Meme Classification using ModernBERT

Michael Ibrahim

Computer Engineering Department, Cairo University, 1 Gamaa Street, 12613, Giza, Egypt

Abstract

The rapid proliferation of memes on social media necessitates robust automated systems for detecting harmful content, particularly in linguistically diverse contexts like Mexican Spanish. This paper presents a benchmark study for the DIMEMEX shared task at IberLEF 2025, leveraging ModernBERT, a transformer model enhanced with rotary positional embeddings (RoPE) and FlashAttention, to classify memes into three categories (hate speech, inappropriate content, neither) and subcategorize hate speech into classism, sexism, racism, and others. Our hierarchical multi-task learning framework achieved macro-F1 scores of 0.44 on Subtask 1 and 0.26 on Subtask 2, underscoring the challenges of fine-grained classification in low-resource, culturally nuanced settings. This work demonstrates ModernBERT's potential for long-context understanding while highlighting the need for multimodal approaches. Future research should prioritize synthetic data generation to address label scarcity, integrate vision-language architectures for joint text-image modeling, and refine ethical safeguards against cultural bias.

Keywords

Meme Classification, ModernBERT, Text Classification, Transformer models

1. Introduction

[1] The exponential growth of user-generated content on social media platforms has intensified the need for automated systems capable of detecting inappropriate or harmful language. Memes, a prevalent form of social media content that combines images and text, often convey complex cultural and linguistic nuances, making the task of identifying inappropriate content particularly challenging. These multimodal artifacts require models to integrate visual and textual semantics while decoding implicit cultural references, a task that remains understudied in low-resource languages like Mexican Spanish The DIMEMEX shared task at IberLEF 2025 [2] addresses this challenge by focusing on the detection of inappropriate memes from Mexico, with subtasks that emphasize the classification of the textual content embedded in memes. Specifically, the first two subtasks involve (1) determining whether a meme's text is inappropriate and (2) categorizing the type of inappropriateness, requiring models that can understand subtle linguistic cues and cultural context in Spanish.

Recent advances in Natural Language Processing (NLP) have been driven by the emergence of transformer-based models, with BERT (Bidirectional Encoder Representations from Transformers) [3] marking a major milestone. BERT's bidirectional context modeling and pretraining on large corpora have significantly improved performance across a wide range of text classification tasks, including sentiment analysis, hate speech detection, and topic categorization [4]. However, the original BERT architecture has limitations in efficiency and handling long contextual sequences, which are often necessary for understanding complex and nuanced texts such as memes. For instance, BERT's fixed 512-token window struggles with memes where sarcasm or hate speech emerges from the interplay of text and image over extended contexts.

ModernBERT, a recent evolution of BERT, incorporates several architectural and training optimizations that address these limitations, including rotary positional embeddings (RoPE) for dynamic positional encoding, Flash Attention 2 for accelerated computation, and extended context windows of up to 8,192 tokens [5]. These improvements enable ModernBERT to process longer texts more

IberLEF 2025, September 2025, Zaragoza, Spain

michael.nawar@eng.cu.edu.eg (M. Ibrahim)
www.linkedin.com/in/michael-ibrahim-90 (M. Ibrahim)
0000-0003-2340-8917 (M. Ibrahim)

effectively and with greater computational efficiency, making it well-suited for tasks involving detailed semantic understanding and fine-grained classification. Studies have demonstrated that ModernBERT outperforms conventional BERT models in various domains, including medical text classification (e.g., Clinical ModernBERT [6]) and long-context retrieval tasks, without sacrificing accuracy.

The application of ModernBERT to text classification tasks, especially in low-resource or domain-specific settings such as Spanish-language memes, benefits from transfer learning and fine-tuning strategies that adapt the model to the target data distribution. For example, synthetic data generation using large language models (LLMs) like GPT-4 has been shown to enhance ModernBERT's performance in low-resource scenarios, achieving F1 scores of 0.89 on domain classification tasks with only 1,000 synthetic examples [7]. Fine-tuning ModernBERT on domain-specific datasets, such as clinical narratives or social media corpora, further enhances its ability to generalize and detect subtle forms of inappropriate content. Moreover, hybrid approaches that combine ModernBERT embeddings with multimodal architectures (e.g., CLIP, UNITER) or convolutional neural networks (CNNs) have been proposed to capture both global context and local textual features, improving classification accuracy in hateful meme detection [1].

The DIMEMEX shared task provides a unique benchmark for evaluating these approaches in the context of Mexican Spanish memes, where cultural and linguistic specificities pose additional challenges. For instance, Mexican memes often employ regional slang, code-mixing, and historical references that require models to internalize both language and cultural knowledge. Leveraging ModernBERT's capabilities, this paper explores its effectiveness in the first two subtasks of DIMEMEX, aiming to achieve robust detection and categorization of inappropriate meme texts.

The remainder of this paper is organized as follows. Section 2 reviews related work on text classification, transformer architectures, and inappropriate content detection. Section 3 details our methodology, including dataset preprocessing and hyperparameter configurations, Section 4 discusses the results, and Section 5 concludes the work with future directions.

2. Related Work

Text classification has been a foundational task in NLP, evolving significantly from traditional machine learning methods to modern deep learning and transformer-based approaches. The first two subtasks of the DIMEMEX shared task at IberLEF 2025, focus on detecting and categorizing inappropriate memes, build upon this rich research landscape. This section elaborates on the progression of methodologies, highlighting the role of ModernBERT and related models in state-of-the-art text classification.

Historically, text classification relied on classical machine learning models such as Support Vector Machines (SVM) [8], Naive Bayes [9], Decision Trees, and Random Forests [10], often using bag-of-words or TF-IDF features. While effective for simpler tasks, these approaches struggled with capturing semantic context and nuances in language, particularly in noisy or informal text such as social media posts or memes.

The introduction of neural networks, especially convolutional neural networks (CNNs) [11] and recurrent neural networks (RNNs) [12], marked a significant improvement by learning dense representations and sequential dependencies in text . Hybrid models combining CNNs and bidirectional LSTMs [13] have shown strong performance on benchmark datasets like SST-2 and AG News. However, these architectures still faced challenges in modeling long-range dependencies and complex contextual relationships.

The transformer architecture [14] revolutionized NLP by enabling models to attend globally to input sequences, overcoming the limitations of RNNs. BERT (Bidirectional Encoder Representations from Transformers) [3] further advanced the field by pretraining deep bidirectional representations on massive corpora and fine-tuning on downstream tasks. BERT's ability to capture rich contextual information bidirectionally has made it the backbone for many text classification tasks, including offensive language and hate speech detection.

In multilingual and Spanish-specific contexts, models like mBERT [3] and BETO [15] have been adapted and fine-tuned, demonstrating strong performance in Iberian languages [16]. The DIMEMEX task leverages these advances by applying ModernBERT, a refined variant of BERT with optimized pretraining and fine-tuning strategies, to classify meme texts for inappropriateness [2].

ModernBERT represents the next generation of BERT-based models, incorporating improvements such as longer context windows, more efficient training objectives, and domain-adaptive pretraining [5]. For example, llm-jp-modernbert [17] extends the context length to 8192 tokens, enabling better handling of long documents . These enhancements translate into superior performance on classification tasks, especially in domains requiring nuanced understanding.

Recent studies have also combined BERT with CNN classifiers to exploit local feature extraction alongside contextual embeddings, yielding improved accuracy in sentiment analysis and social media text classification. Hierarchical BERT models [18] have been proposed to handle multilevel classification effectively, as demonstrated in e-Commerce comment classification, where parent and subclass BERT models are trained sequentially to enhance granularity and accuracy.

Despite the dominance of transformer models, recent comparative studies have revealed that simpler models like logistic regression or SVM with optimized n-gram features can sometimes outperform more complex architectures, particularly when hyperparameter tuning is insufficient . This underscores the importance of rigorous optimization during fine-tuning, including learning rate schedules, batch sizes, and early stopping criteria [19].

Moreover, discriminative encoder-only models like BERT consistently outperform generative decoder-only models (e.g., GPT) on supervised classification tasks [19]. Transfer learning and cross-validation techniques have been instrumental in maximizing BERT's effectiveness for multi-class classification, as evidenced in experiments on datasets such as 20 Newsgroups and Reuters.

The detection of inappropriate or harmful content, especially in social media and memes, has been an active area of research. Multimodal approaches combining textual and visual features have been explored, but text-only models based on BERT remain highly competitive [2]. The DIMEMEX shared task is a notable benchmark focusing on culturally specific memes from Mexico, challenging models to detect subtle forms of inappropriateness and categorize them accurately [2].

The success of ModernBERT in this context is supported by its ability to capture complex semantic cues and contextual dependencies, which are critical for distinguishing nuanced categories of inappropriate content. Its fine-tuning on domain-specific data, combined with hierarchical classification strategies, aligns with best practices identified in recent literature.

3. Methodology

3.1. Dataset and Label Distribution

The Detection of Inappropriate Memes from Mexico (DIMEMEX) dataset [2] provides a benchmark corpus composed of Mexican Spanish memes, annotated with two levels of classification. The first level (Subtask 1) involves a three-way classification: identifying each meme as either *hate speech*, *inappropriate content*, or *neither*. The second level (Subtask 2) applies only to those memes labeled as hate speech and involves the detection of specific subcategories of hate speech.

Subtask 2 comprises six independent binary classification tasks, each corresponding to one hate speech subtype: classism, racism, sexism, and other hate speech. Each meme can belong to one or several of these categories, motivating the need for a fine-grained, multi-aspect approach. The label distribution across these subcategories is highly imbalanced, further complicating effective training and generalization.

In Subtask 1, the label proportions are distributed as follows: 62% "neither," 23% "inappropriate content," and 15% "hate speech." In Subtask 2, among hate speech instances, the six subcategories are sparsely populated, with classism and racism being the most prevalent. Due to this imbalance and task segmentation, we employed targeted strategies for data balancing and architecture design.

3.2. Input Representation and Preprocessing

Each meme instance includes a `description` field that contains the text intended for analysis. This field, provided as part of the dataset, serves as the sole input to the model. No image data or OCR-extracted captions were used.

Text inputs were tokenized using the ModernBERT tokenizer, capable of handling sequences up to 8192 tokens. This capacity was particularly useful for memes containing verbose or contextually dense language. Preprocessing steps included lowercasing, normalization of punctuation, removal of URLs and emojis, and stripping of non-linguistic characters. We deliberately avoided stemming and lemmatization to preserve sociolinguistic signals such as slang, colloquial phrasing, and regional idioms, which are crucial for detecting cultural nuance in hate speech.

3.3. Model Architecture

The architecture is based on ModernBERT, specifically the `jorgeortizfuentes/tulio-modernbert-spanish` variant. This model incorporates two key enhancements: Rotary Positional Embeddings (RoPE) for improved long-range dependency modeling, and FlashAttention2 for accelerated and memory-efficient attention computation.

For Subtask 1, a standard softmax classification head was added on top of the [CLS] token output to predict one of the three mutually exclusive classes. This head was trained using categorical cross-entropy loss.

For Subtask 2, six **independent binary classifiers** were constructed, each corresponding to one of the hate speech subcategories. These classifiers were trained separately on filtered subsets of the data, where only the hate speech-labeled memes were included. Each classifier operates independently, receiving the shared ModernBERT-encoded [CLS] token as input, and applies a sigmoid activation to output the probability of the specific subcategory. This modular, one-vs-rest setup ensured the capacity for nuanced pattern recognition without interference between labels.

3.4. Training Configuration

Fine-tuning was carried out using the following configuration:

- **Pretrained Model**: `jorgeortizfuentes/tulio-modernbert-spanish`
- **Tokenizer**: ModernBERT tokenizer (max length: 8192 tokens)
- **Optimizer**: AdamW
- **Learning Rate**: 5e-5, with linear warmup over the first 10% of training steps
- **Batch Size**: 16
- **Epochs**: 5
- **Loss Function**: Cross-entropy for all tasks (multi-class for Subtask 1, binary for Subtask 2 classifiers)
- **Hardware**: NVIDIA T4 GPU with 16GB memory

To address class imbalance, we used inverse-frequency class weighting, combined with random oversampling of minority classes during training. All experiments followed a stratified 5-fold cross-validation setup, with 80/20 train-validation splits maintained within each fold. Early stopping was disabled to allow full training cycles, and the model checkpoint with the highest validation macro-F1 score was retained.

3.5. Implementation and Inference

The implementation leveraged PyTorch and Hugging Face's Transformers library. Mixed-precision training (FP16) was enabled to reduce memory consumption and accelerate training. For inference, the model architecture was optimized for deployment on a single NVIDIA T4 GPU, supporting real-time classification in content moderation pipelines.

4. Results

We evaluated our ModernBERT-based model across different training configurations for both subtasks in the DIMEMEX challenge using 5-fold cross-validation. Table 1 summarizes the average macro-F1, precision, and recall scores on the validation sets.

Table 1
Cross-Validation Performance across Configurations

Configuration	Subtask	Macro-F1	Precision	Recall
Baseline (no weights)	1	0.40	0.44	0.38
Weighted Loss	1	0.43	0.47	0.42
+ Oversampling	1	0.49	0.50	0.45
Final Run (best config)	2	0.28	0.31	0.27

For Subtask 1, our best configuration combined class-weighted loss and oversampling of minority classes. This configuration yielded the highest macro-F1 score of 0.49, demonstrating notable improvements over the baseline. Oversampling especially boosted the recall of the *hate speech* class, which was severely underrepresented.

Subtask 2 posed greater challenges. Despite leveraging domain-specific pretraining and targeted fine-tuning, performance plateaued at a macro-F1 score of 0.28. This outcome reflects both the inherent difficulty of distinguishing between hate speech subtypes and the compounding effect of error propagation from Subtask 1.

We also monitored training and validation curves across all folds. Loss convergence was stable, though signs of overfitting began appearing around epoch 4 in the baseline setup. Early stopping was not required in the best-performing runs, and performance gains from oversampling were consistently observed across folds.

The results affirm that while ModernBERT's architecture supports improved contextual understanding, its performance is bounded by data scarcity, subtle semantics, and label granularity—challenges that motivate future multimodal and culturally informed approaches.

5. Conclusion and Future Work

This study investigated the application of the ModernBERT architecture to the Detection of Inappropriate Memes from Mexico (DIMEMEX) shared task, which focuses on categorizing Mexican Spanish memes into high-level categories (hate speech, inappropriate content, or neither) and further identifying six distinct subtypes of hate speech. Leveraging rotary positional embeddings and long-context encoding, ModernBERT was adapted to a hierarchical setup with a multi-class classification head for Subtask 1 and six independent binary classifiers for Subtask 2.

Despite the architectural advantages of ModernBERT, the model faced considerable challenges inherent to the task. On the DIMEMEX official test set, our best system achieved a macro-F1 score of 0.44 for Subtask 1 and 0.26 averaged across the six binary classifiers in Subtask 2. These modest scores underscore the difficulty of the problem, which is amplified by extreme class imbalance, the subtleties of cultural and linguistic expression in memes, and the absence of visual context.

Key limitations were identified in both the data and the modeling approach. Most notably, the majority of training examples in Subtask 1 were labeled "neither," limiting the model's exposure to harmful content. Additionally, Subtask 2 required the detection of subtle and often overlapping expressions of hate, which are not always easily separable from informal or regional language. The reliance on text-only inputs further constrained performance, as memes are inherently multimodal, and critical information is frequently embedded in the accompanying image.

To advance this line of research, several promising directions emerge. First, addressing label imbalance through data augmentation techniques—such as GPT-based paraphrasing, synthetic minority oversampling, or contrastive learning—may improve recall for underrepresented classes. Second, the incorporation of visual information via multimodal models (e.g., CLIP, vision transformers) would provide contextual cues often missing from text alone. Third, the pipeline could be restructured by decoupling Subtask 1 and Subtask 2 into standalone classifiers, which may reduce cumulative error and better capture the hierarchical nature of the problem.

Future work should also explore uncertainty-aware training and active learning, which can prioritize ambiguous or borderline cases for manual annotation. This could improve the quality of labels, particularly for the more subjective hate speech subcategories. Lastly, ethical considerations must guide model design and deployment. Bias mitigation strategies—such as adversarial debiasing, dialect-aware calibration, and post-hoc fairness audits—are essential to avoid disproportionate moderation of informal, dialectal, or culturally specific content that is not inherently harmful.

Overall, while the current results reflect the difficulty of automated meme moderation in low-resource, culturally rich contexts, they establish a robust foundation for future innovation. Progress in model architecture, training methodologies, and ethical oversight will be key to developing systems that balance moderation effectiveness with cultural sensitivity and fairness.

6. Declaration on Generative AI

During the preparation of this work, the author used ChatGPT, Grammarly in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the author reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] P. Kapil, A. Ekbal, A transformer based multi task learning approach to multimodal hate speech detection, Natural Language Processing Journal 11 (2025) 100133.

[2] 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).

[3] 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.

[4] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, J. Gao, Deep learning–based text classification: a comprehensive review, ACM computing surveys (CSUR) 54 (2021) 1–40.

[5] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, et al., Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, arXiv preprint arXiv:2412.13663 (2024).

[6] S. A. Lee, A. Wu, J. N. Chiang, Clinical modernbert: An efficient and long context encoder for biomedical text, arXiv preprint arXiv:2504.03964 (2025).

[7] D. Berenstein, Fine-tune modernbert for text classification using synthetic data, 2024. URL: https://huggingface.co/blog/davidberenstein1957/fine-tune-modernbert-on-synthetic-data, hugging Face Blog.

[8] T. Joachims, et al., Transductive inference for text classification using support vector machines, in: Icml, volume 99, 1999, pp. 200–209.

[9] A. McCallum, K. Nigam, et al., A comparison of event models for naive bayes text classification, in: AAAI-98 workshop on learning for text categorization, volume 752, Madison, WI, 1998, pp. 41–48.

[10] L. Breiman, Random forests, Machine learning 45 (2001) 5–32.

[11] Y. Kim, Convolutional neural networks for sentence classiication. emnlp 2014-2014 conference on empirical methods in natural language processing, in: Proceedings of the Conference null, null (2014), 1746ś1751. https://doi. org/10.3115/v1/d14-1181, 2014.

[12] S. Lai, L. Xu, K. Liu, J. Zhao, Recurrent convolutional neural networks for text classification, in: Proceedings of the AAAI conference on artificial intelligence, volume 29, 2015.

[13] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (1997) 1735–1780.

[14] 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).

[15] 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).

[16] 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.

[17] I. Sugiura, K. Nakayama, Y. Oda, llm-jp-modernbert: A modernbert model trained on a large-scale japanese corpus with long context length, 2025. URL: https://arxiv.org/abs/2504.15544. arXiv:2504.15544.

[18] X. Zhang, F. Wei, M. Zhou, Hibert: Document level pre-training of hierarchical bidirectional transformers for document summarization, arXiv preprint arXiv:1905.06566 (2019).

[19] L. Galke, A. Diera, B. X. Lin, B. Khera, T. Meuser, T. Singhal, F. Karl, A. Scherp, Are we really making much progress in text classification? a comparative review, arXiv preprint arXiv:2204.03954 (2022).