Procesamiento del Lenguaje Natural, Revista nº 73, septiembre de 2024, pp. 393-405 recibido 07-07-2024 revisado 17-07-2024 aceptado 25-07-2024

Overview of HOMO-MEX at IberLEF 2024: Hate Speech Detection Towards the Mexican Spanish speaking LGBT+ Population

Resumen de HOMO-MEX en IberLEF 2024: Detección de discursos de odio hacia la población LGBT+ hispanohablante mexicana

Helena Gómez-Adorno,[1] Gemma Bel-Enguix,[2,6] Hiram Calvo,[5] Sergio Ojeda-Trueba,[2] Scott Thomas Andersen,[3] Juan Vásquez,[4] Tania Alcántara,[5] Miguel Soto,[5] Cesar Macias,[5]

[1]Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas, Universidad Nacional Autónoma de México
[2]Instituto de Ingeniería, Universidad Nacional Autónoma de México
[3]Posgrado en Ciencia e Ingeniería de la Computación, Universidad Nacional Autónoma de México
[4]University of Colorado Boulder, Department of Computer Science
[5]Instituto Politécnico Nacional, Centro de Investigación en Computación
[6]Facultat de Filologia i Comunicació, Universitat de Barcelona
helena.gomez@iimas.unam.mx, {gbele, sojedat}@iingen.unam.mx
stasen@comunidad.unam.mx, juan.vasquez-1@colorado.edu
{hcalvo, talcantaram2020, msotoh2021, cmaciass2021}@cic.ipn.mx

Abstract: We present the HOMO-MEX shared task organized at IberLEF 2024, as part of the 40th. International Conference of the Spanish Society for Natural Language Processing (SEPLN 2024). The aim of this task is to promote the development of natural language processing systems capable of detecting and classifying LGBT+phobic content in Mexican-Spanish digital posts and song lyrics. HOMO-MEX 2024 is composed of three subtasks: Task 1 on LGBT+phobia detection on social media posts, Task 2 on fine-grained phobia identification, and Task 3 on LGBT+phobia detection on song lyrics. In this second edition of HOMO-MEX, 40 participants registered on our Codabench platform. Subtask 1 received 19 submissions, subtask 2 received 10 submissions, and Subtask 3 got 17 submissions. Finally, 11 teams presented papers describing their systems. Most systems used transformer-based approaches to tackle the task, while the best-performing teams included data augmentation and preprocessing techniques.

Keywords: hate speech, LGBT+phobia, machine learning, song lyrics.

Resumen: Presentamos la tarea compartida HOMO-MEX organizada en IberLEF 2024, como parte de la 40to. Congreso Internacional de la Sociedad Española de Procesamiento del Lenguaje Natural (SEPLN 2024). El objetivo de esta tarea es promover el desarrollo de sistemas de procesamiento del lenguaje natural capaces de detectar y clasificar contenido LGBT+fóbico en publicaciones y letras de canciones en español. HOMO-MEX 2024 se compone de tres subtareas: Tarea 1 sobre detección de fobia hacia comunidades LGBT+ en publicaciones en redes sociales, Tarea 2 sobre identificación de fobias de grano fino y Tarea 3 sobre detección de fobia hacia comunidades LGBT+ en letras de canciones. En esta edición, 40 participantes se registraron en la plataforma Codabench. Para la subtarea 1 recibimos 19 predicciones, para la subtarea 2 recibimos 10 y para la subtarea 3 recibimos 17. Finalmente, 11 equipos presentaron artículos describiendo sus sistemas. La mayoría de los sistemas utilizaron Transformers para abordar la tarea, y los equipos con mejor desempeño incluyeron técnicas de preprocesamiento y aumento de datos.

Palabras clave: discurso de odio, LGBT+fobia, aprendizaje de máquina, letras de canciones.

ISSN 1135-5948 DOI 10.26342/2024-73-30

1 Introduction

Hate speech detection is a prevailing challenge in the field of Natural Language Processing (NLP). Previous studies have often focused on the use of slurs to detect aggressiveness (ElSherief et al., 2021), while others have focused on the terms used to refer to the affected group (Vásquez et al., 2023), emphasizing the contextual implications of specific terms. However, there is a notable lack of research in Mexican Spanish that examines semantic bias across different text types containing hate speech, such as songs. An example of this is the work of (Calderon-Suarez et al., 2023), who compiles lyrics that show abusive and explicit words against women.

Several shared tasks have been dedicated to hate speech detection, but in a broad sense. For example, the task of detecting hate speech spreaders on *Twitter* (Bevendorff et al., 2021) and hate speech detection in EVALITA 2020 (Sanguinetti et al., 2020). Other tasks aim at identifying sexism online; for example, EXIST: sEXism Identification in Social Networks aims to combat sexism in social networks (Plaza et al., 2023), and EDOS: Explainable Detection of Online Sexism aims to detect sexist content in online platforms (Kirk et al., 2023).

Hate speech and discrimination manifest differently for various marginalized groups. While existing shared tasks have addressed hate speech in general or focused on sexism, homophobia is a different issue. Having a dedicated task allows for a more precise examination of homophobic content, which may have unique linguistic characteristics and societal implications. Besides, different forms of hate speech require tailored solutions.

The first edition of HOMO-MEX (Bel-Enguix et al., 2023) was presented at IberLEF 2023 (Jiménez-Zafra, Rangel, and Montes-y-Gómez, 2023; Montes-y-Gómez et al., 2023) and consisted of two subtasks: i) LGBT+phobia detection to determine if a tweet contained or not LGBT+phobic content, and ii) fine-grained LGBT+phobia identification to determine the specific group being the target of the hate speech. Subtask 1, LGBT+phobia detection, involved a three-class classification with 7 participating teams obtaining a maximum F1-score of 0.843. However, subtask 2, which entailed multi-label identification, received lower performance scores with 6 participating teams.

In this paper, we present the second edition of the HOMO-MEX shared task organized in the framework of the IberLEF 2024 (Chiruzzo, Jiménez-Zafra, and Rangel, 2024), which is aimed at detecting LGBT+phobia on X (*Twitter*) and also in song lyrics. It comprises the same subtasks as last year, plus a new subtask to detect hate speech against the LGBT+ community in song lyrics written in Spanish.

The paper is structured as follows: In Section 2, we provide a detailed description of the creation of both datasets for the tasks: tweets and songs. Section 3 presents the various approaches employed by the participants to address the tasks. The analysis of the results is discussed in Section 4, and in Section 5, we highlight and summarize the notable findings. This section also includes a discussion on the implications of the results and potential directions for future research in detecting and combating LGBT+ phobia on social media.

2 HOMO-MEX 2024 Corpus and Evaluation Framework

For the HOMO-MEX share task at IberLEF 2024 (Chiruzzo, Jiménez-Zafra, and Rangel, 2024), we collected a new corpus for LGBT+phobic detection in song lyrics for subtask 3. However, for subtasks 1 and 2 we used the same corpus of the first edition of HOMO-MEX (Bel-Enguix et al., 2023) but with different partitions. Such corpus is composed of tweets written in Mexican Spanish that contain nouns indicative of the LGBT+ collective. These nouns include slang, slurs, and general terminology used to name the members of the LGBT+ collective. This lexicon and their approximate translation are available in the project's Github[1]. The details about the annotation process of the HOMO-MEX corpus can be found in (Vásquez et al., 2023).

In this 2024 edition, we redistributed the train and test subsets of the HOMO-MEX corpus with the intention of improving the distribution of classes in each subset. For this edition of HOMO-MEX, the training subset contained 80% of the annotated tweets and the test subset the other 20%, of a total of 11,000 tweets. For the correct redistribution of class labels, the 80-20 partition was re-

[1] https://github.com/juanmvsa/HOMO-MEX

tained for each of the labels of the training and test subsets respectively. The final version of this corpus is in our Github[2].

For subtask 3 we developed the HOMO-LYRICS corpus (Soto et al., 2024), which is composed of song lyrics written in Spanish that could contain LGBT+phobic text or not (not phobic). Mostly, LGBT+phobia can be found in two different ways: directly via the use of textual terminology, slang, and general slurs to refer to members of the community and indirectly, via the use of semantic relationships, like metaphors. With the help of metaphors, texts (like lyrics) translate the straight meaning of a voice to a figurative one.

We compiled the songs on three fronts:

- Using the LyricScraper[3] library with the LGBTQ+ HOMO-MEX lexicon.
- Using the Genius[4] API with a Spanish selection; and
- Manually, with selections made by both the LGBTQ+ community and non-community members.

The lexicon of nouns, inflections, such as diminutives and augmentatives, and slang used to filter the lyrics were the same as the ones used in the compilation of the HOMO-MEX corpus (Vásquez et al., 2023) (subtasks 1 and 2). With this process, we extracted 1230 lyrics, of which more details can be found in our Github[5] repository.

Afterward, the songs were labeled to detect LGBT+phobia and non-LGBT+phobia. The annotation process for the lyrics followed a methodology similar to that employed in (Vásquez et al., 2023) and consisted of labeling each lyric in two classes: phobic and not phobic. Initially, the annotators identified the specific lines where LGBT-phobic content was present; it could be more than one line. Then, they selected the label indicating the presence of LGBT+phobia in the lyrics. If the lyrics did not contain any form of such phobia, the annotators selected the label indicating its absence.

Each song lyric was annotated by five annotators. We selected annotators who self-identified as members of the LGBT+ community and non-members. Table 1 shows the demographics of the HOMO-LYRICS corpus as self-identified by the annotators. Once the annotation process was completed, we selected the class label with the majority vote for each lyric. If it had more than 3 votes in the phobic class, we considered it as such. In cases of ties and those that had 2 votes for phobia, we manually verified the lyrics to select the final label. Finally, the labeled corpus was partitioned to create the training and testing datasets. In the same way, as in subtasks 1 and 2, we used 20% for testing and 80% for training.

Categories	Data
Age	22-42 years
Gender Identity	1 non-binary 1 agender 11 women 10 men
Sexual Orientation	10 LGBTQ+ 13 Heterosexual
Native Language	Spanish
Residence	México City
Education Level	1 Undergraduate 17 Graduate 5 Postgraduate

Table 1: Annotators' demographics.

The label distributions of the training and testing subsets for each subtask are shown in Tables 2, 3, and 4.

Set	LP	NLP	I	Total
Train	1,072	5,482	2,246	8,800
Test	267	1,371	562	2,200
Total	1,339	6,853	2,808	11,000

Table 2: Size and label distribution for the LGBT+Phobia detection subset.

Set	L	G	B	T	O	Total
Train	88	894	10	94	77	1163
Test	18	234	3	23	19	297
Total	106	1,128	13	117	96	*X*

Table 3: Size and label distribution for the fine-grained classification subset.

3 Overview of the Submitted Approaches

Twelve teams submitted their working notes detailing their approaches to this shared task,

[2] https://github.com/GIL-UNAM/HOMO-MEX

[3] https://www.azlyrics.com/

[4] https://genius.com/

[5] https://github.com/CCogS-Mx/Spanish-lyrics-dataset-for-LGBT-phobia-screening

Set	LP	NLP	Total
Train	39	945	984
Test	10	236	246
Total	49	1,181	1,230

Table 4: Size and label distribution for the LGBT+Phobia detection in Lyrics subset.

but only 11 were accepted for publication. Eleven teams participated in Task 1, six in Task 2, and ten in Task 3.

Table 5 shows the approaches of each of the teams for all tasks.

Team	Traditional ML	LLM Prompting	Single Transformer	Ensemble Transformers
UC-CUJAE			*X*	
VEL	*X*			
LaboCIC			*X*	
jlpl1			*X*	
CANTeam		*X*		
HOMO-CIC			*X*	
I2C-UHU		*X*		*X*
VerbaNexAI			*X*	
LabTL-INAOE	*X*			
DSVS			*X*	
IntelliLeksika			*X*	
ColPos	*X*			

Table 5: General approach of each participating team for all tasks.

- *Team name: UC-CUJAE* (Hernández-González, Madera-Quintana, and Simón-Cuevas, 2024)
  - **Summary:** These participants present a single pre-trained encoder-based solution to each subtask of the HOMO-MEX task. The authors decided not to preprocess the texts for all three tasks since doing so reduced each model's performance. The authors opted for a single pre-trained model over an ensemble architecture in all three subtasks. For subtask 1, their best performing model was a fine-tuned version of the encoder `RoBERTa`; for subtask 2 and subtask 3, the `xlm-RoBERTa` variation was implemented instead of the base one. The authors also explored artificial data augmentation; however, they observed that data increases do not substantially improve the result, so they decided to use manual thresholds, which highlights an improvement in the results.

- *Team name: VEL* (Kayande et al., 2024)
  - **Summary:** The authors compared several classification algorithms, including XGBoost and LSTM trained on `BETO` word embeddings and fine-tuned configurations of `BETO`. The best results were found with a fully fine-tuned `BETO` model on validation data; however, it appears that this was fairly overfitted as XGBoost had the best results in test data. Their XGBoost-based method achieved the best macro F1-scores of 74.56 on subtask 1 and 47.44 on subtask 3, both on the test data.

- *Team name: LaboCIC* (Espinosa, Sidorov, and Ricárdez Vázquez, 2024)
  - **Summary:** This team participated in subtasks 1 and 3. They first evaluated traditional machine learning models trained on character and word n-grams. Then, they approach the task by finetuning different types of Bert-based models (`RoBerta`, `BERTweet`, `BERT`), achieving better results.

- *Team name: CANTeam* (Quan, Son, and Thin, 2024)
  - **Summary:** This team fine-tuned `Llama2` with the LoRA technique for all three subtasks. The authors described the prompt engineering design and observed during their experiments that providing more information, such as the label's description to the prompt, leads to better model performance. This approach ranked 1st. place in the Multi-label Fine-grained hate

speech detection subtask 2. The authors also presented comparison results with other transformer-based models such as `XLM RoBERTa` and `Multilingual T5`.

- *Team name: HOMO-CIC* (Vazquez et al., 2024)
  - **Summary:** The authors participated in subtask 1 and 3. They evaluated three transformer-based language models: `BERT`, `DistilBERT`, and `RoBERTa`. They used 80% of the data for training, and 20% was used to validate the transformer-based models. Their best results for subtask 1 was achieved with the `RoBERTa` and `BERT` models and for subtask 3 they only reported results with `RoBERTa`.
- *Team name: I2C-UHU* (Román-Pásaro et al., 2024)
  - **Summary:** This team participated in subtask 1 ranking 3rd. place. Their approach is based on the integration of Large Language Models (LLMs) for classification through prompting, alongside an ensemble of Transformers. The authors conducted an exhaustive search for hyperparameters to identify the optimal training parameters for the Transformer models specific to this subtask. Their final classification model combines the results of the best three Transformer models (`XLM`, `mDeBERTa` and `RoBERTa`) with two LLMs (two variants of `Falcon`) through hard voting.
- *Team name: VerbaNexAI* (Gonzalez-Henao et al., 2024)
  - **Summary:** This team ranked 1st. place in subtask 1 with an F1 score of 91%, 3rd. in subtask 2 with an F1 score 93%, and 2nd. in subtask 3 with an F1 score of 56%. The participants fine-tuned a single transformer-based language model, `bert-base-uncased`, which demonstrates that with good data quality and taking care of small details, this model can achieve high classification performance. To maintain good data quality, the authors performed extensive preprocessing that included removing URLs, punctuation marks, and special characters and numbers; they also lowercase the texts and transformed informal language, among others. The authors also performed data augmentation and used a k-fold cross-validation approach to train the `BERT` model.
- *Team name: LabTL-INAOE* (Ramírez-González, Hernández-Farías, and y Gómez, 2024)
  - **Summary:** This team performed an extensive search for the optimal representation for each task based on three different strategies and the use of multiple ML models. One representation included the distance among embeddings. They ranked in the top ten in all three tasks, demonstrating their approach's robustness.
- *Team name: DSVS* (Damián et al., 2024)
  - **Summary:** This team experimented with fine-tuning different Spanish Transformer models, varying the training parameters, and the pre-processing strategies. Their approach also included experimenting with weighted loss functions that were set according to the balance of classes in the dataset. Their approach showed remarkable results on subtasks 1 and 2, in which they ranked 4th. and 3rd., respectively. However, for subtask 3 they ranked 9th. due to time constraints, which prevented them from further improving their model.
- *Team name: IntelliLeksika* (Ramos et al., 2024)
  - **Summary:** This team participated only in subtask 3. They attempted three different preprocessing schemes ranging from 'no-preprocessing' to 'light-weight

preprocessing' to standard NLP preprocessing. They discuss several data representations and models. They used transformer-based approaches for embeddings with BETO and decision trees for classification. The results on the training data are impressive. However, this does not extend very well to the test data.

- *Team name: ColPos* (Ayala Niño, Montes-y-Gómez, and Velasco Cruz, 2024)
  - **Summary:** The authors use a weighted Naive Bayes classifier and the Multinomial variation of this algorithm on subtasks 1 and 3. The authors carried out preprocessing steps such as text normalization (conversion to lowercase), noise token elimination (removal of #, and), and the removal of emojis and "special characters". In order to deal with slang and out-of-vocabulary words, they implemented a search using a `fasText` model to obtain the most similar words within the vocabulary. The implemented weights to the Naive Bayes model were calculated using statistics between discrete variables. Similarly, documents were weighted using a modified word frequency value. This approach normally takes less computing than a deep learning approach, while keeping the performance the same or a bit lower than the baseline.

4 Experimental Evaluation and Analysis of the Results

This section provides a review of the results achieved by participants in the HOMO-MEX shared task, held at Iberlef 2024. We focus on analyzing and comparing the performance of the submitted solutions on the test set, using the macro F1-score as the primary performance metric. This metric was chosen due to its ability to balance precision and recall across different classes, making it particularly suitable for our multi-class classification tasks.

To facilitate the management of the shared task stages and the computation of performance metrics for all submissions, we utilized the Codabench platform [6].

As a baseline approach for our three tasks, we employed the `bert-base-spanish-wwm-cased` model (BETO) (Cañete et al., 2020). This model was chosen for its robust performance in natural language processing tasks involving the Spanish language. Before feeding the text into the model, we did not apply any pre-processing steps to the data. Instead, we focused on varying the number of fine-tuning epochs to observe how different levels of training impacted the performance. By using this approach, we aimed to establish a reference point against which the participants' solutions could be compared. Table 6 shows in detail the parameters used to fine-tune the BETO models for each subtask. In all cases, the validation partition corresponds to the 10% of the training set, the random seed was set to 42 and the batch size is 16.

Subtask	Parameters
Subtask 1	Epochs: 5 Learning rate: Default Epsilon: Default Max length: Default
Subtask 2	Epochs: 10 Learning rate: Default Epsilon: Default Max length: 256
Subtask 3	Epochs: 10 Learning rate: 1×10^{-7} Epsilon: 1×10^{-8}

Table 6: Parameters of BETO models used as the baseline in each subtask.

Table 7 presents an overview of the results achieved by each team. The table shows the evaluation metrics and ranking of each team for this task. Notably, in this edition of the HOMO-MEX shared task, the team VerbaNexAI(Gonzalez-Henao et al., 2024), demonstrated superior performance, surpassing all other approaches and the baseline model.

This edition of the shared task saw participation from 19 teams in subtask 1, each bringing unique methodologies to tackle the challenge of hate speech detection. We have only reported the results for the 12 teams that submitted their working notes, including one team that did not submit their fi-

[6] `https://www.codabench.org/competitions/2229/`

Team name	F1-score	Precision	Recall	Ranking
VerbaNexAI	0.9143	0.9364	0.8963	1
CANTeam	0.8775	0.9291	0.8477	2
I2C-UHU	0.8765	0.9098	0.8531	3
DSVS	0.8713	0.9195	0.8406	4
LabTL-INAOE	0.8563	0.8698	0.8458	5
UC-CUJAE	0.8560	0.9016	0.8258	6
jlpl1	0.8418	0.9064	0.8045	7
Homomex (baseline)	**0.8272**	**0.9074**	**0.7876**	**8**
HomoCIC	0.8219	0.8936	0.7857	9
ColPos	0.7911	0.8310	0.7632	10
LaboCIC	0.7796	0.8699	0.7397	11
VEL	0.7456	0.7962	0.7148	12

Table 7: Result summary for the HOMO-MEX shared task on Task 1.

nal version of the paper. The success of the presented solutions provides valuable insights into the development of more effective hate speech detection systems, emphasizing the importance of innovative approaches and thorough evaluation metrics.

For subtask 2 of the HOMO-MEX shared task, the results are shown in Table 8, which provides a summary of the results achieved by each participating team and our baseline model. This table includes the macro F1-score, hamming loss, and exact match ratio. In this task, the goal was to detect fine-grained categories of hate speech. The CANTeam (Quan, Son, and Thin, 2024) outperformed other participants, as did the baseline model.

In this edition we had 11 participating teams in subtask 2, each employing unique methodologies to tackle the challenge of fine-grained hate speech detection. However, only the results of the 6 teams who submitted their working notes are reported.

Table 9 provides a summary of the results achieved by each participating team and our baseline model in subtask 3 of the HOMO-MEX shared task. This table includes the macro scores for F1-score, precision, and recall. In this task, the goal was to detect homophobic content in lyrics. The team UC-CUJAE (Hernández-González, Madera-Quintana, and Simón-Cuevas, 2024) outperformed other participants.

In this edition we had 17 participating teams in subtask 3, each employing unique methodologies to tackle the challenge of homophobic lyrics detection. Only the outcomes from the 9 teams that provided their working notes have been included in this report.

4.1 Statistical Analysis

For the evaluation of the results obtained by the participating teams, we employed CompStats (Nava-Muñoz, Graff Guerrero, and Escalante, 2023), a Python library designed for evaluating the results achieved in shared tasks.

In Subtask 1, the algorithms demonstrated varying levels of effectiveness in detecting hate speech. The top-performing algorithms, such as those from VerbaNex AI Lab, CANTeam, and I2C-UHU, achieved the highest F1-macro-scores close to 0.90, with narrow confidence intervals indicating consistent performance. Middle-tier algorithms, including LabTL-INAOE and UC-CUJAE, had scores ranging from 0.85 to 0.90. The lower-performing algorithms, such as those from LaboCIC and VEL, scored around or below 0.80, showing greater variability in performance as shown in Figure 1.

For Subtask 3, the performance of algorithms varied significantly in their ability to detect homophobic lyrics. The top performers like UC-CUJAE, VerbaNex AI Lab, and LabTL-INAOE achieved high F1-macro-scores approaching 0.65, with narrow confidence intervals indicating reliable performance. Other algorithms, including ColPos and Homomex (baseline), showed much lower performance, with scores around 0.50 and very tight confidence intervals, suggesting consistent but less effective predictions (Figure 2).

This analysis highlights the varying effectiveness and reliability of the submitted systems across the different subtasks.

Team name	F1-score	Hamming Loss	Exact Match Ratio	Ranking
CANTeam	0.9730	0.0149	0.9291	1
Homomex (baseline)	**0.9488**	**0.0236**	**0.8843**	**2**
DSVS	0.9436	0.0342	0.8470	3
VerbaNexAI	0.9393	0.0299	0.8881	4
UC-CUJAE	0.9346	0.0404	0.8433	5
jlpl1	0.9322	0.0280	0.8993	6
LabTL-INAOE	0.9134	0.0367	0.8507	7

Table 8: Result summary for the HOMO-MEX shared task on Task 2.

Team name	F1-score	Precision	Recall	Ranking
UC-CUJAE	0.5762	0.5604	0.6513	1
VerbaNexAI	0.5683	0.5575	0.6843	2
LabTL-INAOE	0.5667	0.5598	0.5767	3
ColPos	0.4896	0.4797	0.5000	4
Homomex (baseline)	**0.4896**	**0.4797**	**0.5000**	**4**
HomoCIC	0.4896	0.4797	0.5000	4
CANTeam	0.4875	0.4795	0.4958	5
IntelliLeksika	0.4864	0.4794	0.4936	6
DSVS	0.4864	0.4794	0.4936	6
LaboCIC	0.4832	0.4792	0.4873	7
VEL	0.4744	0.4784	0.4703	8

Table 9: Result summary for the HOMO-MEX shared task on Task 3.

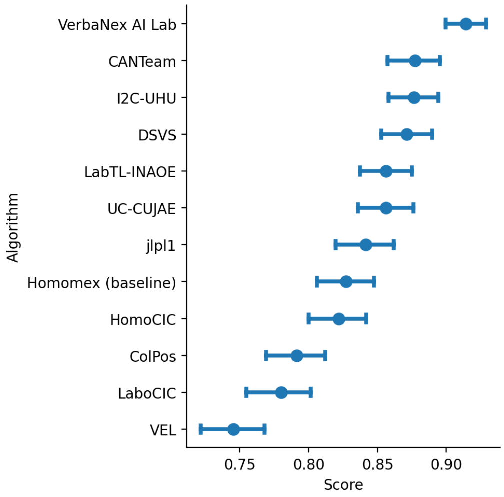

Figure 1: Classification metrics obtained in Subtask 1.

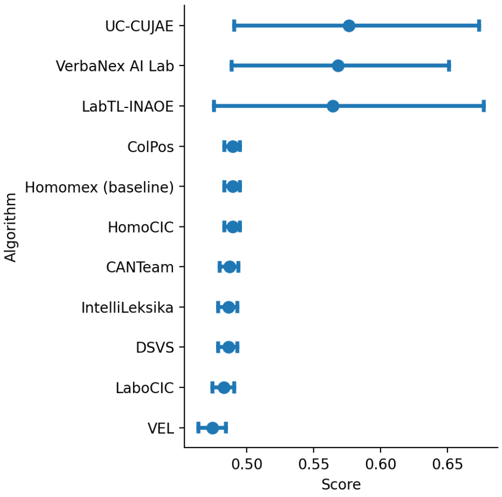

Figure 2: Classification metrics obtained in Subtask 3.

The `plot_difference` function from the CompStats library provides a visual representation of performance differences between the submitted approaches. By comparing each approach to the best-performing model in each subtask, we can identify which models perform similarly and which have significant gaps. The confidence intervals and statistical significance indicators ensure that the differences are observed and statistically validated. This comparison highlights the strengths and weaknesses of the participants' approaches.

In Subtask 1, VerbaNex AI Lab was identified as the best-performing algorithm. Figure 3 illustrates the performance differences between VerbaNex AI Lab and the other submissions. Most systems show significant differences, with notable gaps for teams like VEL and LaboCIC, which performed significantly worse. Teams like CANTeam and I2C-UHU exhibited smaller performance differences but were still significantly outperformed by VerbaNex AI Lab. The confidence intervals provide a clear indication of the reliability of these performance differences.

For Subtask 2, CANTeam was the top performer. Figure 4 shows the performance differences between CANTeam and other submissions. The results indicate that teams like LabTL-INAOE and UC-CUJAE had significant performance differences, with their scores being considerably lower. VerbaNexAI and our baseline had smaller performance gaps but still significant differences. The consistency in the performance difference is demonstrated by the narrow confidence intervals for most teams. Whereas, for Subtask 3, UC-CUJAE was the top performer. Figure 5 shows the performance differences between UC-CUJAE and other submissions. The results indicate that teams like VEL, Labo-CIC, and DSVS had significant performance differences, with their scores being considerably lower. VerbaNex AI Lab and LabTL-INAOE had smaller performance gaps but were still significantly outperformed by UC-CUJAE. The consistency in the performance differences is demonstrated by the narrow confidence intervals for most teams.

4.2 Maximum Possible Accuracy and Coincident Failure Diversity

To evaluate the complementary and diversity of predictions provided by different approaches we used the Maximum Possible Accuracy (MPA) and the Coincident Failure Diversity (CFD). MPA calculates the accuracy of classifications by determining the ratio of correctly classified instances to the total number of instances. For an instance to be considered correctly classified, at least one team must assign the correct label to it. This metric helps us identify instances that have been misclassified by all teams (Tang, Suganthan, and Yao, 2006).

The CFD metric, which ranges from 0 to 1, assesses the diversity among classifiers' predictions (Kuncheva and Whitaker, 2003). A CFD value of 0 indicates that all classifiers are either always correct or always incorrect, showing no diversity in their errors. Conversely, a CFD value of 1 signifies that, at most, one classifier will fail for any randomly chosen instance.

The MPA and CFD metrics results are presented in Table 10. The proposed approaches are grouped based on their methodology. We have created five methodological groups, four of which are described in table 5. The fifth group, called "All Teams", consists of all participating teams. Each group has a minimum of two members, ensuring sufficient representation and comparison among the approaches.

The result of all the approaches in the task (All Teams) consistently shows high MPA values across all subtasks, particularly excelling in subtask 3 with an MPA of 0.9878 and moderate CFD values, indicating balanced diversity in errors. The Single Transformer approaches also demonstrate high accuracy, especially in subtasks 1 and 3 with MPAs of 0.9777 and 0.9797, respectively, but with lower diversity in errors (CFD of 0.1752 and 0.1406). Traditional ML approaches maintain high accuracy in subtasks 1 and 3 (MPAs of 0.9555 and 0.9675) but exhibit the lowest accuracy in subtasks 2 (MPA of 0.8507).

The approaches with LLM prompting, although showing slightly lower MPAs compared to Single Transformers, exhibit extremely low diversity in errors, indicating consistent performance. The Ensemble Transformer approach, with limited data, shows lower accuracy in Subtask 1. Overall, the analysis indicates that while all approaches achieve high accuracy, the diversity of errors varies, with Single Transformer and Traditional ML approaches showing more consistent performance, and LLM prompting demonstrating the least diversity in errors.

5 *Conclusions*

This paper presents the proposed systems, and findings of the second edition of the HOMO-MEX shared task, held at the IberLef 2024. The main goal of this task was the detection of LGBT+ phobia in texts. This edition included three different subtasks. Two of them follow the same scheme as the previous edition of HOMO-MEX, and aim at detecting homophobia in X (*Twitter*) both in a binary classification and a more granular distinction between lesbian, gay, bisexual, and transgender phobia. A third task was added this year, and the goal was identification of homophobia in lyrics written in Spanish.

The results obtained this year in subtasks 1 and 2 show that the automatic identification of LGBT+phobia has made great

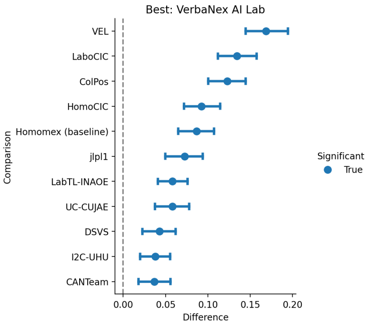

Figure 3: Difference in performance for Subtask 1.

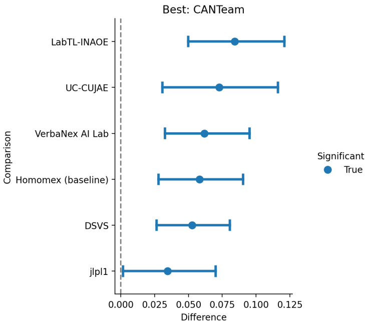

Figure 4: Difference in performance for Subtask 2.

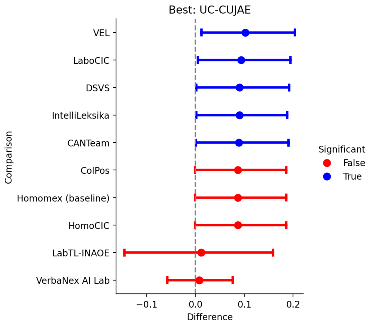

Figure 5: Difference in performance for Subtask 3.

Approach	Substask	MPA	CFD	Number of systems
All teams	1	0.9845	0.2923	11
	2	0.9626	0.1431	6
	3	0.9878	0.2278	10
Single Transformer	1	0.9777	0.1752	6
	2	0.9552	0.0941	4
	3	0.9797	0.1406	6
Ensemble Transformer	1	0.9209	-	1
	2	-	-	0
	3	-	-	0
LLM prompting	1	0.9441	0.0251	2
	2	0.9291	-	1
	3	0.9512	-	1
Traditional ML	1	0.9555	0.1185	3
	2	0.8507	-	1
	3	0.9675	0.0513	3

Table 10: MPA and CFD comparison results among the different proposed approaches for subtasks.

strides in NLP. However, task 3 is still an open topic since the best team reached just an F1 score of 0.57. The diverse results in the same task in two different communicative contexts are due to the language used in each. While lexical-based approaches are very successful in social networks, song lyrics are written with 'poetic' strategies, understood as the non-straight use of linguistic expressions. This implies that the methods of identification cannot be the same as the ones that have given very good results in other linguistic registers.

In general, most teams employed various transformer-based models in their classification pipelines. Despite this, three teams chose traditional ML classification methods. Finally, it is worth pointing out that two teams used prompt engineering as a novel approach. These latter results show that LLMs could be an effective tool for automatic hate speech detection.

LGBT+phobia detection in social networks is far from being a solved task. First, as demonstrated in Subtask 3, more experiments must be performed on other communicative situations, other than social media, such as literature, journalism, law, etc. Second, we are currently evaluating a connection between sentiment analysis and the use of LGBT+ terminology, to determine the contexts these terms are used. Related to this factor, we highlight the need to investigate the impact of the author in the statements, this is, whether the addresser belongs or not to the LGBT+ community can change the polarity and intention of the message. Additionally, further investigation is required to explore the potential biases introduced by annotator sociodemographic factors. Finally, in future corpora involving LGBT+ language, an extended vocabulary from other Spanish-speaking countries should be included in the filtering process.

Acknowledgements

This paper has been partially supported by CONAHCYT CF-2023-G-64 and by DGAPA-UNAM PAPIIT projects IN104424 and TA101722. The authors wish to thank the support of the Instituto Politécnico Nacional (COFAA, SIP-IPN, Grant SIP 20240610), and the Mexican Government (CONAHCyT, SNI). Also, we thank Alejandro Ojeda Trueba for the creation of the HOMO-MEX presentation image. GBE is supported by a grant from the Ministry of Universities of the Government of Spain, financed by the European Union, NextGeneration EU (María Zambrano program).

References

Ayala Niño, D., M. Montes-y-Gómez, and C. Velasco Cruz. 2024. Colpos at homo-mex 2024: Weighted naive bayes for lgbtq+phobia detection in spanish text. 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*.

Bel-Enguix, G., H. Gómez-Adorno, G. Sierra, J. Vásquez, S. T. Andersen, and S. Ojeda-Trueba. 2023. Overview of homo-mex at iberlef 2023: Hate speech detection in online messages directed towards the mexican spanish speaking lgbtq+ population. *Procesamiento del lenguaje natural*, 71:361–370.

Bevendorff, J., B. Chulvi, G. De La Peña Sarracén, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, et al. 2021. Overview of pan 2021: authorship verification, profiling hate speech spreaders on twitter, and style change detection. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings 12*, pages 419–431. Springer.

Calderon-Suarez, R., R. Ortega-Mendoza, M. Montes-y Gómez, C. Toxqui-Quitl, and M. Marquez-Vera. 2023. Enhancing the detection of misogynistic content in social media by transferring knowledge from song phrases. *IEEE Access*, 11:13179–13190.

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Chiruzzo, L., S. M. Jiménez-Zafra, and F. Rangel. 2024. 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*.

Damián, S., D. Vázquez, E. Felipe-Riverón, and C. Yáñez-Márquez. 2024. Dsvs at homo-mex24: Multi-class and multi-label hate speech detection using transformer-based models. 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*.

ElSherief, M., C. Ziems, D. Muchlinski, V. Anupindi, J. Seybolt, M. D. Choudhury, and D. Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. *CoRR*, abs/2109.05322.

Espinosa, D., G. Sidorov, and E. Ricárdez Vázquez. 2024. The labocic at homo-mex 2024: Using bert to classify hate-lgtb speech. 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*.

Gonzalez-Henao, R., D. Marrugo-Tobon, J. Martinez-Santos, and E. Puertas. 2024. Verbanexai lab at homo-mex 2024: Multi-class and multilabel detection of lgbtq+ phobic content using transformers. 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*.

Hernández-González, A., J. Madera-Quintana, and A. Simón-Cuevas. 2024. Uc-cujae at homo-mex 2024: Detecting hate speech against the lgtb+ community using transformers on imbalanced datasets. 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*.

Jiménez-Zafra, S., F. Rangel, and M. Montes-y-Gómez. 2023. Overview of IberLEF 2023: Natural Language Processing Challenges for Spanish and other Iberian Languages. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), co-located with the 39th Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), CEUR-WS.org*.

Kayande, D. D., K. K. Ponnusamy, P. K. Kumaresan, P. Buitelaar, and B. R. Chakravarthi. 2024. Vel at homo-mex 2024: Detecting lgbt+phobia in mexican spanish social media. 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*.

Kirk, H., W. Yin, B. Vidgen, and P. Röttger. 2023. SemEval-2023 task 10: Explainable detection of online sexism. In A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, and E. Sartori, editors, *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2193–2210, Toronto, Canada, July. Association for Computational Linguistics.

Kuncheva, L. and C. Whitaker. 2003. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51:181–207, 05.

Montes-y-Gómez, M., F. Rangel, S. Jimeńez-Zafra, M. Casavantes, B. Altuna, M. Álvarez Carmona, G. Bel-Enguix, L. Chiruzzo, I. de la Iglesia, H. Escalante, M. Garciá-Cumbreras, J. Garciá-Diáz, J. Gonzalez Barba, R. Labadie Tamayo, S. Lima, P. Moral, F. Plaza del Arco, and R. Valencia-Garciá, editors. 2023. *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), co-located with the 39th Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), CEUR-WS.org*.

Nava-Muñoz, S., M. Graff Guerrero, and H. J. Escalante. 2023. Comparison of Classifiers in Challenge Scheme. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13902 LNCS:89–98.

Plaza, L., J. Carrillo-de Albornoz, R. Morante, J. Gonzalo, E. Amigó, D. Spina, and P. Rosso. 2023. Overview of EXIST 2023: sexism identification in social networks. In *Proceedings of ECIR'23*, pages 593–599.

Quan, L. M., B. H. Son, and D. V. Thin. 2024. Canteam at homo-mex 2024: Hate speech detection towards the mexican spanish speaking lgbt+ population with large language model. 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*.

Ramírez-González, M., D. I. Hernández-Farías, and M. M. y Gómez. 2024. Labtl-inaoe at homo-mex 2024: Distance-based representations for lgbt+ phobia detection. 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*.

Ramos, L., C. Palma-Preciado, O. Kolesnikova, M. Saldana-Perez, G. Sidorov, and M. Shahiki-Tash. 2024. Intellileksika at homo-mex 2024: Detection of homophobic content in spanish lyrics with machine learning. 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*.

Román-Pásaro, J., A. Carrillo-Casado, J. Mata-Vázquez, and V. Pachón-Álvarez. 2024. I2c-uhu at homo-mex 2024: Leveraging large language models and ensembling transformers to identify and classify hate messages towards the lgbtq+ community. 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.*

Sanguinetti, M., G. Comandini, E. Di Nuovo, S. Frenda, M. Stranisci, C. Bosco, T. Caselli, V. Patti, and I. Russo. 2020. Haspeede 2@ evalita2020: Overview of the evalita 2020 hate speech detection task. *Evaluation Campaign of Natural Language Processing and Speech Tools for Italian.*

Soto, M., C. Macias, T. Alcántara, H. Calvo, S. Thomas Andersen, S. Ojeda-Trueba, G. Bel-Enguix, and H. a. Gómez-Adorno. 2024. CCogS-Mx/Spanish-lyrics-dataset-for-LGBTQ-phobia-screening: Spanish lyrics dataset for LGBTQ+phobia screening, July.

Tang, E., P. Suganthan, and X. Yao. 2006. An analysis of diversity measures. *Machine learning*, 65(1):247–271.

Vazquez, O. G., M. Cardoso-Moreno, J. A. Torres-León, and D. Jiménez. 2024. Homocic at homo-mex 2024: Deep learning approaches for classifying homophobic content in tweets and songs: Leveraging llm and nl. 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.*

Vásquez, J., S. T. Andersen, G. Bel-Enguix, H. Gómez-Adorno, and S.-L. Ojeda-Trueba. 2023. Homo-mex: A mexican spanish annotated corpus for lgbt+phobia detection on twitter. In *Proceedings of The 7th Workshop on Online Abuse and Harms (WOAH)*, Toronto, Canada, jul. Association for Computational Linguistics.

A Annex 1: Data Collection Ethics

HOMO-LYRICS corpus data usage statement: During this research, the legitimate rights of copyright owners, their agents, and representatives are respected. The data used in this study were obtained exclusively for academic and research purposes. No profit was made from the extraction or use of this data. Any further use of the data beyond the scope of this study will require appropriate consent and authorization from the data owners.

HOMO-MEX corpus: We collect tweets from the social media platform *X* using the *Twitter* API, now known as *X*. This API permits the collection of tweets that have been publicly posted. The authors of the tweets are not notified of their tweets participation in this study, however this process is in accordance to *X*'s privacy policy at the time. We ensure adherence to the requirements *X* sets for use of this API. The tweets collected are based on tagged metadata. All scraped tweets had geolocation tags in Mexico, and a language tag for Spanish. These tweets are supposedly provided to us randomly by the API, we assume a variety of author demographics are represented, such as variations in race, nationality, and socioeconomic background. However, these are not facts that we can verify.

For both corpus, we selected annotators that self identified as members and non-members of the LGBT+ community. They were informed of the purpose of the study and the harmful nature of some of the tweets they would be labeling, and were informed that they could stop participation in the study at any time if they did not wish to continue.