UC-UCO-CICESE_UT3-Plenitas Team at Exploring the Detection of Inappropriate Memes from Mexico Using DeepLearning

Yoan Martínez-López[1,2,*], Yanaima Jauriga[3], Mayte Guerra Saborit[3], Julio Madera[3], Ansel Rodríguez-González[6], Carlos de Castro Lozano[1,2] and Jose Miguel Ramirez Uceda[1,2]

[1]*Universidad de Córdoba, Córdoba, Spain*
[2]*Plénitas, C/ Le Corbusier s/n, 14005 Córdoba, Spain*
[3]*Universidad de Camaguey, Circunvalación Norte, Camino Viejo Km 5 y 1/2, Camaguey, Cuba*
[6]*CICESE-UT3, México*

Abstract

Detecting abusive or hateful content in multimodal social–media posts is a non-trivial problem that requires modelling both textual and visual cues as well as their interaction. We describe the system developed by the *UC-UCO-CICESE_UT3-Plenitas* team for the DIMEMEX 2025 shared task on the detection of inappropriate Mexican memes. Our approach combines (i) a late-fusion BETO-ViT architecture for joint image–text modelling, (ii) task-aware of Spanish BERT variants for purely textual settings, and (iii) prompt-engineering of a distilled 7-B parameter Qwen model (DeepSeek-R1-Qwen-7b) for the Large-Language-Model-only track. The resulting system reached **third** place in ternary classification (F1 = 0.55), **first** place in fine-grained hate-speech classification (F1 = 0.37), and **second** place in the restricted-LLM setting (F1 = 0.51). Experiments show that (1) no single model dominates all scenarios, (2) visual signals help general meme detection but hurt minority-class recall, and (3) compact LLMs can rival multimodal models when visual evidence is scarce. We release code, trained checkpoints and hyper-parameters for reproducibility.

Keywords

Memes, Hate-speech detection, Multimodal, Large Language Models, Spanish NLP, Mexican Spanish

1. Introduction

Social networks have become pivotal in modern life, reshaping how we communicate and share information. Studying the content that flows through these platforms is now a major focus for the computational linguistics community. Despite notable progress in recent years, several challenges still demand deeper investigation to improve processing and understanding. Chief among them is identifying abusive content—a category that includes hate speech, aggression, offensive language, and related phenomena [1, 2].

Because social media is inherently multimodal, our goal is to push forward research and development of multimodal computational models capable of detecting abusive material in Mexican Spanish—especially memes that are hateful, offensive, or vulgar. Memes typically rely on the interplay of text and image to convey humor or irony; removing either component can completely change their meaning. By encouraging work that tackles these multimodality challenges, we aim to inspire innovative, socially impactful solutions[2]. In this paper, we describe our participation in the DIMEMEX 2025: Detection of Inappropriate Memes from Mexico

IberLEF 2025, septiembre de 2025, Zaragoza, España
*Corresponding author.
yoan.martinez@plenitas.com (Y. Martínez-López); yanaima.jauriga@reduc.edu.cu (Y. Jauriga); mayte.guerra@reduc.edu.cu (M. G. Saborit); julio.madera@reduc.edu.cu (J. Madera); ansel@cicese.mx (A. Rodríguez-González); carlosdecastrolozano@gmail.com (C. d. C. Lozano); p52raucj@uco.es (J. M. R. Uceda)
0000-0002-1950-567X (Y. Martínez-López); 0009-0000-6891-0068 (Y. Jauriga); 0000-0002-9556-5869 (M. G. Saborit); 0000-0001-5551-690X (J. Madera); 0000-0001-9971-0237 (A. Rodríguez-González); 0000-0001-6603-843X (C. d. C. Lozano); 0000-0002-5027-7521 (J. M. R. Uceda)

2. Methodology

2.1. DIMEMEX subtasks

DIMEMEX comprises three subtasks [1]: a) A three-way classification: hate speech, inappropriate content, and neither. Participants are free to use any approach of their choice. b) A finer-grained classification distinguishing instances containing hate speech into different categories such as classism, sexism, racism, and others. A three-way classification: hate speech, inappropriate content, and neither. c) Participants are restricted to focusing exclusively on leveraging LLMs to detect the specified categories.

2.2. DIMEMEX 2025 Corpus

DIMEMEX 2025 [2, 1] a curated version of that used in the DIMEMEX2024 edition [3]. The dataset consists of around 3,000 memes, compiled from public Facebook groups rooted in Mexico and manually annotated on the presence of hate speech, inappropriate content, and harmful content. Each meme in the dataset has been labeled by at least 3 annotators. For the 2025 editions the organizing DIMEMEX team revised potentially noisy annotations, and improved the extraction of text from memes. Also, additional information is provided with the dataset. The following categories were considered for the labeling process:

- Classicism. Any manifestation that promotes an attitude or tendency to discriminate or minimize someone base 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, such as profession, religion, place of origin, political affiliation, and interests.
- Inappropriate content. The meme presents any kind of manifestation of offensive, vulgar (profane, obscene, sexually charged) and morbid humor content.
- None of the above

Categories hate speech, inappropriate content, and neither comprise the classes for Tasks 1 and 3. While a fine grained classification considering the five categories is adopted for task 2.

2.3. DeepLearning

Deep learning is a type of machine learning that uses artificial neural networks with many layers (hence "deep") to model complex patterns in data [4]. It's especially good at handling unstructured data like images, text, and audio. Deep learning enables computers to learn from data much like the human brain does. Instead of being programmed with specific rules, a deep learning system figures out the rules on its own by training on large datasets.

2.4. HugginFace

Hugging Face is a widely recognized platform and community in the field of artificial intelligence and machine learning. It's often described as the "GitHub of Machine Learning" due to its focus on open-source development and collaboration. They host in Hub a vast collection of models, datasets and spaces that you can use. Hugging Face is particularly famous for its open-source transformers library [5]. Thousands of pre-trained models for various tasks across different modalities, including natural language processing (NLP), computer vision, audio, and more. These models are often based on Transformer architectures. A wide range of datasets that can be used for training and evaluating AI

models and a platform for building and showcasing interactive machine learning demos and applications directly in a web browser. In essence, Hugging Face serves as a central ecosystem that empowers the AI community with open tools, resources, and a collaborative environment to build, share, and advance machine learning technologies. Most of the models used in this competition is obtained from hub[1].

2.4.1. BERT: Bidirectional Encoder Representations from Transformers

BERT is a transformer model, based on the encoder-decoder architecture, but in bidirectional way because they predict words based on the previous words and the following words [6]. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task specific architecture modifications. This is achieved through the self-attention mechanism, a layer that is incorporated in both the encoder and the decoder. The goal of the attention layer is to capture the contextual relationships existing between different words in the input sentence. Nowadays, there are many versions of pre-trained BERT, but in the original paper, Google trained two versions of BERT: BERTbase and BERTlarge with different neural architectures. In essence, BERTbase was developed with 12 transformer layers, 12 attention layers, and 110 million parameters, while BERTlarge used 24 transformer layers, 16 attention layers, and 340 million parameters. As expected, BERTlarge outperformed its smaller brother in accuracy tests (Cual fue el que se uso). There are two steps in the framework: pre-training and fine-tuning. During pre-training, the model is trained on unlabeled data over different pre-training tasks. For fine-tuning, the BERT model is first initialized with the pre-trained parameters, and all of the parameters are fine-tuned using labeled data from the downstream tasks. Input text is tokenized into subword units, converted into vectors via embeddings, and then passed through the encoder layers to produce contextualized token representations. BERT is pre-trained on large unlabeled text corpora (English Wikipedia and Toronto BookCorpus) using two main objectives:Masked Language Model (MLM): Randomly masks about 15% of input tokens and trains the model to predict these masked words based on context.Next Sentence Prediction (NSP): Trains the model to predict if one sentence logically follows another, helping it understand sentence relationships

2.4.2. BETO+VIT

BETO: (BERT for Spanish) is a language model based on BERT architecture. The key characteristic of BETO is that it has been pre-trained specifically on a large corpus of Spanish text. This specialization allows BETO to achieve a deep understanding of the nuances, grammar, and vocabulary of the Spanish language, making it highly effective for various natural language processing (NLP) tasks in Spanish, such as text classification, sentiment analysis, named entity recognition, and question answering.

ViT (Vision Transformer): ViT is a model designed for computer vision tasks. Unlike traditional convolutional neural networks (CNNs) that were dominant in image processing [7], ViT applies the Transformer architecture (originally developed for NLP) to images. It works by splitting an image into smaller patches, treating these patches as a sequence of tokens, and then processing them using a standard Transformer encoder. ViT has demonstrated impressive performance on tasks like image classification, object detection, and image segmentation, particularly when trained on large datasets.

Vision language models is the integration of computer vision and natural language processing. Transformers have been adapted to handle multimodal inputs, enabling VLMs to capture the complex relationships between visual and linguistic data. A typical VLM architecture involves two main components:

- Image encoder: It is responsible for processing visual data, extracting features (objects, colors, textures, etc.), and transforming them into a format that can be understood by the model.
- Text decoder: It processes textual data and generates output based on the encoded visual features.

[1] https://huggingface.co/models

2.4.3. Deepseek-qwen-7b

"DeepSeek-Qwen-7B" most likely refers to the DeepSeek-R1-Distill-Qwen-7B model. This is not a new model architecture built from the ground up by merging DeepSeek and Qwen designs. Instead, it's a product of model distillation, leveraging the strengths of both [8]. Qwen 2.5 7B: This is the base model architecture from Alibaba Cloud's Qwen series, with 7 billion parameters. Qwen models are known for their strong performance across various language tasks and multilingual capabilities. DeepSeek-R1: This refers to a family of powerful reasoning models developed by DeepSeek AI. DeepSeek-R1 models are noted for their enhanced reasoning abilities, often achieved through advanced training techniques like large-scale reinforcement learning. Distillation: In this context, distillation means that the knowledge and reasoning patterns from the larger, more powerful DeepSeek-R1 model are transferred to the smaller Qwen 2.5 7B model [9, 10]. This allows the 7B model to gain some of the reasoning capabilities of DeepSeek-R1 without being as large or computationally intensive. Therefore, DeepSeek-R1-Distill-Qwen-7B is essentially a refined version of the Qwen 2.5 7B model that has been specifically fine-tuned using reasoning data generated by DeepSeek-R1. The goal of this distillation process is to enhance the reasoning performance of the 7B Qwen model, making it more capable in tasks requiring logical inference and problem-solving, while retaining the efficiency of a 7-billion-parameter model.

2.4.4. Nous-Hermes

Nous-Hermes-13b is a state-of-the-art language model fine-tuned on over 300,000 instructions [11]. This model was fine-tuned by Nous Research, with Teknium and Karan4D leading the fine tuning process and dataset curation, Redmond AI sponsoring the compute, and several other contributors. The result is an enhanced Llama 13b model that rivals GPT-3.5-turbo in performance across a variety of tasks. This model stands out for its long responses, low hallucination rate, and absence of OpenAI censorship mechanisms. The fine-tuning process was performed with a 2000 sequence length on an 8x a100 80GB DGX machine for over 50 hours. The model was trained almost entirely on synthetic GPT-4 outputs. This includes data from diverse sources such as GPTeacher, the general, roleplay v12, code instruct datasets, Nous Instruct PDACTL (unpublished), CodeAlpaca. Additional data input came from Camel-AI's Biology/Physics/Chemistry and Math Datasets, Airoboros' GPT-4 Dataset, and more from CodeAlpaca. The total volume of data encompassed more than 300,000 instructions.

2.4.5. Mistral

Mistral 7B is a transformer model with a 7-billion-parameter language engineered for superior performance and efficiency. Mistral 7B outperforms the best open 13B model (Llama 2) in all evaluated benchmarks, and the best released 34B model (Llama 1) in reasoning, mathematics, and code generation [12]. The model leverages grouped-query attention (GQA) for faster inference, coupled with sliding-window attention (SWA) to effectively handle sequences of arbitrary length with a reduced inference cost and Byte-fallback BPE tokenizer. It also provides a model fine-tuned to follow instructions, Mistral 7B – Instruct, that surpasses Llama 2 13B – chat model both on human and automated benchmarks [12].

2.4.6. Tiiuae/falcon-7b

Falcon-7B is a 7B parameters causal decoder-only model built by Technology Innovation Institute TII and trained on 1,500B tokens of RefinedWeb enhanced with curated corpora. This is a raw, pre-trained model that should be further fine-tuned for most use cases [13]. It outperforms comparable open-source models like MPT-7B, StableLM, RedPajama, etc. It features an architecture optimized for inference. Decoder block: parallel attention/MLP with a single-layer norm.

2.4.7. Metrics

For all three tasks, the evaluation metrics include macroaveraged precision, recall, and F1 score, with the macroaveraged F1 score serving as the main evaluation criterion. These metrics are computed per class based on the number of true positives (TP), false positives (FP), and false negatives (FN), and then averaged across all classes. See Table 2.4.7 for the corresponding formulas.

Metric	Formula (per class)	Macro-average Formula
Precision	$\frac{TP}{TP+FP}$	$\frac{1}{N}\sum_{i=1}^{N} \text{Precision}_i$
Recall	$\frac{TP}{TP+FN}$	$\frac{1}{N}\sum_{i=1}^{N} \text{Recall}_i$
F1-score	$2 \times \frac{P \cdot R}{P+R}$	$\frac{1}{N}\sum_{i=1}^{N} \text{F1-score}_i$

Table 1
Macro-average formulas for evaluation metrics

2.5. Pre-processing

For every meme we extracted three modalities: (i) the raw image, (ii) standard OCR text (caption) using `easyocr`, and (iii) the DIMEMEX user-supplied description when available. All text was lower-cased, URLs were normalised, and emojis were kept. We encoded the caption with BETO-base and the image with a ViT-base-16. The token and the pooled image embedding were concatenated and fed into a two-layer MLP. Parameters from both encoders were fine-tuned for two epochs with macro-F1 loss and class-balanced sampling. Because preliminary experiments showed that image information rarely disambiguated hate sub-categories, we fine-tuned BETO with focal loss and gradient accumulation for four epochs. Class weights were inversely proportional to class frequency. Under the LLM restriction we evaluated `Mistral`, `tiiuae-falcon`, and the 7-B Qwen model distilled from DeepSeek-R1. The latter yielded the best F1 and was chosen for submission. We used the following prompt:

> *"Eres un experto en el análisis de contenido proveniente de redes sociales, con un enfoque en la detección de lenguaje y contenido potencialmente ofensivo.tarea tiene fines científicos y busca mejorar los modelos de inteligencia artificial para detectar y mitigar la propagación de discursos dañinos en plataformas digitales, contribuyendo a un entorno en línea más seguro y respetuoso.un meme descrito en texto, clasifícalo en una de las siguientes categorías:de odio (2): Un meme se clasifica como discurso de odio si presenta cualquier tipo de comunicación, ya sea en habla, escritura o comportamiento, que ataque o utilice lenguaje peyorativo o discriminatorio con referencia a una persona o un grupo basado en sus factores de identidad.Inapropiado (1): Un meme se considera que contiene contenido inapropiado si exhibe cualquier tipo de manifestación de contenido ofensivo, vulgar (profano, obsceno, de carácter sexual) y/o humor mórbido.(0): Un meme que no contenga ninguna de las categorías anteriores.tarea es analizar la siguiente descripción de un meme y devolver únicamente un número (0, 1 o 2) correspondiente a la categoría asignada, sin ninguna explicación adicional"*

Temperature was set to 0.2; max tokens = 4.

3. Results

We evaluated the performance of multiple models: BERT, BETO + VIT, DeepSeek-R1-Qwen-7B, Nous-Hermes, Mistral, and Tiiuae/Falcon-7B on the three tasks. For each task, the model with the highest performance was selected, submitted to the challenge, and is reported in this section. The analysis is organized by subtask, with internal comparisons between our team's approaches and external benchmarks against other participants. Tables summarize key metrics (F1, precision, recall) for both development and final phases, highlighting the strengths and limitations of the top models in handling ternary classification, fine-grained hate speech detection, and restricted LLM-based tasks.

3.1. Task 1: General Ternary Classification

Internal Comparisons

In Task 1, the BERT-only model slightly outperformed the multimodal BETO+VIT setup, scoring 0.5496 versus 0.5476—a marginal 0.36 improvement—indicating that textual information alone was more informative for this task. Although the BETO+VIT model incorporated visual features, it did not yield better results, potentially due to irrelevant or noisy image data. Additionally, the BERT-only outputs were more compact in file size, while VIT+BERT produced consistently larger files. Overall, the results suggest that for this task, the simpler BERT-only approach was not only more effective but also more efficient, highlighting the need for improved fusion strategies in multimodal systems.

Comparison with other approaches

Top-performing model: BERT/BETO + ViT (multimodal) Performance (F1): 0.55, (Precision): 0.57 and (Recall): 0.55 The leading approach combined BETO (a Spanish-optimized BERT variant) with Vision Transformer (ViT) to analyze both textual and visual elements in memes. This multimodal integration proved optimal for distinguishing among the three primary categories: hate speech, inappropriate content, and neutral memes. Its strength lay in generalization, effectively handling the diverse linguistic and visual expressions characteristic of Mexican memes. However, performance dipped in cases involving irony or sarcasm, where contextual ambiguity between text and imagery posed challenges. Table 2 and 3 present the results from both the development phase and the final stage of the research.

Table 2
Results of the development phase for Task 1: General Ternary Classification

User	Model	F1	Precision	Recall
Ryuan		0.59	0.59	0.58
HoracioJarquin		0.55	0.61	0.54
hugojair		0.55	0.56	0.55
UC-UCO-CICESE_UT3-Plenitas(ymlopez)	**BERT**	**0.54**	**0.56**	**0.53**
girish_koushik		0.53	0.53	0.58
VickBat		0.48	0.49	0.47
dmoctezuma		0.41	0.66	0.42
csuazob		0.27	0.54	0.34

Table 3
Results of the final phase for Task 1: General Ternary Classification

User	Model	F1	Precision	Recall
Ryuan		0.58	0.58	0.58
Onarion		0.57	0.58	0.56
UC-UCO-CICESE_UT3-Plenitas(ymlopez)	**BERT**	**0.55**	**0.57**	**0.55**
VickBat		0.52	0.54	0.51
michaelibrahim		0.44	0.46	0.43
John94		0.43	0.45	0.42
VeronicaNeriMendoza		0.42	0.42	0.36
csuazob		0.34	0.54	0.34
AngelBaron		0.33	0.35	0.34

3.2. Task 2: Fine-Grained Hate Speech Classification

Internal Comparisons

In Task 2, the BERT-only model outperformed the multimodal BETO+VIT approach, achieving a higher score of 0.3704 compared to 0.3266, indicating that text alone was more effective than combining text and image data. The lower performance of BETO+VIT suggests that the addition of visual features

may have introduced noise or irrelevant information. File sizes also reflected this difference, with BERT-only outputs varying more widely due to preprocessing, while BETO+VIT files were consistently larger due to image embeddings. Overall, the text-only approach proved more effective for this task, and any future use of multimodal models would require improved fusion strategies or better visual feature extraction.

Comparison with other approaches

Top-performing model: BERT-only Performance (F1): 0.37; (Precision): 0.40 and (Recall): 0.36 For this subtask, the team employed BERT to detect hate speech subcategories (e.g., classism, racism, sexism). While the F1 score was lower than in Task 1, the model achieved a precision of 0.40 in sensitive categories like racism and sexism, underscoring its efficacy in identifying specific discriminatory content. Limitations emerged in minority classes (e.g., "others"), where lower recall suggested a need for data augmentation or class-balancing techniques to improve coverage. Immediately following this sentence is the point at which Table 4 and 5 are included in the input file; compare the placement of the table here with the table in the printed output of this document.

Table 4
Results of the development phase for Task 2: Fine-Grained Hate Speech Classification

User	Model	F1	Precision	Recall
HoracioJarquin		0.41	0.49	0.40
UC-UCO-CICESE_UT3-Plenitas(ymlopez)	**BERT**	**0.41**	**0.46**	**0.38**
hugojair		0.40	0.51	0.36
girish_koushik		0.32	0.33	0.32

Table 5
Results of the final phase for Task 2: Fine-Grained Hate Speech Classification

User	Model	F_1	Precision	Recall
UC-UCO-CICESE_UT3-Plenitas(ymlopez)	**BERT**	**0.37**	**0.40**	**0.36**
csuazob		0.27	0.29	0.26
michaelibrahim		0.26	0.35	0.25

3.3. Task 3: Ternary Classification Using Restricted LLMs

Internal Comparisons

In Task 3, DeepSeek-R1-Qwen-7B emerged as the top-performing language modelDeepSeek-R1-Qwen-7B a score of 0.5069, clearly outperforming all other tested models. Nous-Hermes followed with a solid, though lower, score of 0.4774. Mid-tier models like Mistral and Tiiulo performed similarly but significantly worse, scoring around 0.29. The lowest performance came from DeepSeek-R1-Qwen-7B, which achieved only 0.1745, indicating it may be unsuitable for the task without substantial optimization. File sizes ranged from approximately 16.8 KB to 23.8 KB, showing no clear link between size and effectiveness. Overall, DeepSeek-R1-Qwen-7B proved most effective, while models like Mistral, Tiiulo, and especially DeepSeek-R1-Qwen-7B may benefit from fine-tuning or further adaptation.

Comparison with other approaches

Top-performing model: DeepSeek-R1-Distill-Qwen 7B Performance (F1): 0.51, (Precision): 0.54 and (Recall): 0.50 In this task, which restricted solutions to large language models (LLMs), the distilled DeepSeek-Qwen 7B model (enhanced with DeepSeek-R1's reasoning techniques) outperformed competitors like Mistral 7B and Falcon-7B. Its advantages included reduced hallucination tendencies, computational efficiency, and adaptability to Mexican Spanish. Nevertheless, its slightly lower F1 compared to Task 3's multimodal approach suggests that LLMs could benefit from localized fine-tuning or hybrid architectures incorporating visual analysis. See Table 6 and 7.

Table 6
Results of the development phase for Task 3: Ternary Classification Using Restricted LLMs

User	Model	F1	Precision	Recall
HoracioJarquin		0.55	0.61	0.54
girish_koushik		0.48	0.50	0.53
Ryuan		0.43	0.43	0.44
UC-UCO-CICESE_UT3-Plenitas(ymlopez)	**DeepSeek-R1-Qwen-7B**	**0.41**	**0.43**	**0.41**
hugojair		0.40	0.42	0.40
dmoctezuma		0.33	0.33	0.33

Table 7
Results of the final phase for Task 3: Ternary Classification Using Restricted LLMs

User	Model	F1	Precision	Recall
LuisArellano		0.54	0.63	0.55
UC-UCO-CICESE_UT3-Plenitas(ymlopez)	**DeepSeek-R1-Qwen-7B**	**0.51**	**0.54**	**0.50**
csuazob		0.49	0.50	0.50

4. Conclusions

DIMEMEX 2025 experiments confirm that aligning model strategy with task demands is essential. Multimodal fusion (text + image) produced the strongest general-purpose result—Task 1: 3rd place, F1 = 0.55—while task-specific fine-tuning enhanced category-level precision—Task 2: 1st place, F1 = 0.37, despite its more granular hate speech classification. In restricted settings, lightweight LLMs proved competitive—Task 3: 2nd place, F1 = 0.51—demonstrating their effectiveness when visual cues are limited. However, challenges such as irony, sarcasm, and minority-class imbalance persist, indicating that future improvements may stem from hybrid LLM-plus-vision architectures, culturally tailored data augmentation, and adversarial training. No single model dominates across all scenarios, but a flexible, task-aware approach consistently yields the best outcomes. Our results demonstrate that demonstrate that:

- Multimodality is critical for generalization. The fusion of text (BETO) and image (ViT) features in Task 1 yielded the highest F1 score (0.55), underscoring the importance of leveraging both modalities for broad meme classification.
- Task-specific tuning enhances precision. While Task 2's F1 score was lower (0.37), fine-tuning for hate speech subcategories improved the model's ability to detect racism and sexism, revealing trade-offs between generalization and granularity.
- LLMs show promise but require refinement. DeepSeek-Qwen 7B's strong performance in Task 3 (F1 = 0.51) indicates that LLMs are viable in restricted settings; however, integrating visual components could help close the performance gap with fully multimodal models.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4 and Grammarly in order to: Grammar and spelling check. After using these services, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] H. Jarquín-Vásquez, I. Tlelo-Coyotecatl, M. Casavantes, D. I. Hernández-Farías, H. J. Escalante, L. Villaseñor-Pineda, M. Montes, et al., Panorama de dimemex en iberlef2025: Detección de memes

inapropiados de méxico, in: Actas del Foro Ibérico de Evaluación de Lenguas (IberLEF 2025), celebrado conjuntamente con la 41ª Conferencia de la Sociedad Española de Procesamiento del Lenguaje Natural (SEPLN 2025), CEUR-WS.org, 2025.
[2] L. y. J.-Z. S. M. González-Barba, José Ángel y Chiruzzo, Panorama general de IberLEF 2025: Desafíos del procesamiento del lenguaje natural para el español y otras lenguas ibéricas, in: Actas del Foro Ibérico de Evaluación de Lenguas (IberLEF 2025), celebrado conjuntamente con la 41ª Conferencia de la Sociedad Española de Procesamiento del Lenguaje Natural (SEPLN 2025), CEUR-WS.org, 2025.
[3] H. Jarquín-Vásquez, I. Tlelo-Coyotecatl, M. Casavantes, D. I. Hernández-Farías, H. J. Escalante, L. Villaseñor-Pineda, M. Montes, et al., Overview of dimemex at iberlef 2024: Detection of inappropriate memes from mexico, Procesamiento del Lenguaje Natural 73 (2024) 335–345.
[4] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, nature 521 (2015) 436–444.
[5] S. M. Jain, Hugging face, in: Introduction to transformers for NLP: With the hugging face library and models to solve problems, Springer, 2022, pp. 51–67.
[6] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, arXiv preprint arXiv:2308.02976 (2023).
[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).
[8] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al., Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, arXiv preprint arXiv:2501.12948 (2025).
[9] I. Ahmed, S. Islam, P. P. Datta, I. Kabir, N. U. R. Chowdhury, A. Haque, Qwen 2.5: A comprehensive review of the leading resource-efficient llm with potentioal to surpass all competitors, Authorea Preprints (2024).
[10] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al., Qwen3 technical report, arXiv preprint arXiv:2505.09388 (2025).
[11] N. Kotonya, S. Krishnasamy, J. Tetreault, A. Jaimes, Little giants: Exploring the potential of small llms as evaluation metrics in summarization in the eval4nlp 2023 shared task, arXiv preprint arXiv:2311.00686 (2023).
[12] D. S. Chaplot, Albert q. jiang, alexandre sablayrolles, arthur mensch, chris bamford, devendra singh chaplot, diego de las casas, florian bressand, gianna lengyel, guillaume lample, lucile saulnier, lélio renard lavaud, marie-anne lachaux, pierre stock, teven le scao, thibaut lavril, thomas wang, timothée lacroix, william el sayed, arXiv preprint arXiv:2310.06825 (2023).
[13] L. Basyal, M. Sanghvi, Text summarization using large language models: a comparative study of mpt-7b-instruct, falcon-7b-instruct, and openai chat-gpt models, arXiv preprint arXiv:2310.10449 (2023).

A. Online Resources

The results of the DIMEMEX 2025 are available via

- DIMEMEX 2025,