HULAT_UC3M at MiSonGyny 2025: Detecting Misogyny in Song Lyrics with Transformer Ensembles and Instruction-Tuned Generative Models

Juan-Sebastián Toledo, Isabel Segura-Bedmar

Human Language and Accessibility Technologies Group (HULAT), Computer Science and Engineering Department, Universidad Carlos III de Madrid, Leganés, 28911, Madrid, Spain

Abstract

This paper describes our participation in the MiSonGyny 2025 shared tasks, which focus on detecting misogynistic content in Spanish song lyrics. The competition featured two subtasks: binary detection of misogyny (task 1) and fine-grained misogyny speech classification (task 2). To address these challenges, we explored a range of models, including traditional machine learning classifiers, transformer-based encoders, and instruction-tuned generative models. For task 1, our top-performing system was a stacked ensemble combining Longformer, mDeBERTaV3, and RoBERTa-BNE, with logistic regression as the meta-learner. For task 2, the instruction-tuned Qwen3-14B generative model with LoRA achieved the best performance. On the final leaderboard, our systems ranked first in both tasks, obtaining a macro-F1 score of 0.88 for misogyny detection and a macro-F1 of 0.59 for fine-grained classification. These results demonstrate that transformer ensembles and instruction-tuned large language models can deliver state-of-the-art performance for detecting misogynistic speech in Spanish song lyrics.

Keywords

Misogyny detection, Transformers, LoRa, instruction tuning

1. Introduction

Misogyny, which literally means "hatred towards women", can be expressed through a wide range of social and cultural forms, including the perpetuation of male privilege, systemic gender-based discrimination, and violence against women [1]. Far from being detached from broader societal issues, music often reflects and amplifies misogynistic discourses, embedding harmful stereotypes and sexist messages within its lyrical content. This has a profoundly negative impact, as it directly contributes to the normalisation of misogynistic behaviours and beliefs within society, thereby perpetuating gender stereotypes and reinforcing inequality.

In recent years, there has been a growing interest in the field of Natural Language Processing (NLP) focused on the detection of sexist content on social media platforms [2, 3, 4]. However, the automatic identification of misogyny in song lyrics remains a largely underexplored area. Studies addressing this task in songs written in Spanish are even scarcer, despite the cultural significance and wide consumption of Spanish-language music. The IberLef-Misogyny 2025 shared task [5, 6] aims to address this gap by promoting research on NLP methods for the detection of misogyny in Spanish song lyrics.

For its first edition in 2025, MiSonGyny has proposed two tasks aimed at the automatic detection and classification of misogyny in Spanish song lyrics. The competition is structured around two subtasks. The first task is a binary classification problem, where systems must predict whether a given lyric contains misogyny speech or not. The second task extends the problem to a fine-grained classification challenge. Systems must assign one of four predefined labels that characterise different forms of misogynistic speech: Sexualisation (S), Violence (V), Hate (H), or Not Related (NR).

The objective of this paper is to present our participation in both tasks of MiSonGyny 2025 and to

IberLEF 2025, September 2025, Zaragoza, Spain

*Corresponding author.

100549168@alumnos.uc3m.es (J. Toledo); isegura@inf.uc3m.es (I. Segura-Bedmar)

https://hulat.inf.uc3m.es/nosotros/miembros/isegura (I. Segura-Bedmar)

0000-0001-5120-9486 (J. Toledo); 0000-0002-7810-2360 (I. Segura-Bedmar)

describe the models, training strategies, and evaluation results that led to our top-ranking submissions in both subtasks.

For task 1, we explored three independent strategies, each evaluated separately to identify the best performing model. First, (1) we trained four traditional machine learning classifiers: Support Vector Machines, Random Forest, Logistic Regression, and Gradient Boosting, to establish a performance baseline. Next, (2) we fine-tuned three BERT-based encoders: longformer-base-4096-bne-es [7], mDeBERTaV3 [8], and roberta-base-bne [9]. Finally, (3) we built a stacked ensemble using the three fine-tuned transformers from the previous approach, with a logistic regression as the meta-learner. The stacked ensemble in approach (3) achieved first place in task 1.

To tackle task 2, we also applied and evaluated three strategies: First, (1) we trained the same four traditional machine learning classifiers as in task 1, to provide a baseline. Then, (2) we instruction-tuned two autoregressive language models: Qwen3-14B [10] and LLaMA-3.1-8B [11], both using LoRa [12], a low-rank adaptation method that enables parameter-efficient updates while preserving the original weights. Finally, (3) we designed a hierarchical pipeline in which a Longformer model first filtered out non-misogynistic lyrics, and Qwen3 then performed three-class subtype classification on the remaining texts. The instruction-tuned Qwen3 model from approach (2) achieved the highest macro-F1 score on the official leaderboard, securing first place in task 2 and highlighting the effectiveness of our approach in this challenging and under-explored domain.

The rest of this paper is organised as follows. Section 2 details the dataset, data augmentation techniques, and the modelling approaches used for both tasks. Section 3 presents results and discusses the performance of each system. Conclusions are finally drawn in Section 4.

2. Approaches

This section presents our methodologies and design choices behind the models developed for tasks 1 and 2 in the MiSonGyny 2025 competition.

2.1. Dataset Overview

The dataset provided by the organisers consists of a collection of Spanish song lyrics from multiple musical genres. A team of 15 human annotators labelled these lyrics across two distinct phases, as defined by the shared task guidelines.

Table 1 summarises the main statistics of the dataset used for task 1. As shown, the dataset is imbalanced, with 642 misogynistic (M) and 1,462 non-misogynistic (NM) songs, resulting in an imbalance ratio of 2.28. Furthermore, misogynistic songs tend to be longer, with a median length of 683 tokens compared to 337 tokens for non-misogynistic lyrics. The 90th percentile values further highlight this disparity, with 90% of misogynistic songs being under 1,262 tokens compared to 696 tokens for non-misogynistic songs.

Table 1
Dataset statistics for task 1

	Non-Misogynistic	Misogynistic	Entire Dataset
Number of songs	1,462	642	2,104
Min token length	23	42	23
Max token length	2,571	4,513	4,513
90th percentile	696	1,262	970
Median token length	337	683	392

Table 2 presents the dataset statistics for task 2. As shown in the table, the class distribution is highly imbalanced, with the majority of songs labelled as 'Not Related' and 'Sexualisation', while 'Hate' and 'Violence' are underrepresented. Regarding song length, those classified as Sexualisation generally tend to be the longest (median length of 751 tokens), while those classified as 'Not Related' are the

shortest (median length of 346 tokens). The 'Violence' and 'Hate' categories fall in between, with median lengths of 558 and 452 tokens, respectively. This pattern is also evident in the 90th percentile values: 'Sexualisation', 'Violence', and 'Hate' reach 1,307, 1,141, and 1,093 tokens, respectively, whereas Not Related songs remain significantly shorter, with 90% of samples falling below 693 tokens.

Table 2
Dataset statistics for task 2

	Sexualisation	Violence	Hate	Non Related	Entire Dataset
Number of songs	435	129	78	526	1168
Min token length	96	42	125	40	40
Max token length	4,513	3,708	3,448	2,571	4,513
90th percentile	1,307	1,141	1,093	693	1,124
Median token length	751	558	452	346	490

During the development phase, preceding the final evaluation of the task, we applied a stratified split to divide the original dataset into training (70%), validation (10%), and testing (20%) partitions. This setup allowed us to develop various approaches and select the best-performing ones for the final evaluation.

2.2. Data Augmentation

Given the limited amount of labelled data available, we employed data augmentation techniques to expand the training set by generating new instances from existing ones, and ultimately improve the model's ability to generalise. Therefore, much of our effort was devoted to experimenting with different techniques and hyperparameter settings, continuously monitoring their impact on the validation macro-F1 score.

Although recent work has explored a wide range of state-of-the-art augmentation methods for English text, these studies offered little direct guidance for Spanish song lyrics [13]. One possible explanation may be the linguistic richness of the Spanish language, combined with the unique stylistic and structural properties inherent to song lyrics, such as poetic structure, figurative language, colloquialisms, slang, and frequent repetition, among others.

These characteristics reduce the transferability of text specific augmentation pipelines and motivated us to design and evaluate our own strategies, based on two sampling approaches: in the first (1), we oversampled the minority classes to approximately equalize the class distribution; in the second approach (2), we oversampled all classes proportionally to maintain the original distribution while increasing the total number of training instances by a predefined factor. Each modelling strategy was tested with both sampling approaches, and only the configuration that achieved the highest validation score was selected for submission.

Below, we detail the two augmentation techniques used in both tasks:

- **Back-translation (BT)**: This technique relies on translating the original Spanish lyric into a pivot language and then translating back to Spanish. BT is commonly used to generate paraphrased variants of the original input while preserving its semantic content [14]. According to [15], this method results in minimal label changes, with only 0.3% to 1.5% of the augmented texts displaying changes in meaning that could compromise label integrity. In our implementation, we utilised both the Google Cloud Translation API and the DeepL API, selecting English and Portuguese as intermediate languages. English was chosen due to the effectiveness of state-of-the-art machine translation systems for Spanish-English, which produce low-noise outputs with semantic quality comparable to average human translators [16]. Portuguese, on the other hand, was selected due to its linguistic proximity to Spanish, sharing approximately 89% of its vocabulary, while still introducing lexical variety [17].

- **Random insertion of punctuation marks**: As a second data augmentation technique, we applied AEDA (An Easier Data Augmentation) [18], which involves inserting punctuation marks from a predefined set: {".", "?", "!", ",", "’"}, at random positions within a given text. Unlike traditional augmentation techniques such as synonym replacement, random deletion, or random swap [19], which may lead to information loss or semantic distortion, AEDA preserves the original meaning of the text.
  This technique is particularly well suited to song lyrics, which frequently deviate from standard conventions of grammar, punctuation, and sentence structure [20]. By randomly inserting punctuation, we simulate the stylistic variability in song lyrics without altering the underlying semantics.
  In our implementation, punctuation was inserted at a frequency of 0.3 relative to the total token count in each input text, as proposed by [18]. Prior work in hate speech detection has shown that combining AEDA with BT improves macro-F1 in a multilingual sexism dataset [21], supporting its applicability in our tasks.

2.3. Task 1: Misogyny Speech Detection

2.3.1. Classical Machine Learning Classifiers

As an initial approach for both tasks, we trained a set of classical machine learning classifiers with TF-IDF bag-of-words features to establish a solid baseline for later Transformer models. Because TF-IDF imposes no sequence-length limit, every song lyric is represented in its entirety, in contrast to the 512-token truncation of standard BERT-based encoders.

The preprocessing pipeline applied to the raw text was as follows:

1. **Text normalisation:** Convert all characters to lowercase and standardise spacing and encoding.
2. **Special character removal:** We removed symbols such as #, $, %, , etc., as they provide no meaningful contribution to the classification objective.
3. **Stop word removal:** We removed Spanish stop words using NLTK [22], a powerful open source Python library for NLP.
4. **Lemmatisation:** Tokens were reduced to their canonical forms using Spacy [23].

Once the input songs were cleaned and preprocessed, we transformed them into TF-IDF vectors using Scikit-Learn [24], a widely used Python library for machine learning. TF-IDF is a very well-known text representation method that reflects the importance of a word in a document relative to a collection or corpus. It combines two components: Term Frequency (TF), which measures how frequently a term appears in a document, and Inverse Document Frequency (IDF), which downscales terms that appear frequently across many documents. We then trained and evaluated four classifiers- Support Vector Machine (SVM), Gradient Boosting, Random Forest, and Logistic Regression- selected for their robust performance in text classification tasks.

To optimise model performance, we performed Bayesian hyperparameter optimisation (HPO) using the Optuna framework [25], along with 5-fold stratified cross-validation to ensure consistent evaluation. While HPO was performed for all four models, we report only the hyperparameters for the Random Forest classifier in Table 3, as this model was ultimately selected for inclusion in our final system submission. Reporting its configuration in this section supports the reproducibility of the best-performing pipeline from this approach.

The Bayesian search converged on a moderately deep forest as shown in Table 3. The model consists of 371 trees with a maximum depth of 50. This configuration gives the RF enough capacity to learn various decision paths without overfitting. Node growth is controlled by a minimum split size of 2 and a minimum of 6 samples per leaf. On the text side, the optimiser selected a unigram TF-IDF representation, retaining tokens that appear in at least two documents but in no more than 68% of the corpus.

Table 3
Best hyperparameters for the Random Forest classifier.

Parameter	Value
vect_max_df	0.6825
vect_min_df	2
vect_ngram_range	(1, 1)
tfidf_use_idf	True
rf_n_estimators	371
rf_max_depth	50
rf_min_samples_split	2
rf_min_samples_leaf	6
rf_max_features	sqrt
rf_criterion	entropy

2.3.2. BERT-Based Models

Transformer-based architectures have become the foundation for state-of-the-art LLMs such as BERT, LLaMA and GPT, among others. First introduced by Vaswani et al. in "Attention is All You Need" [26], the Transformer was proposed as a sequence transduction model with an encoder-decoder architecture built entirely around self-attention layers. This innovation enabled the model to capture long-term dependencies, attend to different parts of the input simultaneously, and benefit from parallelizable training. Subsequent works adapted this architecture for various NLP tasks by specialising either the encoder or the decoder component.

One of the most influential adaptations is BERT, introduced by Devlin et al. [27]. This architecture relies exclusively on the encoder component of the original Transformer and is designed to learn bidirectional representations through a masked language modelling objective. Once pre-trained, BERT can be fine-tuned to perform specific tasks such as information extraction, semantic search and text classification [28], which is the focus of this study.

As part of our experimentation, we studied several BERT-based models, each selected to address specific challenges identified in our dataset. These include:

- Input texts exceeding 512 tokens, which risk truncation in standard models.
- Short English phrases embedded in Spanish lyrics, which introduce multilingual variability.
- The complexity of misogyny detection, particularly in musical contexts where misogynistic content is often implicit or nuanced rather than explicit.

Below, we briefly describe the models we fine-tuned for this task:

- **PlanTL-GOB-ES/longformer-base-4096-bne-es**: This model is based on the Longformer architecture, which was specifically designed to handle long sequences by replacing the standard self-attention mechanism with a combination of local windowed attention and task-motivated global attention [29]. This architectural modification allows the model to process beyond the 512-token limits of standard BERT-based models. The version used in this study was pre-trained on a large Spanish corpus as part of the MarIA project and supports input sequences of up to 4,096 tokens, enabling our system to handle long lyrics without truncation [30].
- **microsoft/mdeberta-v3-base**: A multilingual DeBERTa model pre-trained on multiple languages, including Spanish and English, useful for handling content in multiple languages [31].
  PlanTL-GOB-ES/roberta-base-bne: A RoBERTa variant developed as part of the MarIA project by PlanTL and the National Library of Spain [30]. The model was pre-trained on a 570 GB dataset of Spanish texts crawled from `.es` domains, covering several topics including: fine arts, politics, and feminism. Given this extensive training data, roberta-base-bne is likely to be familiar with lexical and syntactic patterns found in song lyrics. This makes the model a strong candidate for detecting misogynistic content in Spanish song lyrics.

These models can be found in Hugging Face public Hub [7, 8, 9].

Once the models were selected, we designed a text preprocessing pipeline adapted for each architecture. The first stage involved two basic text cleaning techniques: text normalisation and removal of special characters. Each lyric was normalised to Unicode, and non-informative symbols (such as "$", "[", "]", "-") were removed. No additional token-level operations, such as stop-word removal or lemmatisation, were applied because each model's subword tokeniser already handles grammatical variations.

For the second stage, we performed data sampling from the original and augmented datasets to identify a class distribution that would optimise macro F1-score on the validation set and improve the model's ability to generalise to unseen data. For mDeBERTaV3 and RoBERTa, we oversampled the minority (misogynistic) class to achieve a 1:1 class ratio, resulting in a balanced dataset of misogynistic and non-misogynistic instances. Validation macro-F1 peaked at this ratio, confirming that the models benefited from additional positive evidence.

In contrast, the Longformer show better performance when trained on a larger dataset, even at the cost of class imbalance. Therefore, instead of oversampling just the minority class, we used the entire augmented dataset while preserving the original class distribution. To reduce the impact of class imbalance, we applied a weighted loss function, assigning higher penalties to misclassified misogynistic song lyrics. This approach enabled the Longformer to benefit from a larger and more diverse dataset while still addressing class imbalance.

All three models were fine-tuned using Bayesian optimisation in Optuna [25] (20 trials per model), with each trial running for up to 12 epochs. This setup was chosen based on the approximate time required for full training, with the goal of finding a balance between computational costs and search space exploration. To control overfitting, we implemented early stopping by monitoring the macro-F1 on the validation set after every 100 training steps. The best configurations are reported in Table 4.

Table 4
Hyperparameters used for fine-tuning BERT-based models on Task 1.

Parameter	longformer-base-4096-bne-es	mdeberta-v3-base	roberta-base-bne
learning_rate	1.20e-05	7.82e-06	4.44e-06
weight_decay	0.0278	0.2791	0.0017
batch_size	6	8	32
warmup_ratio	–	0.1182	0.0946
hidden_dropout	0.1431	0.0520	0.0531
attn_dropout	0.2150	0.3023	0.2736
classifier_dropout	0.1312	0.3771	0.3934
optimizer	adamw_torch	adamw_torch	adamw_torch
lr_scheduler_type	linear	linear	linear

2.3.3. Ensemble of Transformers

As mentioned in the previous section, each base model possesses attributes that make it particularly effective in certain scenarios. To build a more robust system, we implemented an ensemble strategy that combines the predictions of the individual transformer models described above. Ensemble methods have been widely used to reduce variance and improve generalisation [32]. Moreover, recent work has shown that ensembles are especially effective when the individual models are diverse and learn different sets of features [33].

Our ensemble approach for task 1 is based on a stacked architecture, a type of ensemble modelling where the predictions from several base models are used as input features for a meta-learner, which then makes the final prediction. The final ensemble included the three best-performing fine-tuned models from the previous section: Longformer, mDeBERTaV3, and RoBERTa. All base models were trained using the same data split: 70% for training, 10% for hyperparameter tuning, and the remaining

20% (unseen by any base model) was reserved for training the meta-learner, a logistic regression model. Each transformer model produces a class probability vector for the misogyny and non-misogyny classes. Concatenating these three vectors results in a six-dimensional feature vector, which is then passed to the logistic regression responsible for making the final prediction. We optimised the hyperparameters for the meta-learner with Optuna and evaluated performance using stratified 5-fold cross-validation. The overall architecture of the ensemble strategy is illustrated in Figure 1.

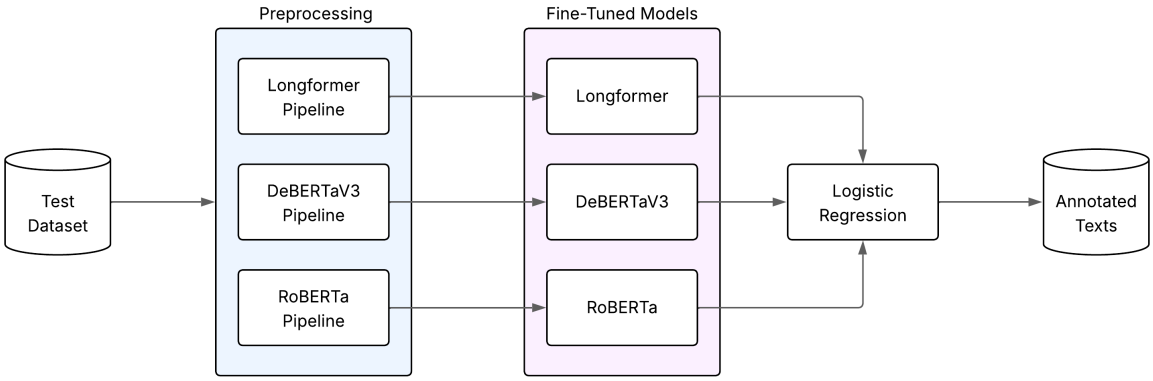

Figure 1: Stacked ensemble architecture with a logistic regression meta-learner.

2.4. Task 2: Fine-grained Misogyny Speech Detection

2.4.1. Classical Machine Learning Classifiers

For task 2, we reused the classical machine learning pipeline described in Section 2.3.1, adapting it to a multiclass classification setting. The preprocessing steps, feature extraction using TF-IDF, and the set of models evaluated remained the same. The main difference was that models were trained to classify samples into one of four misogyny subtypes: Sexualization (S), Violence (V), Hate (H), or Not Related (NR).

As in the first task, we explore four classifiers: SVM, Gradient Boosting, Random Forest, and Logistic Regression. Hyperparameter optimisation was carried out using Optuna, with stratified 5-fold cross-validation, and macro F1-score as the primary evaluation metric. The best results were achieved with the SVM classifier, whose optimal hyperparameters are detailed in Table 5.

Table 5
Best hyperparameters for the SVM classifier on Task 2.

Parameter	Value
vect_max_df	0.5109
vect_min_df	5
vect_ngram_range	(1, 1)
tfidf_use_idf	False
tfidf_norm	l2
svm_C	0.9179
svm_kernel	sigmoid
svm_shrinking	False
svm_gamma	0.9896

2.4.2. Instruction-Tuned Generative Models

In contrast to encoder-based models commonly used in classification tasks, generative language models leverage autoregressive decoding to generate text token by token, conditioned on a prompt [34]. While primarily designed for open-ended generation tasks, these models have recently demonstrated strong performance in zero and few-shot classification as well as instruction following, particularly when fine-tuned with task-specific instructions [35]. For task 2, we adopted this approach and investigated the application of two state-of-the-art generative models to detect misogynistic phrases in lyrics.

The first model we selected was Qwen3-14B, using the quantised variant `unsloth/Qwen3-14B-Base-unsloth-bnb-4bit` [10]. For the second model, we used the LLaMA-3.1-8B architecture, specifically the `meta-llama/Llama-3.1-8B` [11]. Both models were fine-tuned using the parameter-efficient fine-tuning (PEFT) technique, Low-Rank Adaptation (LoRA), enabling effective training with reduced memory and computational requirements.

Qwen3 is the latest generation of foundational models in the Qwen family, improving upon its predecessor, Qwen2.5, in several ways. It was pre-trained on a massive corpus of 36 trillion tokens across 119 languages, which is three times the language coverage of Qwen2.5.

In terms of architecture, Qwen3 models are divided into two groups: mixture-of-experts (MoE) and dense models. The Qwen3-14B model used in our work belongs to the dense family, which has a similar architecture to Qwen2.5. However, Qwen3 introduces two key modifications aimed at improving training stability: first, the removal of bias terms from the query, key, and value (QKV) projections; and second, the incorporation of QK-Norm, a normalisation technique applied to query and key vectors within the attention computation. This normalisation step prevents divergence due to uncontrolled attention logit growth, thereby improving numerical stability [36].

The model's pre-training process follows a three-stage pipeline:

- **General Stage**: Acquisition of broad linguistic knowledge and multilingual fluency. The model was first exposed to a very large, multilingual corpus, giving it broad language coverage, including Spanish vocabulary and stylistic variations.
- **Reasoning Stage**: Specialisation in knowledge-intensive domains, improving the model's ability to follow complex prompts and produce consistent outputs.
- **Long Context Stage**: Pre-training on high-quality sequences up to 32,768 tokens, enabling long-context understanding, crucial for processing extended song lyrics.

Our approach using Qwen3 is described below:

- **Preprocessing**: The preprocessing pipeline that showed the best results was the same used for Longformer in task 1. Each lyric was normalised to Unicode, and non-informative symbols were removed. We then oversampled both majority and minority classes proportionally to preserve the original class distribution while increasing the number of training samples.
- **Prompting**: We defined the prompt templates based on the Alpaca instruction format proposed in [37], which segments prompts into three components: *instruction*, *input*, and *response*. Moreover, we found that explicitly including the label descriptions within the instruction section improved model performance. Examples of prompts used for task 2 are shown in Table 6.
- **Architecture Adaptation**: Before fine-tuning, we constrained the model's output by modifying the final layer (`lm_head`) to only produce logits for the four valid task 2 labels: `S`, `H`, `V`, and `NR`. This adjustment helped the model stay focused on the classification task and allowed for more precise evaluation by eliminating unrelated token generation.
- **Fine-Tuning**: We used the `FastLanguageModel` class from the `Unsloth` library, which supports quantised models and integrates with Hugging Face's `TRL` library, enabling supervised fine-tuning via the `SFTTrainer` class. Table 7 shows the configuration used in the fine-tuning process.

- **LoRA**: To enable efficient fine-tuning of Qwen3, we applied LoRA, a parameter-efficient method that injects trainable low-rank matrices into selected linear layers of the model while keeping the original weights frozen. This technique significantly reduces memory usage without compromising performance [12]. We targeted the following modules during LoRA adaptation: `lm_head`, `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`. These modules cover the core attention and feed-forward components of the transformer, as well as the output classification head. The detailed LoRA configuration and hyperparameters are shown in Table 8.

Table 6
Prompt for instruction fine-tuning.

Prompt:	`### Instruction:`
	Classify the misogyny subtype of the following lyric.
	Categories:
	• S: describe or suggest sexual acts, sexual language, or insinuations
	• V: physical or verbal aggression, threats, or violent actions
	• H: offensive or discriminatory language, expressions of contempt, or hostility towards a group or individual
	• NR: none of the above
	Return only one label from: S, V, H, NR
	`### Input:`
	{lyrics}
	`### Response:`

Response:	H

Table 7
Training configuration used for Qwen3 fine-tuning.

Parameter	Value
`train_batch_size`	8
`gradient_accumulation_steps`	2
`warmup_steps`	10
`max_grad_norm`	0.3
`optimizer`	adamw_8bit
`weight_decay`	0.001
`learning_rate`	1e-4
`lr_scheduler_type`	cosine
`num_train_epochs`	3

Table 8
LoRA configuration used for Qwen3 fine-tuning.

Parameter	Value
`rank`	16
`alpha`	16
`dropout`	0
`bias`	none
`use_gradient_checkpointing`	unsloth
`use_rslora`	True

LLaMA-3.1-8B is the smallest instruction-tuned model in the Meta LLaMA-3 family, consisting of an 8 billion parameter decoder-only architecture with a 128K token context window and strong zero-shot

performance in Spanish [38]. The 3.1 version was trained for multilingual dialogue, including Spanish and English. We incorporate this generative architecture into our approach to test whether the results reached with Qwen3-14B can be achieved under lower memory budgets.

Our approach using Llama-3.1-8B is described below:

- **Preprocessing and Prompting**: We applied the same preprocessing and prompting strategy as described for Qwen3. No architectural modifications were applied to constrain the outputs. Instead, the predictions were obtained by extracting the token that followed the response section of each output.
- **Fine-Tuning**: We used the HuggingFace library to fine-tune the pre-trained model with the LoRa method.
- **LoRa**: We targeted all linear layers within the transformer architecture, including: {`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`}.

2.4.3. Hierarchical Transformer

In addition to standalone generative models, we experimented with a hierarchical classification approach that divides the task into two sequential stages: a binary classifier to detect the presence of misogyny, followed by a subtype classifier applied only when misogyny is present. The architecture is illustrated in Figure 2.

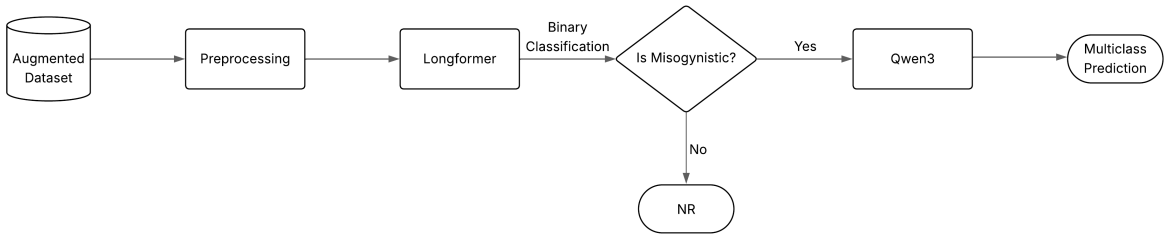

Figure 2: Hierarchical classifier.

- **Preprocessing**: We applied the same preprocessing pipeline as in previous approaches: each lyric was normalised to Unicode and stripped of non-informative symbols. Oversampling was applied proportionally across classes to preserve the original distribution and increase training data size.
- **Binary Classification**: We fine-tuned the Longformer model to predict whether each lyric contains misogynistic content. This model was optimised using the same hyperparameter tuning strategy described in task 1, including a 5-fold stratified cross-validation and Optuna search. During preliminary exploratory data analysis, we found that every lyric annotated as 'Not Related' (NR) in task 2 was also labelled as 'Not Misogynist' (NM) in the task 1 dataset. Consequently, we designed the first phase of the Hierarchical approach as a misogyny detection problem, where the binary classifier treats NR as the negative class and groups the three subtype labels ('Sexualisation', 'Violence' and 'Hate') into the positive class.
- **Decision**: If the input text is classified as `NM`, the pipeline stops and outputs the label `NR`. Otherwise, the lyric is passed to the next stage for subtype classification.
- **Subtype Classification**: We used the Qwen3-14B model to classify the misogynistic lyric into one of three subtypes: `S`, `H`, or `V`. The model was fine-tuned using the same setup described earlier, with the only difference being that the output space was reduced to these three categories.

Implementation details. All experiments were conducted using Python 3.11. Model training was performed on a single NVIDIA RTX A6000 GPU with 48GB of VRAM. The source code and configuration files are publicly available at: https://github.com/JUTO97/IberLEF_2025_Misongyny.

3. Results and Discussion

This section first reports the performance of our systems on the 20% hold-out partition used during development and then summarises our position on the official MiSonGyny 2025 leaderboard.

During the development phase, we applied a stratified split of the dataset into training (70%), validation (10%), and test (20%) partitions. The evaluation on this internal test set allowed us to compare alternative architectures and hyperparameters. The model selection was primarily driven by the macro-F1 score, which balances precision and recall across classes and is therefore suitable for class imbalance in both tasks. For completeness, we also report precision, recall, and accuracy.

Because these scores were obtained from models trained on only 70% of the data, they may slightly underestimate the final performance; nevertheless, they proved to be robust enough to select the best candidates. The final models were subsequently retrained on the entire dataset before being submitted for official scoring.

While the competition leaderboard records only the single best run per team and task, we include below the results for all of our submitted runs for task 1 and task 2, alongside the official ranking.

3.1. Task 1: Misogyny Speech Detection

3.1.1. Results on our Internal Test Split

In this subsection, we report the scores we obtained on the 20% test split; these results determined which models were ultimately sent to the competition.

As shown in Table 9, among the traditional machine learning models, the Random Forest classifier achieved the highest macro-F1 score (0.7673), establishing a strong baseline for comparison against more complex architectures. This score significantly outperforms the top-ranked system reported in the HOMO-MEX 2024 [39] share task (subtask 3), which focused on detecting LGBT+phobic content in Spanish song lyrics. In task 3, the first place achieved an F1 score of 0.5762. Although the two tasks are not directly comparable due to differences in label definitions and dataset composition, the contrast still offers a useful point of reference for evaluating model effectiveness in a lyric-based classification task.

Table 9
Our results for task 1 on the hold-out set from the development phase. Bold values indicate the top result within each approach.

Model	Accuracy	Precision	Recall	F1
Traditional ML Models				
Random Forest	**0.7682**	**0.7843**	**0.7752**	**0.7673**
XGBoost	0.7838	0.7494	0.7149	0.7271
Logistic Regression	0.7743	0.7332	0.7235	0.7279
SVM	0.7862	0.7478	0.7386	0.7428
BERT-based Transformers				
mDeBERTaV3	0.7888	0.7715	0.7794	0.7747
RoBERTa	0.8112	0.8002	0.7835	0.7902
Longformer	**0.8157**	**0.8038**	**0.7911**	**0.7964**
Ensembles				
Stack	**0.8520**	**0.8448**	**0.8461**	**0.8454**
Soft Voting	0.8247	0.8156	0.7981	0.8052

Among the BERT-based Transformer models, we observe less variability in F1 performance compared to traditional ML models. Our best system in this category was the Longformer, with a macro-F1 of 0.7964, followed by RoBERTa (0.7902) and mDeBERTaV3 (0.7747). These results suggest that input length plays a crucial role in detecting misogyny. Longformer's ability to process extended context likely helped to identify misogynistic content in sections of lyrics that standard Transformer models, limited to 512 tokens, would have truncated.

Additionally, while mDeBERTaV3 is a powerful multilingual model, it was pre-trained on less Spanish text than longformer-base-4096-bne-es and roberta-base-bne, both trained on high-quality Spanish corpora. This performance gap suggests that task 1 benefits from both a combination of extended context and language-specific pre-training, as misogynistic content is often implicit or dispersed throughout the lyrics.

Finally, we evaluated a third group of models, which are ensembles built from the BERT-based Transformers described above. The best result was obtained using a stacked ensemble with logistic regression as the meta-learner, achieving a macro F1 score of 0.8454, about +4.9 percentage points (pp) above the Longformer alone. In contrast, our soft voting ensemble achieved a macro F1 score of 0.8052, which is slightly better than the best individual model (Longformer) but significantly lower than the stacked approach. These results suggest that ensemble models can effectively capture complementary patterns across architectures, particularly when guided by a meta-learner.

The three best models, one from each approach, identified in Table 9 (Random Forest, Longformer, and the stacked ensemble), were retrained on the entire dataset before submission to the official leaderboard. Their final hyperparameter configuration is documented in Tables 3 and 4, respectively.

3.1.2. Official Results

This subsection discusses the scores released on the MiSonGyny 2025 leaderboard. A total of 13 teams participated in task 1, each allowed up to 10 submissions.

According to the ranking released by the organisers, our submissions delivered a strong performance in the final evaluation (see Table 10). All of our Transformer-based models (Stacked-Ensemble and the two Longformer variants) outperformed every system submitted by other contestants.

Table 10
Our results for task 1 on the final test dataset

Place	Submission_Id	F1-Score	Model_Type
1	289972	0.8811	Stack-Ensemble
2	286732	0.8523	Longformer
3	288684	0.8496	Longformer
38	286119	0.7602	Random Forest

Our stacked ensemble was the top-performing system, securing first place with a macro-F1 of 0.8811. This is +2.9 pp higher than our best single model, Longformer (0.8523), and +4.5 pp above the best system from the second-best team (*atorojaen*, 0.8359). Our second Longformer variant ranked third overall, with a score of 0.8496.

The official baseline model submitted under the username *misongyny* achieved a macro-F1 score of 0.7434 (13.8 pp relative to our ensemble), placing 41st overall. Our best classical machine learning model (Random Forest) surpassed the baseline, scoring 0.7602 and ranking 38th. Overall, these results validate our methodology and confirm the robustness of the ensemble approach compared to both individual Transformers and traditional ML classifiers.

3.2. Task 2: Fine-grained Misogyny Speech Detection

3.2.1. Results on our Internal Test Split

In this subsection, we report the scores we obtained on the 20% test split; these results determined which models were ultimately sent to the competition.

Among the traditional ML models, the best result came from the SVM, which achieved a macro-F1 of 0.4230 (see Table 11). Although this score remains relatively low, it outperformed all the other classical models. These results reinforce the notion that traditional models struggle to capture the semantic nuances necessary for detecting misogynistic content, likely due to the high variability of song lyrics and the limited representational power of TF-IDF features in multi-class contexts.

The second group comprised instruction-tuned generative transformers. Qwen3-14B achieved the highest overall performance with a macro-F1 score of 0.5762, followed by the hierarchical model (0.5486) and LLaMA-3.1-8B (0.5354). The superior performance of Qwen3 (+2.7 pp over the hierarchical system and +4.1 pp over LLaMA) can be attributed to several factors, such as: its ability to process long-context sequences; the modification of its output layer, which constrains predictions to task-specific labels only; and the use of instruction-tuned prompts that embed label semantics directly into the model's input.

While Qwen3 stood out as the most effective system, the other generative models also showed promising results. For instance, LLaMA achieved comparable results, though its performance was slightly lower in both Recall and macro F1 score (−2.4 pp and −4.1 pp, respectively). This may be explained by its smaller model size of 8B parameters relative to Qwen3's 14B. The hierarchical model performed reasonably well but still fell short of the generative models, particularly Qwen3, in terms of Precision and macro F1 score (−4.3 pp and −2.8 pp, respectively).

Comparing the two model approaches, instruction-tuned generative transformers performed significantly better over traditional ML models: Qwen3 outperformed SVM by +15.3 pp on macro-F1. Even the weakest generative model (LLaMA) had a +11.2 pp advantage. These results prove that instruction-tuned large language models performed better in fine-grained misogyny classification than feature-based classifiers.

Table 11
Our results for task 2 on the hold-out set from the development phase. Bold values indicate the top result within each approach.

Model	Accuracy	Precision	Recall	F1
Traditional ML Models				
Random Forest	0.5897	0.4165	0.3963	0.3917
XGBoost	0.6410	0.4300	0.4191	0.4178
Logistic Regression	0.4829	0.4545	0.4513	0.4045
SVM	**0.5897**	**0.4166**	**0.4368**	**0.4230**
Instruction-Tuned Generative Models				
Qwen3-14B	**0.7447**	**0.5839**	**0.5757**	**0.5762**
LLaMA-3.1-8B	0.7031	0.5556	0.5516	0.5354
Mixed Architecture				
Hierarchical	**0.7132**	**0.5412**	**0.5635**	**0.5486**

A closer look at the class predictions of Qwen3-14B (see Figure 3) reveals that the model performed well on the 'Non-Related' (NR) and 'Sexualisation' (S) classes, correctly classifying 91 and 73 samples, respectively. However, performance on the 'Hate' (H) and 'Violence' (V) classes was lower, with most misclassifications occurring between the S and NR classes. This pattern reflects the difficulty in distinguishing certain forms of misogyny, particularly when there are few instances available for training, as is the case with classes V and H (see Table 2).

The three best models identified in Table 11 (SVM, Qwen3-14B, and the hierarchical system) were retrained on the entire dataset before submission to the official leaderboard. Their final hyper-parameter settings are documented in Tables 5, 7, and 6, respectively. The following subsection reports how these three runs performed on the official MiSonGyny 2025 task 2 leaderboard.

3.2.2. Official Results

This subsection reviews the official MiSonGyny 2025 leaderboard for task 2. A total of 9 teams participated, each allowed up to 10 submissions.

According to the ranking released by the organisers, our team achieved first place with a macro-F1 of 0.5895 using Qwen3-14B (see Table 12). Within our own submissions, this score is +3.3 pp higher than the hierarchical model (0.5564) and +14.7 pp higher than the SVM model (0.4428), showing the clear

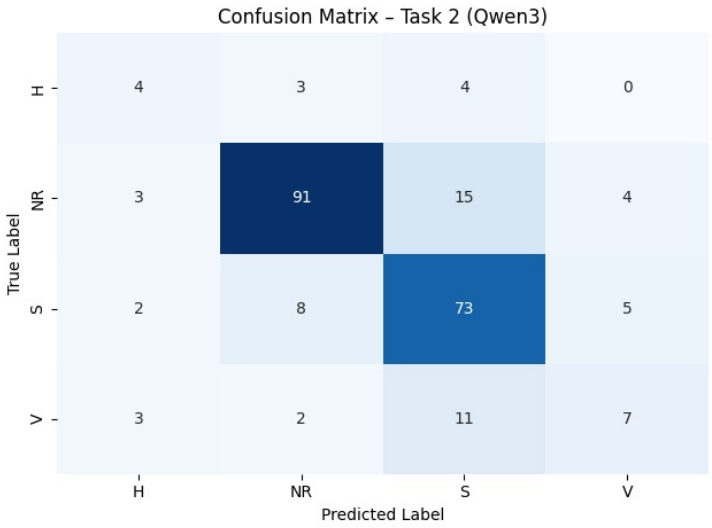

Figure 3: Confusion matrix for the Qwen3-14B model on the task 2 testing set.

advantage of instruction-tuned generative transformers over traditional ML alternatives for fine-grained misogyny detection in song lyrics.

Table 12
Our results for task 2 on the final test dataset.

Place	Submission_Id	F1-Score	Model_Type
1	289972	0.5895	Qwen3
3	288780	0.5564	Hierarchical
20	286119	0.4428	Support Vector Machine

The hierarchical pipeline, which consists of a Longformer for binary classification (misogyny detection) followed by Qwen3 for subtype prediction (see Figure 2), secured third place overall. Meanwhile, our SVM model finished 20th yet still surpassed the competition baseline.

Compared with other teams, Qwen3 had a clear advantage, finishing +2.8 pp ahead of the best run from the second-ranked team (0.5613) and +17.4 pp above the official baseline system submitted under the username misongyny (0.4151, 27th place).

Overall, these results demonstrate the robustness of our approach across modelling paradigms. The clear margin between our top submissions and the rest of the leaderboard further emphasises the strength of our instruction-tuned generative transformer and its applicability to fine-grained misogyny speech detection.

4. Conclusions

This paper described the participation of the HULAT_UC3M team in the IberLEF-MiSonGyny 2025 shared tasks. Throughout our work, we explored a broad spectrum of modelling approaches, ranging from traditional machine learning models to state-of-the-art instruction-tuned generative architectures.

One of the main contributions of this work was the identification of effective data augmentation strategies to mitigate the challenges posed by small and highly imbalanced datasets. Furthermore, we leveraged instruction fine-tuning combined with LoRA, a parameter-efficient technique that enables the training of large language models on a single GPU, while still capturing the nuanced and often subjective nature of misogyny in Spanish song lyrics. These strategies led to strong results, with our systems ranking first in both binary detection (task 1) and fine-grained classification (task 2).

Although the final outcomes were highly successful, achieving them was not straightforward. Instead, the process involved several missteps. For instance, we learned that simply scaling model size can lead to overfitting when data is insufficient, and that combining multiple augmentation techniques indiscriminately may degrade dataset quality and harm model performance, highlighting the need for careful curation and validation. These insights were critical to improving our final pipeline and provided valuable guidance for researchers in this field.

For future work, we plan to extend our study to three complementary directions. First, traditional augmentation techniques such as back-translation, EDA or AEDA could be applied in a lyric-aware manner, preserving elements such as rhyme, tempo, metre and line breaks in order to generate high-quality synthetic lyrics. Second, because misogyny labelling is subjective, our next models will incorporate explanation mechanisms (e.g., attention heat-maps or exemplar retrieval) so that end-users and annotators can inspect which phrases triggered a prediction, thereby facilitating error analysis. Finally, we aim to expand the linguistic coverage by incorporating a multilingual lyric dataset and by experimenting with Mixture-of-Experts (MoE) models with Chain-of-Thought (CoT) prompting strategies, which may further enhance reasoning capabilities and fine-grained misogyny discrimination in song lyrics.

5. Acknowledgments

Grant PID2023-148577OB-C21 (Human-Centered AI: User-Driven Adapted Language Models-HUMAN_AI) by MICIU/AEI/ 10.13039/501100011033 and by FEDER/UE.

Declaration on Generative AI

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

References

[1] C. Kramarae, D. Spender, Routledge international encyclopedia of women: Global women's issues and knowledge, Routledge, 2004.

[2] H. Kirk, W. Yin, B. Vidgen, P. Röttger, 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, E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2193–2210. URL: https://aclanthology.org/2023.semeval-1.305/. doi:10.18653/v1/2023.semeval-1.305.

[3] L. Plaza, J. C. de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2023-learning with disagreement for sexism identification and characterization (extended overview)., CLEF (Working Notes) (2023) 813–854.

[4] L. Plaza, J. Carrillo-de Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of exist 2024—learning with disagreement for sexism identification and characterization in tweets and memes, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2024, pp. 93–117.

[5] T. Alcántara, M. Soto, C. Macias, O. Garcia-Vazquez, A. Espinosa-Juarez, H. Calvo, J. E. Valdez-Rodríguez, E. Felipe-Riveron, Overview of MiSonGyny at IberLEF 2025: Misogyny Speech Detection in Spanish Language Song Lyrics, Procesamiento del Lenguaje Natural 75 (2025).

[6] 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.

[7] H. Face, Longformer base trained with data from the national library of spain (bne), https://huggingface.co/PlanTL-GOB-ES/longformer-base-4096-bne-es, ????. [Accessed 01-06-2025].
[8] H. Face, microsoft/mdeberta-v3-base, https://huggingface.co/microsoft/mdeberta-v3-base, ????. [Accessed 01-06-2025].
[9] H. Face, Roberta base trained with data from the national library of spain (bne), https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne, ????. [Accessed 01-06-2025].
[10] H. Face, unsloth/qwen3-14b-base-unsloth-bnb-4bit, https://huggingface.co/unsloth/Qwen3-14B-Base-unsloth-bnb-4bit, ????. [Accessed 02-06-2025].
[11] H. Face, Llama-3.1-8b, https://huggingface.co/meta-llama/Llama-3.1-8B, ????. [Accessed 02-06-2025].
[12] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, 2021. URL: https://arxiv.org/abs/2106.09685. arXiv:2106.09685.
[13] A. Mohasseb, E. Amer, F. Chiroma, A. Tranchese, Leveraging advanced nlp techniques and data augmentation to enhance online misogyny detection, Applied Sciences 15 (2025). URL: https://www.mdpi.com/2076-3417/15/2/856. doi:10.3390/app15020856.
[14] D. R. Beddiar, M. S. Jahan, M. Oussalah, Data expansion using back translation and paraphrasing for hate speech detection, Online Social Networks and Media 24 (2021) 100153. URL: https://www.sciencedirect.com/science/article/pii/S2468696421000355. doi:https://doi.org/10.1016/j.osnem.2021.100153.
[15] M. S. Jahan, M. Oussalah, D. R. Beddia, J. kabir Mim, N. Arhab, A comprehensive study on nlp data augmentation for hate speech detection: Legacy methods, bert, and llms, 2024. URL: https://arxiv.org/abs/2404.00303. arXiv:2404.00303.
[16] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Łukasz Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, J. Dean, Google's neural machine translation system: Bridging the gap between human and machine translation, 2016. URL: https://arxiv.org/abs/1609.08144. arXiv:1609.08144.
[17] A. Currey, A. Karakanta, J. Dehdari, Using related languages to enhance statistical language models, in: J. Andreas, E. Choi, A. Lazaridou (Eds.), Proceedings of the NAACL Student Research Workshop, Association for Computational Linguistics, San Diego, California, 2016, pp. 116–123. URL: https://aclanthology.org/N16-2017/. doi:10.18653/v1/N16-2017.
[18] A. Karimi, L. Rossi, A. Prati, AEDA: An easier data augmentation technique for text classification, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 2748–2754. URL: https://aclanthology.org/2021.findings-emnlp.234/. doi:10.18653/v1/2021.findings-emnlp.234.
[19] J. Wei, K. Zou, EDA: Easy data augmentation techniques for boosting performance on text classification tasks, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 6382–6388. URL: https://aclanthology.org/D19-1670/. doi:10.18653/v1/D19-1670.
[20] F. Tegge, K. Parry, The impact of differences in text segmentation on the automated quantitative evaluation of song-lyrics, PLOS ONE 15 (2020) 1–16. URL: https://doi.org/10.1371/journal.pone.0241979. doi:10.1371/journal.pone.0241979.
[21] Y. Fang, L. Lee, J.-D. Huang, Nycu-nlp at exist 2024: Leveraging transformers with diverse annotations for sexism identification in social networks, CEUR Workshop Proceedings 3740 (2024) 1003–1011. Publisher Copyright: © 2024 Copyright for this paper by its authors.; 25th Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2024 ; Conference date: 09-09-2024 Through 12-09-2024.
[22] S. Bird, E. Loper, NLTK: The natural language toolkit, in: Proceedings of the ACL Interactive

Poster and Demonstration Sessions, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 214–217. URL: https://aclanthology.org/P04-3031/.

[23] M. Honnibal, I. Montani, spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, 2017. To appear.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[25] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2023. URL: https://arxiv.org/abs/1706.03762. arXiv:1706.03762.

[27] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL: https://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[28] C. Sun, X. Qiu, Y. Xu, X. Huang, How to fine-tune bert for text classification?, 2020. URL: https://arxiv.org/abs/1905.05583. arXiv:1905.05583.

[29] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, 2020. URL: https://arxiv.org/abs/2004.05150. arXiv:2004.05150.

[30] A. G. Fandiño, J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022). URL: https://upcommons.upc.edu/handle/2117/367156#.YyMTB4X9A-0.mendeley. doi:10.26342/2022-68-3.

[31] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2021. arXiv:2111.09543.

[32] J. Briskilal, C. Subalalitha, An ensemble model for classifying idioms and literal texts using bert and roberta, Information Processing Management 59 (2022) 102756. URL: https://www.sciencedirect.com/science/article/pii/S0306457321002375. doi:https://doi.org/10.1016/j.ipm.2021.102756.

[33] Z. Allen-Zhu, Y. Li, Towards understanding ensemble, knowledge distillation and self-distillation in deep learning, 2023. URL: https://arxiv.org/abs/2012.09816. arXiv:2012.09816.

[34] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, Q. V. Le, Finetuned language models are zero-shot learners, 2022. URL: https://arxiv.org/abs/2109.01652. arXiv:2109.01652.

[35] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, H. Hajishirzi, Self-instruct: Aligning language models with self-generated instructions, 2023. URL: https://arxiv.org/abs/2212.10560. arXiv:2212.10560.

[36] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, Z. Qiu, Qwen3 technical report, 2025. URL: https://arxiv.org/abs/2505.09388. arXiv:2505.09388.

[37] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Stanford alpaca: An instruction-following llama model, https://github.com/tatsu-lab/stanford_alpaca, 2023.

[38] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Roziere, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Wyatt, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hupkes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic,

F. Guzmán, F. Zhang, et al., The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[39] H. Gómez-Adorno, G. Bel-Enguix, H. Calvo, S. Ojeda-Trueba, S. T. Andersen, J. Vásquez, S. Ojeda-Trueba, T. Alcántara, M. Soto, C. Macías, Overview of HOMO-MEX at IberLEF 2024: Hate Speech Detection Towards the Mexican Spanish speaking LGBT+ Population, Procesamiento del Lenguaje Natural 73 (2024) 393–405.