# HARGP-BETO: Hierarchical Text Interactions Model for Abuse Detection in Mexican Spanish Memes

Qiyuan Jin[1,*], Xiang Zhou[1]

[1] The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, 999077, Hong Kong

## Abstract

The automatic detection of abusive content in memes presents unique challenges in low-resource languages like Mexican Spanish, where cultural nuances and data scarcity compound traditional NLP difficulties. This study introduces HARGP-BETO, a novel hierarchical framework that combines advanced attention mechanisms with adaptive feature fusion for detecting hate speech and inappropriate content in Mexico Spanish memes. Our approach integrates dual-segment encoding with local-global attention interactions, enhanced through gated fusion and multi-level pooling strategies. Experimental results on the DIMEMEX corpus demonstrate the framework's effectiveness with macro-F1 score of 0.6139 and 68.24% accuracy, improving performance compared to existing baseline methods. While excelling at majority class detection (75.4% recall for non-abusive content), analysis reveals persistent challenges with minority class discrimination, particularly for hate speech categories. The results validate text-based approaches as computationally efficient alternatives to multimodal meme systems, while highlighting directions for addressing cultural specificity and data imbalance in Mexico Spanish abusive content detection.

## Keywords

Natural Language Processing, Hybrid Attention, Mexico Spanish, Abusive Content Detection

## 1. Introduction

Social media is playing an ever more significant role in people's daily lives [1]. Memes, combining images with text, are spreading rapidly and have become a popular way for people to communicate and express emotions. Nevertheless, some memes contain negative elements like insults, attacks on specific groups or individuals, and hate speech. Such memes with harmful content can foster bad trends such as prejudice and discrimination, causing serious negative impact, particularly on the healthy development of teenagers. The harmony of the online world is a pressing concern. As a result, in recent years, the detection and analysis of abusive content in social media have emerged as a hot topic in computational linguistics [2].

In English-speaking contexts, with the development of NLP, a variety of automatic detection methods for hate speech and abusive content have emerged, making remarkable progress [3, 4, 5]. However, in Spanish-speaking contexts, this challenge is much more severe. The unique linguistic features of Spanish, coupled with a lack of datasets, which include the influence of local languages and the emergence of localized neologisms [6], have collectively hindered progress in this area. Although previous work has made some headway and started to fill this gap [7], such as shared task DIMEMEX [8, 9], robust baselines for abusive meme classification are still limited.

To further advance this research, we conduct a study on the DIMEMEX dataset, with the aim of determining whether memes contain hate speech, inappropriate content, or neither. This paper aims to develop models that can accurately detect and classify different types of abusive content in Mexican Spanish memes. By addressing this challenge, we hope to create a safer and more respectful online world for Spanish-speaking users in Mexico and beyond.

*IberLEF 2025, September 2025, Zaragoza, Spain*

[*] Corresponding author.

meqiyuanj@ust.hk (Q. Jin); aafzoe@ust.hk (X. Zhou)
https://y99-stack.github.io/ (Q. Jin); https://zoeeehub.github.io/ (X. Zhou)
0009-0001-0182-2132 (Q. Jin); 0009-0002-4957-7663 (X. Zhou)

Moreover, since text serves as the primary and direct information carrier on social media platforms and memes often contain text conveying core meanings and emotions, this gives unique advantages to text-based models. Even though memes are essentially multi-modal, text alone can provide rich semantic information to identify whether a meme is abusive, which is particularly important when there are nuances in emotional expression. Compared to multi-modal approaches, text-based detection has several advantages, such as easier data acquisition and annotation, lower computational costs, and better interpretability. For all these reasons, this paper focuses on developing an effective text-based computational model to accurately detect and classify abusive content in Mexican Spanish memes [9].

## 2. Related Work

The tendency of hate speech to go viral significantly compounds the challenges of content moderation detection. Numerous scholars are dedicated to developing automated detection algorithms over years for monitoring negative content on social media. Davidson et al. [10] implemented unsupervised learning, utilizing crowd-sourced hate speech lexicons to train multi-class classifiers; Founta et al. [11] devised an incremental iterative methodology employing crowd-sourcing to annotate large-scale tweet collections with abuse-related labels; Kiela et al. [12] established the Hateful Memes Challenge Dataset, addressing gaps in multi-modal hate speech classification datasets; Bai et al. [13] introduced STATE ToxiCN, the first span-level Chinese hate speech dataset. As the world's third most spoken language, Spanish has also witnessed significant progress in hate speech detection. Exploratory work collectively drives domain advancement. For instance, scholars such as José Antonio García-Díaz [14], Montesinos-Cánovas [15], and Vallecillo-Rodríguez [16] have made notable contributions through their research.

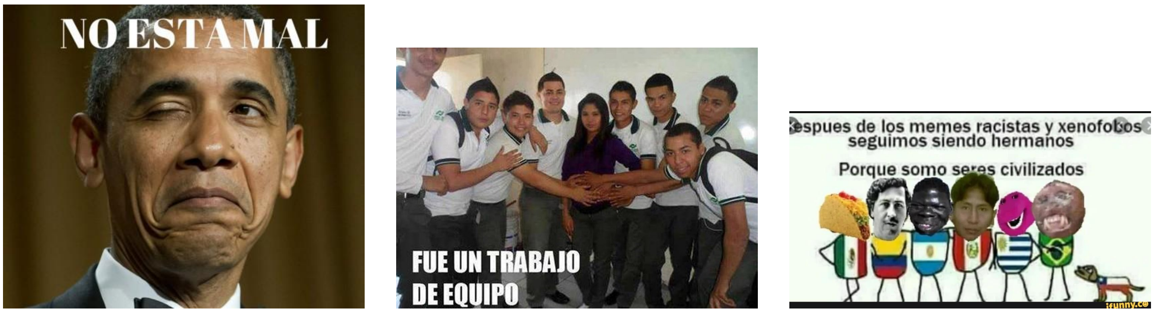
Mexico maintains critical significance within Spanish-speaking regions in the world. Its unique geography fosters complex digital linguistic ecosystems where slang, cultural allusions, regional dialects, and loanwords prevail [16]. These lingual-cultural phenomena substantially increase detection difficulties for hateful or abusive content. Many researchers have devoted significant effort to the field of offensive content detection in Mexican Spanish and have achieved some significant research outcomes. Gemma Bel-Enguix [17] introduce the T-MexNeg corpus, which is the first corpus annotated with negation in Twitter in Mexican Spanish. MEX-A3T task in IberLEF 2019 [18] and IberLEF 2020 [19] conferences focused on the detection of aggressive tweets. Subtask 3 and Subtask 4 in MeOffendES 2021 [20, 21] are related to the identification of offensive language targeting the Mexican variant of Spanish. Similarly, they continue to contribute to this field in 2023 - 2025 [22, 8, 23]. These studies and conferences demonstrate the evolution and progress in Mexican Spanish offensive content detection these years.

Nevertheless, previous work has primarily concentrated on pure text analysis. Memes serve as emerging carriers of hate speech with concealment that impedes identification. Developing efficient detection for Mexican Spanish demands extensive new studies.

## 3. Dataset

According to the introduction of the organizers, the DIMEMEX corpus consists of more than 3000 memes, compiled from public Facebook groups rooted in Mexico. Given that a significant amount of emotions, opinions and statements can be conveyed through memes, the dataset has been manually annotated to detect hate speech, inappropriate content, and neither within them, shown in Figure 1. Based on the task definition, this dataset can be used for a three-class classification problem, distinguishing between hate speech, inappropriate content, and neither. Additionally, it enables a more nuanced classification, differentiating instances of hate speech into various categories like classism, sexism, racism, and others. However, our primary focus here is on the three-class classification problem.

Each meme's information comes from two parts, text and image. The meme text is extracted from the images via a state-of-the-art OCR technique, and each meme has a unique meme ID. Table 1 outlines


NO ESTA MAL
FUE UN TRABAJO DE EQUIPO
espues de los memes racistas y xenofobos seguimos siendo hermanos
Porque somo seres civilizados
ifunny.co

(a) neither (b) inappropriate content (c) hate speech

**Figure 1:** Samples of images from the DIMEMEX dataset across three labels

the training dataset for the three - class classification problem, containing 2263 memes. In the training phase of this study, 85% of the dataset is used as the train dataset to train and optimize the model, while the remaining 15% serves as the validation set for preliminary performance evaluation and model tuning. During different stages of the competition, model predictions are made on the official test datasets, and the results are submitted to obtain performance scores. This ongoing assessment ensures the generalizability and practical effectiveness of the model.

**Table 1**
Dataset Statistics for DIMEMEX

| Labels | Samples | % |
|---|---|---|
| hate-speech | 386 | 17.06% |
| inappropriate content | 472 | 20.86% |
| none | 1405 | 62.09% |

## 4. Methodology

In this section, we describe our data preprocess and proposed framework, HARGP-BETO, for multi-aspect meme classification.

### 4.1. Data Augmentation

This paper utilizes a multimodal DIMEMEX dataset comprising text-image pair samples. Each sample $x_i$ is structured as follows:

$$x_i = (T_{ocr}, D_{ctx}, I_{meme}, y) \tag{1}$$

where $T_{ocr}$ represents the OCR text extracted from memes, $D_{ctx}$ is the description associated with the memes, $I_{meme}$ is the image of memes, and $y$ denotes the classification labels.

To construct the input for our model, we design a specialized data preprocessing pipeline with the following key steps. For dynamic text augmentation, we apply synonym replacement (using the nlpaug library) on 20% of the meme text instances during training to enhance model generalization:

$$T'_{ocr} = Aug_s(T_{ocr}), P(\text{augment}) = 0.2 \tag{2}$$

This approach helps to mitigate overfitting by introducing lexical variations while maintaining semantic consistency.

We conducted quantitative analysis of text length distributions using BERT tokenizershown in Table 2. Through hyperparameter experimentation, we chose the optimal maximum input lengths: 56 tokens for text and 240 tokens for descriptions. To effectively integrate these dual-text sources while preserving their distinct semantic roles, we design a specialized input format that explicitly segments the content,

**Table 2**
Statistics for text length

| Type | 50th percentile | 90th percentile | 99th percentile |
|---|---|---|---|
| OCR Text | 17 | 31 | 60 |
| Description | 128 | 164 | 247 |

as Equation (3) . This method extends standard BETO processing by incorporating explicit boundary markers between OCR-extracted meme text and descriptions, and balances information preservation and GPU memory constraints and reduce padding waste to the greatest extent.

$$\text{Input} = [\text{CLS}] \oplus T_{\text{ocr}}^{1:L_t} \oplus [\text{SEP}] \oplus D_{\text{ctx}}^{1:L_d} \oplus [\text{SEP}] \tag{3}$$

As our research primarily focuses on text, we only perform basic processing on the image information for contrastive experiments. The images of memes $I$ undergo some basic processing methods to obtain $I'$ and the standardized visual features are extracted through a ViT processor.

$$V_{\text{pixel}} = \text{ViTProcessor}(I') \in \mathbb{R}^{3\times224\times224} \tag{4}$$

Finally, text and images multimodal alignment can be achieved:

$$\mathcal{D}_{\text{final}} = (\mathbf{I}_{\text{text}}, \mathbf{M}_{\text{attn}}, \mathbf{T}_{\text{type}}, V_{\text{pixel}}, y) \tag{5}$$

These visual features exclusively support baseline comparisons in Section 5.1, including the ViT-only model and multimodal CLIP benchmark. This underscores our core methodological focus: demonstrating that textual signals, when properly processed through our proposed model, provide sufficient discriminative power for abuse detection without visual dependency.

## 4.2. Model

The model is a hierarchical text classification framework combining dual-segment interaction and adaptive feature fusion. The architecture comprises three key modules, shown in Figure 2:

1. BETO-based Dual-segment Encoder,
2. Hybrid attention with local-global interactions,
3. Gated hierarchical pooling.

**BETO-based Dual-segment Encoder**: The BETO encoder generates embeddings $\mathbf{H} \in \mathbb{R}^{B\times(L_t+L_d+3)\times d}$ [24]. We then divide the output of BETO output into text sequence $T \in \mathbb{R}^{B\times L_t}$ and description sequence $D \in \mathbb{R}^{B\times L_d}$ [25]. These are passed through a norm layer to get the normalized OCR text feature$T_{norm} \in \mathbb{R}^{B\times L_t\times d}$ and normalized descriptions features $D_{norm} \in \mathbb{R}^{B\times L_d\times d}$ , where $L_t$ is maximum input length of text, $L_d$ is maximum input length of descriptions.

**Hybrid Attention Mechanism**: This module jointly models the local contextual patterns of OCR text and the global cross - interaction between OCR text and descriptions [26, 27]. This hybrid attention mechanism combines the advantages of local causal attention and global cross-segment attention, and uses an adaptive feature fusion method to dynamically combine local and global features.

1. Local Causal Attentions

The window-restricted multi-head attention [28] capture local dependencies by restricting attention to a fixed-size window, thereby focusing on fine-grained contextual relationships.

$$\mathbf{Q}_L = \mathbf{T}_{\text{norm}}\mathbf{W}_Q^L, \quad \mathbf{K}_L = \mathbf{T}_{\text{norm}}\mathbf{W}_K^L, \quad \mathbf{V}_L = \mathbf{T}_{\text{norm}}\mathbf{W}_V^L \tag{6}$$

We introduce a causal mask $M \in \{0, -\infty\}^{L_t\times L_t}$ to to restrict attention to a window of size 3 for locality.

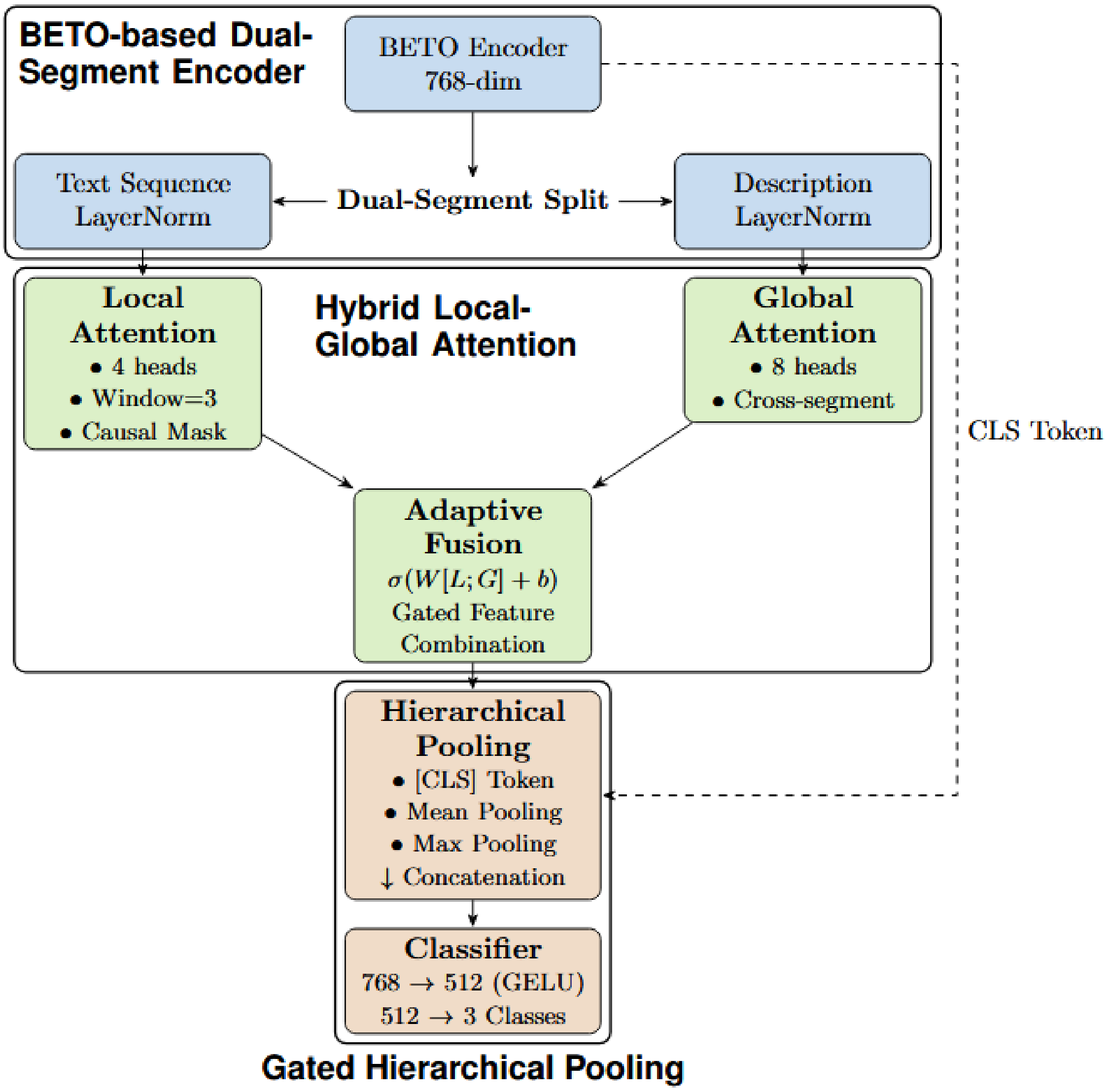


**Figure 2:** An overview of our proposed framework, HARGP-BETO

$$\mathbf{M}_{i,j} = \begin{cases} 0 & \text{if } j \leq i+1 \text{ (window size = 3)} \\ -\infty & \text{otherwise} \end{cases} \tag{7}$$

The local attention mechanism is then computed as Equation(8) . Additive masking, which is equivalent to multiplicative masking, uses large negative values to ensure numerical stability and hardware efficiency, making it more memory-bandwidth friendly.

$$\mathbf{A}_{\text{local}} = \text{Softmax}\left(\frac{\mathbf{Q}_L \mathbf{K}_L^\top}{\sqrt{d}} + \mathbf{M}\right) \mathbf{V}_L \in \mathbb{R}^{B \times L_t \times d} \tag{8}$$

2. Global Cross-Segment Attention

To obtain the better interactions between OCR text and descriptions, we employ a global multi-head attention mechanism here.

$$\mathbf{Q}_G = \mathbf{T}_{\text{norm}} \mathbf{W}_Q^G, \quad \mathbf{K}_G = \mathbf{D}_{\text{norm}} \mathbf{W}_K^G, \quad \mathbf{V}_G = \mathbf{D}_{\text{norm}} \mathbf{W}_V^G \tag{9}$$

$$\mathbf{A}_{\text{global}} = \text{Softmax}\left(\frac{\mathbf{Q}_G \mathbf{K}_G^\top}{\sqrt{d}}\right) \mathbf{V}_G \in \mathbb{R}^{B \times L_t \times d} \tag{10}$$

3. Adaptive Feature Fusion

To dynamically combine local and global features, we introduce a learnable gate that adaptively fuses these features of the input data [29, 30, 31].

$$\mathbf{G} = \sigma\left(\mathbf{W}_g[\mathbf{A}_{\text{local}}; \mathbf{A}_{\text{global}}]\right) \in \mathbb{R}^{B \times L_t \times d} \tag{11}$$

$$\mathbf{A}_{\text{fused}} = \mathbf{G} \odot \mathbf{A}_{\text{local}} + (1 - \mathbf{G}) \odot \mathbf{A}_{\text{global}} \in \mathbb{R}^{B \times L_t \times d} \tag{12}$$

where $\sigma$ denotes the sigmoid activation, and $\odot$ represents element-wise multiplication. This adaptive fusion allows the model to leverage both local and global information effectively.

**Hierarchical Pooling**: To enhance the text understanding capability of the model, we adopt a hierarchical pooling strategy inspired by previous work [32].

This strategy concatenates the CLS token $h_{\text{CLS}}$ from the beginning of the input sequence with the mean-pooled $\bar{A}$ and max-pooled $\tilde{A}$ outputs from the hybrid attention mechanism, resulting in multilevel pooling features. The concatenated features are then transformed using a GELU activation function to produce a comprehensive feature representation $\mathbf{P}$ [33, 34]. This allows the model to utilize global semantics, overall statistical features, and salient local information.

$$\mathbf{P} = \text{GELU}\left(\mathbf{W}_p[\mathbf{h}_{\text{CLS}}; \bar{\mathbf{A}}; \tilde{\mathbf{A}}]\right) \in \mathbb{R}^{B \times d} \tag{13}$$

where:

$$\bar{\mathbf{A}} = \frac{1}{L_t} \sum_{i=1}^{L_t} \mathbf{A}_{\text{fused},i}, \quad \tilde{\mathbf{A}} = \max_i(\mathbf{A}_{\text{fused},i})$$

# 5. Experiments

## 5.1. Ablation experiment

We evaluate model performance on the DIMEMEX dataset, monitoring precision, recall, F1-score, and accuracy. All models are initialized with BETO, with its parameters frozen during training; only the newly added modules are optimized. Each experiment is repeated five times, with results reported as averages. Our ablation studies focus on two key aspects: (1) comparing the effectiveness of using Cross Attention alone versus Hybrid Attention, and (2) assessing the contribution of Hierarchical Pooling.

To investigate the impact of each module on the performance of the model, we design a series of ablation experiments. Starting with BETO as the baseline model, we progressively introduce different components: cross-segment attention, residual connections, gated fusion, and hierarchical pooling. The experimental setup and results are shown in Table 3.

HARP-BETO introduces a paradigm shift by replacing cross-segment attention with hybrid attention, which integrates local causal attention and global cross-modal attention. This architecture captures both fine-grained contextual patterns within OCR text and high-level interactions between text and descriptions. HARGP-BETO further enhances this design by incorporating gated fusion to dynamically weight local and global features, as Equation (11) and Equation (12) , resulting in more adaptive feature representation. Both models employ hierarchical pooling for final feature aggregation.

We compare our model with both single-modal ViT and multimodal CLIP models. The results are shown in Table 4.

The ablation study in Table 4 systematically evaluates the impact of key architectural modules on the performance of the model. Our architectural decision to build on BETO's text-based foundation stems from three key observations: (1) The strong performance of Spanish-specialized BETO (61.77% precision)

**Table 3**
Ablation Experiments Model Architecture

| Model | Cross-Segment Attention | Hybrid Attention (Local+Global) | Residual Connection | Gated Fusion | Hierarchical Pooling |
|---|---|---|---|---|---|
| A-BETO | ✓ | × | × | × | × |
| AR-BETO | ✓ | × | ✓ | × | × |
| ARG-BETO | ✓ | × | ✓ | ✓ | × |
| ARGP-BETO | ✓ | × | ✓ | ✓ | ✓ |
| HARP-BETO | × | ✓ | ✓ | × | ✓ |
| HARGP-BETO | × | ✓ | ✓ | ✓ | ✓ |

**Table 4**
Results of different models on the DIMEMEX dataset

| Model | F1-Score | Precision | Recall | Accuracy |
|---|---|---|---|---|
| Vit | 0.4276 | 0.5623 | 0.4576 | 0.5458 |
| CLIP | 0.5769 | 0.5793 | 0.6054 | 0.6792 |
| BETO | 0.5648 | 0.6177 | 0.5988 | 0.6884 |
| A-BETO | 0.5957 | 0.6239 | 0.6084 | 0.6724 |
| AR-BETO | 0.5991 | 0.6190 | 0.6149 | 0.6561 |
| ARG-BETO | 0.6006 | 0.6246 | 0.6198 | 0.6860 |
| ARGP-BETO | 0.6053 | 0.6226 | 0.6171 | 0.6792 |
| HARP-BETO | 0.6112 | 0.6281 | 0.6194 | 0.6770 |
| HARGP-BETO | 0.6139 | 0.6258 | 0.6242 | 0.6862 |

versus general multilingual CLIP (57.93%) validates the necessity of language-specific modeling for nuanced Mexico Spanish. (2) Although CLIP's visual-text alignment improves recall (+1.66% over BETO), its limited Spanish pretraining causes precision degradation in culture-specific contexts. (3) ViT's poor performance (42.76% F1) confirms that visual patterns alone lack sufficient semantic signals for abuse content detection.

We dissect the role of each component based on their incremental contributions. A-BETO demonstrates that cross-attention enhances feature interaction between segments but lacks hierarchical refinement. Introducing a residual connection in AR-BETO stabilizes training during backpropagation to some extent, generating a gain of 0. 34% F1 despite the increase in the parameter count. For ARG-BETO, gating mechanism mitigates noise but underperforms for minority classes with Class 1 (inappropriate content) $F1 = 0.49$. For ARGP-BETO, multiscale pooling diversifies feature aggregation, particularly benefiting Class 0 (neither) ($F1 = 0.82 \rightarrow$ +4% vs. AR-BETO). HARP-BETO combines local and global attention improving Class 1 recall (43% → 50%), but precision drops due to overfitting in sparse samples. As for HARGP-BETO, adding gating to hybrid attention output refines hybrid features, which perform best in these models.

## 5.2. Confusion Matrix

The HARGP-BETO model demonstrates robust performance in integrating multiscale features through its hybrid attention mechanism, which effectively combines local and global semantic patterns to achieve a balanced macro-F1 score of 0.6139, as seen in Figure 3. The model excels in classifying the majority class (Class 0 with 172 TP and 75.4% recall), showcasing its ability to leverage hierarchical attention and gated fusion for stable feature aggregation. The multilevel pooling strategy through concatenated representations [CLS], mean-pooled, and max-pooled further enhances discriminative power by capturing various statistical signals, contributing to an overall accuracy of 68.24%.

However, the model has limitations due to data imbalance in the dataset. Class 1 has a low recall of 46.6%, with 23 samples misclassified as Class 0. Class 2 shows moderate performance (61.1% recall) with

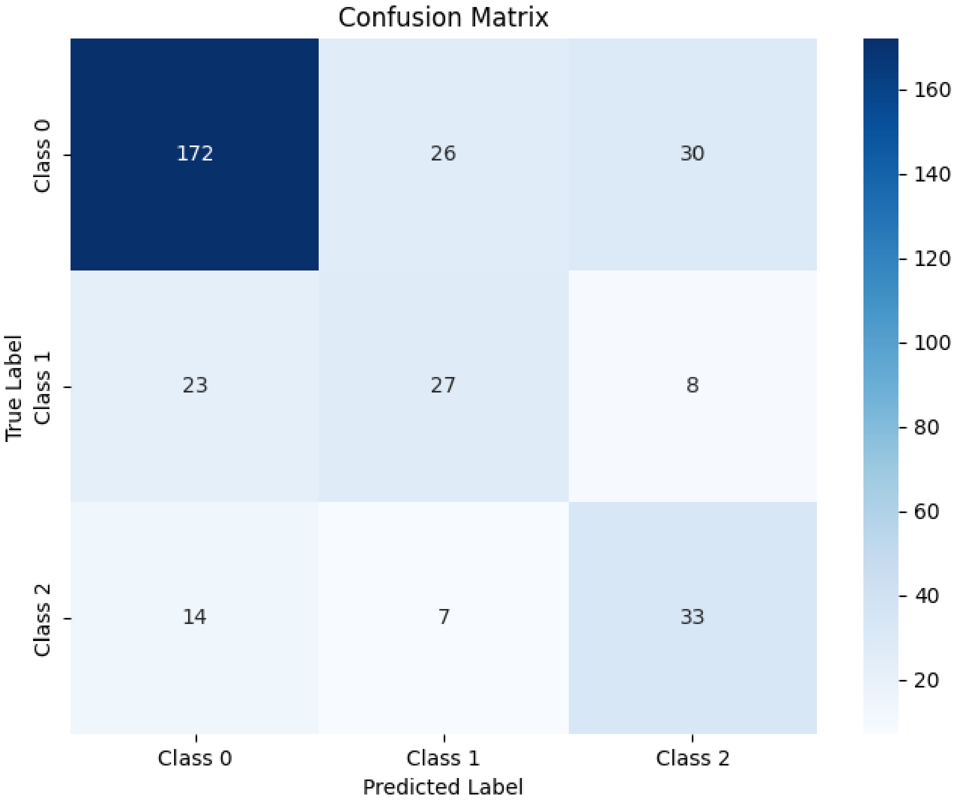


**Figure 3:** Confussion matrix with HARGP-BETO predictions. Class 0: neither, Class 1: inappropriate content, Class2: hate speech

14 misclassifications to Class 0. This suggests that the gating mechanism may be biased towards the weight allocation of majority classes, suppressing the characteristics of low-frequency classes. In the future, dynamic category weights, contrastive learning losses and hierarchical attention optimization need to be adopted to alleviate the imbalance and enhance the fine-grained discrimination ability.

## 6. Conclusion

This study presents HARGP-BETO, a novel hierarchical framework for detecting abusive content in Mexican Spanish memes, leveraging hybrid attention mechanisms and adaptive feature fusion to address the challenges of multimodal and imbalanced data. The proposed HARGP-BETO model achieves a macro-F1 score of 0.6139 and an accuracy of 68. 24%, demonstrating its effectiveness in integrating local and global semantic patterns through gated fusion and multilevel pooling. The hierarchical architecture, particularly the hybrid attention design, significantly improves feature interaction between OCR text and contextual descriptions, enabling robust performance on the majority class (Class 0: 75.4% recall) while maintaining balanced precision-recall trade-offs. These advancements highlight the potential of text-based approaches in abusive meme detection, which are computationally efficient compared to multimodal methods. Additionally, text-based models are more conducive to leveraging post-hoc attribution analysis and visualization tools after training, making them more easily interpretable than purely image-based methods.

However, the model's performance on minority classes (e.g., Class 1 recall= 46.6%) highlights persistent challenges rooted in data imbalance and feature ambiguity. Future work should focus on dynamic class-aware gating, contrastive learning for minority-class discrimination, and enhanced local attention mechanisms to mitigate bias. By refining these aspects, the framework could be extended to other low-resource languages, fostering safer internet spaces while preserving cultural and linguistic nuances in abusive content detection.

## Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

## References

[1] P. F. Bruning, B. J. Alge, H.-C. Lin, Social networks and social media: Understanding and managing influence vulnerability in a connected society, Business Horizons 63 (2020) 749–761.

[2] M. J. Díaz-Torres, P. A. Morán-Méndez, L. Villasenor-Pineda, M. Montes, J. Aguilera, L. Meneses-Lerín, Automatic detection of offensive language in social media: Defining linguistic criteria to build a mexican spanish dataset, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, 2020, pp. 132–136.

[3] N. S. Mullah, W. M. N. W. Zainon, Advances in machine learning algorithms for hate speech detection in social media: a review, IEEE Access 9 (2021) 88364–88376.

[4] M. Kumar, et al., Exploring hate speech detection: challenges, resources, current research and future directions, Multimedia Tools and Applications (2025) 1–37.

[5] A. Toktarova, D. Syrlybay, B. Myrzakhmetova, G. Anuarbekova, G. Rakhimbayeva, B. Zhylanbaeva, N. Suieuova, M. Kerimbekov, Hate speech detection in social networks using machine learning and deep learning methods, International Journal of Advanced Computer Science and Applications 14 (2023).

[6] A. Mojedano Batel, M. Adams, P. Pezik, Native dialect influence detection (ndid): Differentiating between mexican and peninsular l1 spanish in l2 english, Language and Law/Linguagem e Direito 9 (2022) 120–145.

[7] R. N. M. Mercado, H. F. C. Chuctaya, E. G. C. Gutierrez, Automatic cyberbullying detection in spanish-language social networks using sentiment analysis techniques, International Journal of Advanced Computer Science and Applications 9 (2018).

[8] 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.

[9] J. Á. González-Barba, L. Chiruzzo, S. M. Jiménez-Zafra, 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, 2025.

[10] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, 2017. URL: https://arxiv.org/abs/1703.04009. arXiv:1703.04009.

[11] A.-M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, N. Kourtellis, Large scale crowdsourcing and characterization of twitter abusive behavior, 2018. URL: https://arxiv.org/abs/1802.00393. arXiv:1802.00393.

[12] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The hateful memes challenge: Detecting hate speech in multimodal memes, 2021. URL: https://arxiv.org/abs/2005.04790. arXiv:2005.04790.

[13] Z. Bai, S. Yin, J. Lu, J. Zeng, H. Zhu, Y. Sun, L. Yang, H. Lin, State toxicn: A benchmark for span-level target-aware toxicity extraction in chinese hate speech detection, 2025. URL: https://arxiv.org/abs/2501.15451. arXiv:2501.15451.

[14] J. A. García-Díaz, S. M. Jiménez-Zafra, M. A. García-Cumbreras, R. Valencia-García, Evaluating feature combination strategies for hate-speech detection in Spanish using linguistic features and transformers, Complex &amp; Intelligent Systems 9 (2022) 2893–2914.

[15] E. Montesinos-Cánovas, F. Garcia-Sánchez, J. A. Garcia-Díaz, G. Alcaraz Mármol, R. Valencia García, Spanish hate-speech detection in football (2023).

[16] H. Gomez-Adorno, G. Bel-Enguix, G. Sierra, J.-C. Barajas, W. Álvarez, Machine learning and deep learning sentiment analysis models: Case study on the sent-covid corpus of tweets in mexican spanish, Informatics 11 (2024). URL: https://www.mdpi.com/2227-9709/11/2/24.

[17] G. Bel-Enguix, H. Gómez-Adorno, A. Pimentel, S.-L. Ojeda-Trueba, B. Aguilar-Vizuet, Negation detection on mexican spanish tweets: The t-mexneg corpus, Applied Sciences 11 (2021). URL: https://www.mdpi.com/2076-3417/11/9/3880. doi:10.3390/app11093880.

[18] M. E. Aragón, M. Á. Álvarez-Carmona, M. M. y Gómez, H. J. Escalante, L. V. Pineda, D. Moctezuma, Overview of mex-a3t at iberlef 2019: Authorship and aggressiveness analysis in mexican spanish tweets, in: IberLEF@SEPLN, 2019. URL: https://api.semanticscholar.org/CorpusID:267061516.

[19] M. E. Aragón, H. J. Jarquín-Vásquez, M. M. y Gómez, H. J. Escalante, L. V. Pineda, H. Gómez-Adorno, J. P. Posadas-Durán, G. Bel-Enguix, Overview of mex-a3t at iberlef 2020: Fake news and aggressiveness analysis in mexican spanish, in: IberLEF@SEPLN, 2020, pp. 222–235. URL: https://ceur-ws.org/Vol-2664/mex-a3t_overview.pdf.

[20] F. M. Plaza-del Arco, M. Casavantes, H. J. Escalante, M. T. Martín-Valdivia, A. Montejo-Ráez, M. Montes-y Gómez, H. Jarquín-Vásquez, L. Villaseñor-Pineda, Overview of MeOffendEs at IberLEF 2021: Offensive Language Detection in Spanish Variants, Procesamiento del Lenguaje Natural 67 (2021) 183–194. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6388, number: 0.

[21] F. M. Plaza-del Arco, M. Casavantes, H. Escalante, M. Martin-Valdivia, A. Montejo-Ráez, M. Montes-y Gómez, H. Jarquín-Vásquez, L. Villasenor-Pineda, Overview of the meoffendes task on offensive text detection at iberlef 2021, Procesamiento del Lenguaje Natural 67 (2021).

[22] H. Jarquín-Vásquez, D. I. Hernández-Farías, L. J. Arellano, H. J. Escalante, L. Villaseñor-Pineda, M. Montes, F. Sanchez-Vega, et al., Overview of da-vincis at iberlef 2023: Detection of aggressive and violent incidents from social media in spanish, Procesamiento del Lenguaje Natural 71 (2023) 351–360.

[23] T.-C. I. H.-F. D. I. E. H. J. V.-P. L. M.-y.-G. M. Jarquín-Vásquez, Horacio, Overview of DIMEMEX at IberLEF2025: Detection of Inappropriate Memes from Mexico, Procesamiento del Lenguaje Natural 75 (2025).

[24] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 4171–4186.

[25] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[26] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, arXiv preprint arXiv:2004.05150 (2020).

[27] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, et al., Big bird: Transformers for longer sequences, Advances in neural information processing systems 33 (2020) 17283–17297.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[29] F. Liu, S.-Y. Shen, Z.-W. Fu, H.-Y. Wang, A.-M. Zhou, J.-Y. Qi, Lgcct: A light gated and crossed complementation transformer for multimodal speech emotion recognition, Entropy 24 (2022) 1010.

[30] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, Advances in neural information processing systems 32 (2019).

[31] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using rnn encoder-decoder for statistical machine translation, arXiv preprint arXiv:1406.1078 (2014).

[32] A. Tao, K. Sapra, B. Catanzaro, Hierarchical multi-scale attention for semantic segmentation, arXiv

preprint arXiv:2005.10821 (2020).

[33] A. Aberdam, R. Litman, S. Tsiper, O. Anschel, R. Slossberg, S. Mazor, R. Manmatha, P. Perona, Sequence-to-sequence contrastive learning for text recognition, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 15302–15312.

[34] D. Hendrycks, K. Gimpel, Gaussian error linear units (gelus), arXiv preprint arXiv:1606.08415 (2016).

## A. Online Resources

The sources for the CEUR-art style are available via

- GitHub