CLTL at DIMEMEX Shared Task: Fine-Grained Detection of Hate Speech in Memes

Yeshan Wang[1], Ilia Markov[1]

[1]Computational Linguistics & Text Mining Lab (CLTL)
Vrije Universiteit Amsterdam
1081 HV Amsterdam, The Netherlands

Abstract

We present the CLTL system developed for the DIMEMEX Shared Task on detecting fine-grained types of hate speech in Mexican Spanish memes. The competition consisted of two tasks. Task 1 involved classifying memes into hate speech, inappropriate, or harmless categories, while Task 2 required further classification of hateful memes into classism, sexism, racism, or other types. We explored the effectiveness of combining state-of-the-art language models with the Swin Transformer-based visual model to create a multimodal system using the Multilayer Perceptron fusion module for classification. Our experiments demonstrated that the XLM-T model combined with Swin Transformer V2 achieved the highest results, with an F1 score of 57.88 for Task 1 and 43.65 for Task 2, ranking 1st in both tasks in the competition.

Keywords

Multimodal Hate Speech Detection, Hateful Memes Detection, Mexican Spanish

1. Introduction

Memes have become a prevalent form of online expression, effectively conveying complex ideas through shareable and engaging content that combines images with concise text. While often humorous and entertaining, memes are commonly used to spread hate speech and reinforce harmful stereotypes [1]. The rapid growth of social media platforms has exacerbated the dissemination of hateful memes, prompting increased research into detecting hatefulness expressed in such content, including fine-grained types of inappropriate multimodal content such as hateful [2], misogynistic [3], harmful [4], and offensive [5]. However, most of the previous studies have primarily focused on English-language memes, with only a few extending their research to less-resourced languages, e.g., [6]. This gap is particularly evident in the context of Mexican Spanish. Although Mexico has the largest number of Spanish speakers in the world (125.95 million) [7], Mexican Spanish is considered one of the notable resource-constrained languages [8]. The scarcity of large annotated datasets and domain-specific models poses significant challenges for multimodal content moderation in Mexican Spanish.

To bridge this gap, the DIMEMEX challenge [9] at IberLEF 2024 [10] provides a platform for developing and evaluating models capable of detecting hateful memes in Mexican Spanish. This challenge includes two tasks aimed at (1) determining the presence of hate speech and inappropriate content in memes, and (2) classifying hateful memes into fine-grained classes: classism, sexism, racism, and others. To the best of our knowledge, this is the first shared task on hateful memes detection in Mexican Spanish.

We carried out a variety of experiments using state-of-the-art Spanish language models and multilingual language models combined with the visual model Swin Transformer V2 [11] to extract features for encoding textual and visual cues from memes. These features were then concatenated via the Multilayer Perceptron (MLP) fusion module for multimodal classification. Without the need for text preprocessing and feature engineering, our system achieved first place in both tasks, outperforming the second-runner by 2 F1 points in Task 1 and 7 F1 points in Task 2.

IberLEF 2024, September 2024, Valladolid, Spain
y.wang11@student.vu.nl (Y. Wang); i.markov@vu.nl (I. Markov)

2. Task Description & Dataset

The DIMEMEX challenge comprises two tasks:

2.1. Task 1. Detection of Hate Speech, Inappropriate, and Harmless Memes

This task involves a three-way classification, where each meme belongs to one of the following classes: hate speech, inappropriate, or harmless content. The description of each of the categories is provided below:

- **Hate speech**: The meme presents *"Any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group based on who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factors."* [12]
- **Inappropriate content**: The meme presents any kind of manifestation of offensive, vulgar (profane, obscene, sexually charged) and/or morbid humor content.
- **Harmless**: The meme does not contain any form of hate speech or inappropriate content.

2.2. Task 2. Finer-Grained Detection of Hate Speech in Memes

This task entails a finer-grained classification, where each meme belongs to one of the following types of hate speech: classism, sexism, racism, or others. The description of each type is provided below:

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

2.3. Dataset Description

The DIMEMEX dataset used in the shared task consists of around 2,200 memes compiled from public Facebook groups rooted in Mexico. The dataset is annotated for the hate speech categories described above by at least three annotators. Each item in the dataset comprises an original image file of the meme alongside its extracted textual content. We used the training data provided by the organizers and randomly split it into 85% for training and 15% for evaluating our approaches. The statistics of the training and evaluation sets used in this work are provided in Tables 1 and 2. It can be observed that the dataset is highly imbalanced in terms of the represented classes, with the hateful memes constituting the minority class with 17.06% of the entire dataset.

3. Methodology

We conducted comprehensive experiments to evaluate the effectiveness of several state-of-the-art language models combined with the Swin Transformer V2 vision model [11] for extracting contextualized embeddings from textual and visual inputs, respectively. These embeddings were then concatenated using the Multilayer Perceptron (MLP) fusion module [13] with a prediction layer on top to classify each instance into one of the predefined categories, as illustrated in Figure 1. This strategy has demonstrated state-of-the-art performance in several multimodal classification tasks, such as detecting propagandistic

Table 1
Statistics of the training and evaluation sets used for Task 1.

Label	Total	Train		Eval	
		# Num	%	# Num	%
Hate Speech	**386**	328	17.06	58	17.06
Inappropriate Content	**472**	401	20.85	71	20.88
Harmless	**1,405**	1,194	62.09	211	62.06
Total	**2,263**	**1,923**	**100**	**340**	**100**

Table 2
Statistics of the training and evaluation sets used for Task 2.

Label	Total	Train		Eval	
		# Num	%	# Num	%
Classicism	**44**	37	1.9	7	2.1
Racism	**114**	97	5	17	5
Sexism	**156**	133	6.9	23	6.8
Other	**72**	61	3.2	11	3.2
Inappropriate Content	**472**	401	20.85	71	20.88
Harmless	**1,405**	1,194	62.09	211	62.06
Total	**2,263**	**1,923**	**100**	**340**	**100**

memes in Arabic [14] and identifying multimodal hate speech related to the Russia-Ukraine conflict in English [15]. We did not apply any preprocessing steps, directly feeding the input representations provided with the dataset into the transformer models. All models were fine-tuned on the training set and evaluated on the evaluation set as described in Section 2.3.

Our experiments were conducted on the Google Colaboratory platform with an NVIDIA L4 GPU. We used the PyTorch framework and the AutoGluon library [13] for our implementation. We set uniform hyperparameters for all the examined models: a base learning rate of 1e-4, a decay rate of 0.9 using cosine decay scheduling, a batch size of 8, a maximum of 10 training epochs, and optimization via the AdamW optimizer.

3.1. Language Models

Our approach involves freezing the vision model (Swin Transformer V2) and testing a variety of language models to evaluate the effectiveness of our multimodal classification system on both tasks in the competition. Specifically, we examined the following categories of language models:

3.1.1. Cross-lingual language models

- **XLM-T** [16]: a multilingual transformer model specifically designed for processing social media texts, which includes informal language, slang, and mixed languages. It leverages the structure of the XLM-RoBERTa-Large and was re-trained on more than 1 billion tweets in diverse languages up to December 2022.
- **Multilingual-E5** [17]: a transformer-based language model designed for multilingual applications. It extends the XLM-RoBERTa-Large model by incorporating weakly-supervised contrastive pre-training on 1 billion multilingual text pairs.

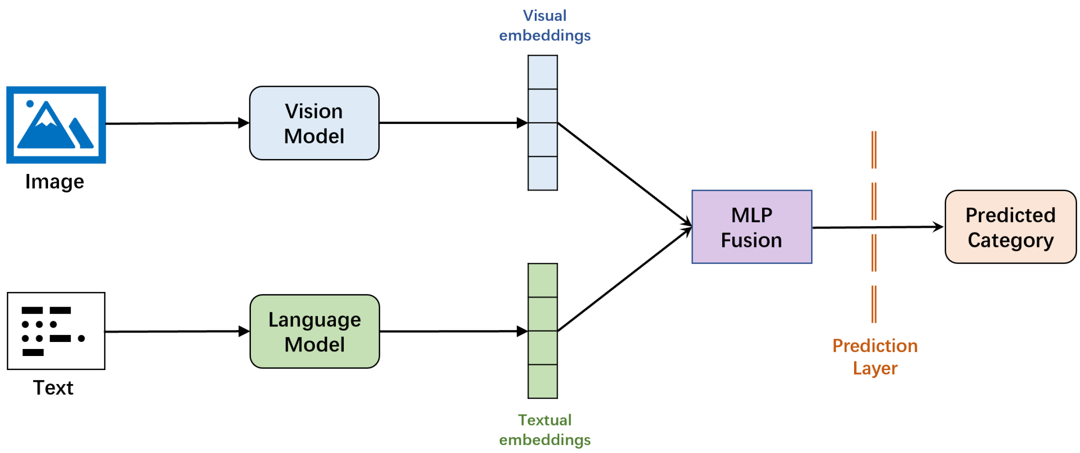

Figure 1: An overview of the multimodal classification system.

3.1.2. Monolingual Spanish language models

- **RoBERTa-base-BNE** [18]: a Spanish language model based on the RoBERTa architecture [19]. The model was pre-trained using the largest Spanish corpus known to date, with a total of 570 GB of clean and deduplicated texts compiled from the web crawling of the National Library of Spain from 2009 to 2019.
- **BETO** [20]: a BERT-based language model pre-trained exclusively on a large corpus from various sources like Wikipedia and the OPUS Project. For our experiments, we used a specific version of the BETO model designed for sentiment analysis in Spanish.[1]

3.2. Vision Model

The vision model used to process meme images is Swin Transformer V2 [11]. This model builds on the original Swin Transformer architecture [21] and introduces a window-based attention mechanism for efficient image processing across various scales and resolutions. It partitions the image into non-overlapping patches and processes these sequentially at each stage. Additionally, it employs a self-supervised pre-training method to reduce the need for extensive labeled data. The combination of this vision model with other language models has achieved state-of-the-art performance in the multimodal hate speech detection task related to the Russia-Ukraine conflict in English [15].

3.3. Multimodal Fusion & Prediction

We utilized a multimodal architecture that combines visual and language models, serving as image and text encoders to extract respective embeddings from original images and texts of memes. The extracted embeddings were then concatenated using the Multilayer Perceptron (MLP) fusion module [13], where the top vector representations from the different models are combined into a single vector. Following this, a prediction layer was added to classify each instance into one of the predefined categories for Tasks 1 and 2.

4. Results

We evaluated the performance of various multimodal models on both the evaluation and test sets for Task 1 and Task 2. The macro-averaged F1 score was used as the official evaluation metric for the both tasks.

[1]https://huggingface.co/ignacio-ave/beto-sentiment-analysis-spanish

Table 3
Results for Task 1 on the evaluation and test sets.

Set	Multimodal model	macro-F1
Eval	XLM-T + Swin Transformer V2	**58.71**
	Multilingual-E5 + Swin Transformer V2	56.11
	BETO + Swin Transformer V2	53.87
	RoBERTa-base-BNE + Swin Transformer V2	52.73
Test	XLM-T + Swin Transformer V2	57.88

Table 4
Results for Task 2 on the evaluation and test sets.

Set	Multimodal model	macro-F1
Eval	XLM-T + Swin Transformer V2	**39.20**
	RoBERTa-base-BNE + Swin Transformer V2	34.64
	Multilingual-E5 + Swin Transformer V2	33.74
	BETO + Swin Transformer V2	29.65
Test	XLM-T + Swin Transformer V2	43.65

The results for Task 1 are presented in Table 3. The combination of XLM-T and Swin Transformer V2 achieved the highest macro-averaged F1 score of 58.71 on the evaluation set, outperforming the second-best model (Multilingual-E5 + Swin Transformer V2) by a margin of 2.6 F1 points. On the test set, the combination of XLM-T and Swin Transformer V2 models maintained its superior performance with an F1 score of 57.88.

Table 4 shows the results for Task 2. The XLM-T combined with Swin Transformer V2 once again achieved the highest F1 score of 39.2 on the evaluation set, with a substantial increase in performance compared to other models. The performance of this combination was substantially better on the test set, with an F1 score of 43.65.

The results suggest that multilingual models (XLM-T and Multilingual-E5), when combined with the Swin Transformer V2 visual model, perform consistently better than monolingual Spanish models (BETO and RoBERTa-base-BNE) in most cases. This may be attributed to the peculiarities of the task covering Mexican Spanish, as the monolingual models were primarily trained on European Spanish data and may not generalize well to the Mexican Spanish variety. Further investigation is required to verify this conclusion.

To further analyze the performance of the best model combination (XLM-T + Swin Transformer V2), Figure 2 presents the confusion matrices for Task 1 and Task 2 on the evaluation set. For Task 1, the model shows the best performance in detecting "Harmless" memes (label 2), with the vast majority of true "Harmless" instances being correctly classified (180 out of 211). However, there is some noteworthy confusion between "Hate Speech" (label 0) and "Inappropriate Content" (label 1). For Task 2, the model again performed best in detecting "Harmless" memes (label 5) with 172 correct predictions out of 191. There is a notable confusion in identifying "Racism" (label 1) and "Sexism" (label 2), where instances are often misclassified into other categories, such as "Other" (label 3) and "Inappropriate Content" (label 4). These observations highlight the model's strengths and areas for improvement, particularly in distinguishing between specific fine-grained types of hate speech in memes.

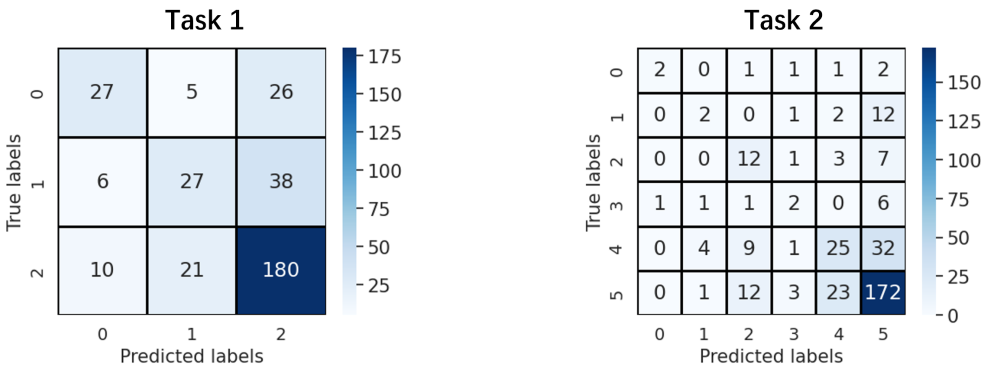

Figure 2: Confusion matrices for the best-performing model (XLM-T + Swin Transformer V2) on the evaluation set for Task 1 and Task 2.

5. Conclusion

In this study, we presented the CLTL system developed for fine-grained hate speech detection in Mexican Spanish memes at the DIMEMEX Shared Task. We evaluated several state-of-the-art language models, including XLM-T, Multilingual-E5, RoBERTa-base-BNE and BETO, in combination with the Swin Transformer V2 visual model to extract textual and visual embeddings, which were then fused using the Multilayer Perceptron module for classification. The results showed that the combination of XLM-T and Swin Transformer V2 models achieved first rank on the leaderboard for both tasks, with macro-averaged F1 scores of 57.88 and 43.65 on the official test sets for Task 1 and Task 2, respectively. One of the directions for future work would be to include image captioning features to potentially further enhance the multimodal system's ability to identify the fine-grained categories of hate speech in Mexican Spanish memes.

References

[1] T. Chakraborty, S. Masud, Nipping in the bud: detection, diffusion and mitigation of hate speech on social media, SIGWEB Newsl. 2022 (2022). URL: https://doi.org/10.1145/3522598.3522601. doi:10.1145/3522598.3522601.

[2] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The Hateful Memes Challenge: Detecting hate speech in multimodal memes, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 2611–2624. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1b84c4cee2b8b3d823b30e2d604b1878-Paper.pdf.

[3] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, J. Sorensen, SemEval-2022 task 5: Multimedia automatic misogyny identification, in: G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, S. Ratan (Eds.), Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 533–549. URL: https://aclanthology.org/2022.semeval-1.74. doi:10.18653/v1/2022.semeval-1.74.

[4] S. Pramanick, D. Dimitrov, R. Mukherjee, S. Sharma, M. S. Akhtar, P. Nakov, T. Chakraborty, Detecting harmful memes and their targets, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 2783–2796. URL: https://aclanthology.org/2021.findings-acl.246. doi:10.18653/v1/2021.findings-acl.246.

[5] S. Suryawanshi, B. R. Chakravarthi, M. Arcan, P. Buitelaar, 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, D. Kadar (Eds.), Proceedings of the Second Workshop on Trolling, Ag-

gression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 32–41. URL: https://aclanthology.org/2020.trac-1.6.

[6] S. Suryawanshi, B. R. Chakravarthi, P. Verma, M. Arcan, J. P. McCrae, P. Buitelaar, A dataset for troll classification of TamilMemes, in: G. N. Jha, K. Bali, S. L., S. S. Agrawal, A. K. Ojha (Eds.), Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 7–13. URL: https://aclanthology.org/2020.wildre-1.2.

[7] S. R. Department, Countries with the largest number of native spanish speakers worldwide in 2022, 2024. URL: https://www.statista.com/statistics/991020/number-native-spanish-speakers-country-worldwide/.

[8] A. Guevara-Rukoz, I. Demirsahin, F. He, S.-H. C. Chu, S. Sarin, K. Pipatsrisawat, A. Gutkin, A. Butryna, O. Kjartansson, Crowdsourcing Latin American Spanish for low-resource text-to-speech, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 6504–6513. URL: https://aclanthology.org/2020.lrec-1.801.

[9] H. Jarquín Vásquez, I. Tlelo-Coyotecatl, I. Hernández Farías, M. Casavantes, H. J. Escalante, L. Villaseñor-Pineda, M. Montes y Gómez, Overview of DIMEMEX at IberLEF 2024: Detection of Inappropriate Memes from Mexico, in: Procesamiento del Lenguaje Natural, September, 2024.

[10] L. Chiruzzo, S. M. Jiménez-Zafra, F. Rangel, Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS.org, 2024.

[11] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, B. Guo, Swin Transformer V2: Scaling up capacity and resolution, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11999–12009. doi:10.1109/CVPR52688.2022.01170.

[12] U. Nations, Understanding hate speech | united nations, 2023. URL: https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech.

[13] X. Shi, J. Mueller, N. Erickson, M. Li, A. Smola, Multimodal AutoML on structured tables with text fields, in: 8th ICML Workshop on Automated Machine Learning (AutoML), 2021. URL: https://openreview.net/forum?id=OHAIVOOl7Vl.

[14] Y. Wang, I. Markov, CLTL at ArAIEval shared task: Multimodal propagandistic memes classification using transformer models, in: The Second Arabic Natural Language Processing Conference (ArabicNLP 2024), Association for Computational Linguistics, Bangkok, 2024.

[15] Y. Wang, I. Markov, CLTL@Multimodal hate speech event detection 2024: The winning approach to detecting multimodal hate speech and its targets, in: A. Hürriyetoğlu, H. Tanev, S. Thapa, G. Uludoğan (Eds.), Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024), Association for Computational Linguistics, St. Julians, Malta, 2024, pp. 73–78. URL: https://aclanthology.org/2024.case-1.9.

[16] F. Barbieri, L. Espinosa Anke, J. Camacho-Collados, XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 258–266. URL: https://aclanthology.org/2022.lrec-1.27.

[17] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei, Multilingual E5 text embeddings: A technical report, arXiv/2402.05672 (2024). URL: https://doi.org/10.48550/arXiv.2402.05672. doi:10.48550/ARXIV.2402.05672. arXiv:2402.05672.

[18] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del Lenguaje Natural (2022) 39–60. URL: https://doi.org/

10.26342/2022-68-3. doi:10.26342/2022-68-3.

[19] 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/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[20] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, 2023. arXiv:2308.02976.

[21] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin Transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10012–10022. URL: https://openaccess.thecvf.com/content/ICCV2021/html/Liu_Swin_Transformer_Hierarchical_Vision_Transformer_Using_Shifted_Windows_ICCV_2021_paper.html.