1 score reaching only 0.37, and the lowest 0.26. These outcomes underscore the inherent difficulty of accurately identifying different forms of hate speech and highlight the need for developing multimodal models capable of accurately identifying nuanced instances of hate speech.

Subtask 2:			
Team	**Precision**	**Recall**	**F1-Score**
UC-UCO-CICESE	**0.40**	**0.36**	**0.37**
Baseline (TXT + IMG)	0.40	0.35	0.36
Baseline (TXT)	0.30	0.34	0.29
csuazob	0.29	0.26	0.27
Baseline (IMG)	0.26	0.28	0.27
MICHAEL	0.35	0.25	0.26

Table 5: Results of the participant teams in Subtask 2. Results in bold correspond to the best results of each measure.

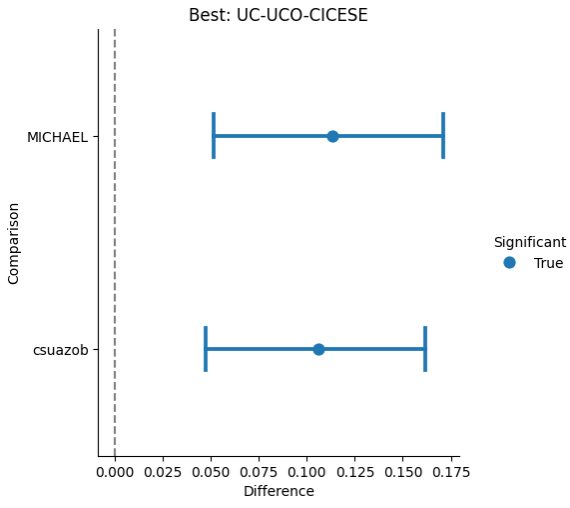

Figure 2: 95% confidence intervals for subtask 2 using bootstrapping.

Figure 2 displays the confidence intervals for Subtask 2, comparing the top-ranked team against the remaining participants. In contrast to Figure 1, where no statistically significant differences were observed between the first, second, and third places, this figure reveals a clear and statistically significant difference between the first place and both the second and third places. This observation is consistent with the larger performance gap seen in Subtask 2, where the macro-average F_1 score difference between the first and second place reached 0.10. Moreover, no significant difference is observed between the second and third places, which is in line with their minimal performance gap of just 0.01 in the macro-average F_1 score.

The results obtained by the participating teams in Subtask 3 are presented in Table 6. The highest-performing team was LuisArellano, followed by the UC-UCO-CICESE and csuazob teams. As observed, all participating teams outperformed our proposed baseline, which relied on a zero-shot approach leveraging both textual and visual information. The performance gap is notable: when comparing the baseline with the lowest-performing team, a difference of 0.17 in macro-average F_1 score is observed. This disparity may be due to the use of more advanced reasoning models based on distillation techniques, as reported by the UC-UCO-CICESE team (Martínez-López et al., 2025).

In contrast to the results from Subtask 2, the differences between the top-ranked team and the second and third teams in Subtask 3 were smaller, with gaps of 0.03 and 0.05 in macro-average F_1 score, respectively. Moreover, when comparing the best results from Subtasks 1 and 3, the difference in macro-average F_1 score was only 0.04, with Subtask 1 achieving the highest score. These findings highlight the adaptability of LLMs to diverse tasks, including the detection of hate speech and inappropriate content in Mexican Spanish memes.

Subtask 3:			
Team	**Precision**	**Recall**	**F1-Score**
LuisArellano	**0.63**	**0.55**	**0.54**
UC-UCO-CICESE	0.54	0.50	0.51
csuazob	0.50	0.50	0.49
Baseline (ZS-LLM)	0.53	0.39	0.32

Table 6: Results of the participant teams in Subtask 3. Results in bold correspond to the best results of each measure.

Figure 3 displays the confidence intervals for Subtask 3, comparing the top-ranked team with the remaining participants. In contrast to Figure 2, where statistically significant differences were observed between the first place and both the second and third places, this figure shows that no significant

differences exist among the top three teams. This result is consistent with the smaller performance gap observed in Subtask 3, where the difference between the first and third place was only 0.05 in the macro-average F_1 score.

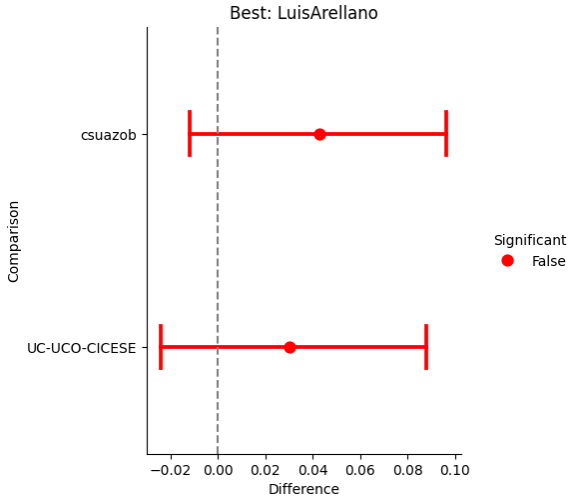

Figure 3: 95% confidence intervals for subtask 3 using bootstrapping.

3.3 Analysis

To perform a more comprehensive analysis of the participants' results, we began by examining the complementarity and diversity present in their predictions. To quantify complementarity, we employed the *Maximum Possible Accuracy (MPA)* metric, defined as the ratio of correctly classified instances to the total number of test instances. An instance is considered correctly classified if at least one of the participating teams predicted it correctly. As a reference point for assessing this complementarity, we report the *Best Performance Accuracy (BPA)*, extracted from the top-performing team in each subtask. To measure diversity, we used the *Coincident Failure Diversity (CFD)* metric (Tang, Suganthan, and Yao, 2006), which captures the error diversity among participants' predictions. This metric ranges from 0, indicating that all teams make the same prediction (either correct or incorrect), to 1, representing completely distinct misclassifications across teams.

The results of applying the MPA and CFD metrics to evaluate the performance of all participating teams across the three subtasks for identifying hate speech and inappropriate content are presented in Table 7. The table reports the results for both the overall classification and per-class analysis.

ST	Class	BPA	MPA	CFD	#T
1	all classes	0.661	0.946	0.698	10
1	hate	0.500	0.848	0.327	10
1	inappropriate	0.459	0.866	0.343	10
1	neither	0.773	1.000	0.431	10
2	all classes	0.641	0.793	0.322	3
2	sexism	0.311	0.466	0.148	3
2	racism	0.333	0.575	0.129	3
2	classism	0.231	0.385	0.117	3
2	other	0.000	0.103	0.133	3
3	all classes	0.668	0.851	0.285	3
3	hate	0.187	0.544	0.199	3
3	inappropriate	0.659	0.785	0.125	3
3	neither	0.805	0.957	0.128	3

Table 7: Comparison of BPA, MPA, and CFD results between the different general approaches. The first column refers to the subtask number while the last one refers to the number of systems' results involved in the calculation.

Regarding the results of Subtask 1 (rows 2-5), the MPA values obtained by all teams are remarkably higher than the BPA achieved by the HARGP-BETO team. This suggests that the systems and approaches employed by all participants exhibit a high degree of complementarity. In this subtask, the class *neither* was the easiest to identify, with an MPA of 1, while *hate speech* was the most challenging. In contrast, the results for Subtask 2 (rows 6-10) show a smaller increase in MPA values compared to the BPA of the UC-UCO-CICESE team. This may be attributed to the smaller number of participants in this subtask and the inherent difficulty in distinguishing between different categories of hate speech. Based on this analysis, the most difficult classes to distinguish in Subtask 2 were *classism* and *other*.

Finally, the results for Subtask 3 (rows 11-14) reveal a greater increase in MPA values compared to Subtask 2, indicating a higher degree of complementarity among the systems proposed by participants. As in Subtask 1, the class *neither* was the easiest to identify, while *hate speech* remained the most difficult. Additionally, Table 7 also presents the CFD values, which reflect the diversity of prediction errors across teams. Higher diversity was observed in Subtask 1 compared to Subtasks 2 and 3 (rows 2–5 vs. rows 6–14), which is consistent with the improvements noted in the MPA metric across the

subtasks.

Qualitative Analysis

NOTE: This subsection includes examples that may be considered offensive by some readers. These examples are presented solely for research purposes and do not reflect the views or opinions of the authors.

To further analyze the outcomes of the participating systems, we leveraged the results obtained from measuring the complementarity of the proposed approaches using the MPA metric. This analysis aimed to identify those instances that were consistently misclassified by all systems. A manual qualitative analysis was then conducted on a subset of these memes. In the following paragraphs, we briefly describe the main observed features, along with representative examples of such instances.

Category: Hate	**Category: Hate**
	
Translation: I don't know much about birds, but I already know who the wife is...	Translation: When you drink the Coca-Cola too fast from the glass.
Category: Inappropriate	**Category: Inappropriate**
	
Translation: The glasses here mean that... #THEY-CAN-SUCK-IT.	Translation: Did you know... Ray Charles and Stevie Wonder couldn't even stand the sight of each other!.

Table 8: Samples of memes that were incorrectly classified by all participating teams in Subtask 1 and Subtask 3.

Table 8 presents examples of memes that were misclassified by all participating teams in Subtasks 1 and 3. A common factor among many of these memes is the need for extra-linguistic context to enable correct interpretation. For example, the fourth meme in the lower right corner, which was labeled as inappropriate content, requires the additional knowledge that the individuals depicted are visually impaired, in order to understand the dark humor conveyed by the meme. For this instance, most teams predicted the class as *neither*.

Another essential factor for accurately detecting inappropriate content and hate speech in memes is the correct interpretation of both textual and visual modalities. This is illustrated by the second meme in the upper right corner, where the textual caption *"When you drink the Coca-Cola too fast from the glass"* appears harmless on its own. However, when combined with the visual content, it conveys a discriminatory message based on skin color. In this case, the majority of the teams also misclassified the instance as *neither*.

Category: Classism	**Category: Racism**
	
Translation: Knows the sea — Doesn't know the sea.	Translation: F**k it life goes on.
Category: Sexism	**Category: Other**
	
Translation: When you get back from work and dinner isn't served xD.	Translation: Daddy, what do Jehovah's Witnesses believe in? — They believe we're going to open the door for them.

Table 9: Samples of memes that were incorrectly classified by all participating teams in Subtask 2.

Finally, Table 9 presents examples of memes that were misclassified by all participating teams in Subtask 2. Once again, the necessity of extra-linguistic context is evident for correctly classifying the different types of hate speech in memes. For instance, the second meme in the upper right corner, labeled as *racism*, adds the suffix "*tl*" to certain words—a linguistic marker commonly used to mock speakers of Indigenous languages in

Mexico, given its frequent use in Nahuatl.

In addition, the correct interpretation of both textual and visual components is crucial, as illustrated by the first meme in the upper left corner. In this example, the placement of the captions "Knows the sea" and "Doesn't know the sea" conveys a classist message based on skin color. 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 the overview of the second edition of 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. This shared task included three subtasks: Subtask 1 involved a three-way classification to distinguish between hate speech, inappropriate content, or neither; Subtask 2 required a finer-grained classification of hate speech into specific categories such as classism, sexism, racism, and others; and Subtask 3 constrained participants to use only LLMs for the same three-way classification as in Subtask 1. 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; most of the approaches utilized text as the preferred modality to work with, but there were also efforts to incorporate text and image modalities. Notably, Transformer-based approaches exhibited a dominant presence and outperformed traditional machine learning methods; the participating teams employed a variety of pre-trained language models and vision models. Some teams involved multimodal models like CLIP and a variety of LLMs including Qwen-7B, Nous-Hermes, Mistral, and Tiiulo. Regarding the introduced innovations by the teams, data augmentation with generative AI, and the use of local and global attention, plus a hierarchical approach, were notable.

The evaluation results highlight the overall difficulty of the three proposed subtasks, indicating substantial room for improvement through the development of more sophisticated approaches. A noteworthy aspect was the integration of LLMs by participants in Subtask 3. Although their performance was slightly lower than the best-performing approaches in Subtask 1, the adaptability of LLMs demonstrated competitive potential for the detection of hate speech and inappropriate content. As anticipated, the fine-grained classification required in Subtask 2 presented greater difficulty compared to the three-way classification addressed in Subtasks 1 and 3, 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 considers extending 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 scholarship 972915. This work was also partially supported by "HARMONIA" project - M4-C2, I1.3 Partenariati Estesi - Cascade Call - FAIR - CUP C63C22000770006 - PE PE0000013 under the NextGenerationEU programme.

References

Cabada, R. Z., M. L. B. Estrada, V. M. B. Beltrán, A. G. Robles, and N. L. López. 2025. ITC's participation at dimemex: Data augmentation using generative ai for better detection of hate speech in mexican memes. In J. Á. González-Barba,

L. Chiruzzo, and S. M. Jiménez-Zafra, editors, *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS. org*, CEUR Workshop Proceedings. 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. Warmsley, 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.

Gelber, K. and L. McNamara. 2015. Evidencing the harms of hate speech. *Social Identities*, 22:1–18, 12.

González-Barba, J. Á., L. Chiruzzo, and S. M. Jiménez-Zafra. 2025. Overview of IberLEF 2025: Natural Language Processing Challenges for Spanish and other Iberian Languages. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS. org*.

Ibrahim, M. 2025. Meme classification using modernbert. In J. Á. González-Barba, L. Chiruzzo, and S. M. Jiménez-Zafra, editors, *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS. org*, CEUR Workshop Proceedings. CEUR-WS.org.

Jarquín-Vásquez, H., I. Tlelo-Coyotecatl, M. Casavantes, D. I. Hernández-Farías, H. J. Escalante, and L. V.-P. y Manuel Montes-y Gómez. 2024. Overview of dimemex at iberlef 2024: Detection of inappropriate memes from mexico. *Procesamiento del Lenguaje Natural*, 73(0):335–345.

Jin, Q. and X. Zhou. 2025. HARGP-BETO: Hierarchical text interactions model for abuse detection in mexican spanish memes. In J. Á. González-Barba, L. Chiruzzo, and S. M. Jiménez-Zafra, editors, *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS. org*, CEUR Workshop Proceedings. CEUR-WS.org.

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.

Martínez-López, Y., Y. Jauriga, M. G. Saborit, J. Madera, A. Rodríguez-González, C. de Castro Lozano, and J. M. R. Uceda. 2025. UC-UCO-CICESE_UT3-Plenitas team at exploring the detection of inappropriate memes from mexico using deeplearning. In J. Á. González-Barba, L. Chiruzzo, and S. M. Jiménez-Zafra, editors, *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS. org*, CEUR Workshop Proceedings. CEUR-WS.org.

Moctezuma, D., T. Ramirez-delreal, E. Tellez, M. Graff, and G. Ruiz. 2025. DIMEMEX2025: Solution based on open-clip of the infotec+centrogeo team. In J. Á. González-Barba, L. Chiruzzo, and

S. M. Jiménez-Zafra, editors, *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS. org*, CEUR Workshop Proceedings. 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.

Neri-Mendoza, V., J. Rojas-Simon, Y. Ledeneva, Y. A. Santos-Bobadilla, A. A. Gil-García, Á. Baron-García, and R. A. Garcia-Hernández. 2025. UAEMemex participation at dimemex 2025: Exploring lexical and semantic information to detect hate, inappropriate, and harmless memes. In J. Á. González-Barba, L. Chiruzzo, and S. M. Jiménez-Zafra, editors, *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS. org*, CEUR Workshop Proceedings. 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.