UMUTeam at MiSonGyny 2026@IberLEF 2026: Hierarchical Multi-Task Learning for Misogyny Detection in Song Lyrics

Tomás Bernal-Beltrán[1], Ronghao Pan[1], Jorge Gómez-Navalón[1], José Antonio García-Díaz[1], Ghassan Beydoun[2] and Rafael Valencia-García[1,*]

[1]Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100, Spain
[2]School of Information, Systems and Modelling, Faculty of Engineering and IT, University of Technology Sydney, Sydney, Australia

Abstract

Misogyny detection in song lyrics is a challenging task due to the presence of figurative language, slang, ambiguity and different manifestations of gender-based harmful discourse. This paper describes the UMUTeam participation in the MiSonGyny 2026 shared task at IberLEF 2026. The task focuses on the automatic detection and characterization of misogynistic content in song lyrics and is divided into three subtasks: binary misogyny detection, multilabel misogyny type classification and binary stereotype identification. Our approach formulates the problem as a hierarchical multi-task classification task. The proposed system relies on a shared Transformer encoder and three task-specific prediction heads, while incorporating hierarchical conditioning from the main misogyny detection task to the remaining subtasks. In addition, the system combines class-weighted training, validation-based threshold optimization and hierarchical consistency constraints during inference. On the official test set, our system ranked first in the three subtasks, achieving F1-scores of 0.8856, 0.7385 and 0.8283 for Subtasks 1, 2 and 3, respectively. These results show that hierarchical multi-task learning is an effective strategy for modeling related dimensions of misogynistic discourse in song lyrics.

Keywords

Misogyny Detection, Sexism Detection, Song Lyrics, Multi-Task Learning, Hierarchical Classification, Multilabel Classification, Natural Language Processing

1. Introduction

Misogyny, sexism and gender-based stereotypes constitute harmful forms of discourse that contribute to the normalization of discrimination, hostility and violence against women. Their automatic detection has become an important research topic within Natural Language Processing (NLP), especially as part of the broader field of abusive and hate speech detection [1]. This line of research is particularly relevant in online environments, where harmful content can be produced and disseminated at large scale and where detection systems must also be considered from ethical and human-rights perspectives [2].

However, misogynistic discourse is not limited to social media. Song lyrics are also a relevant cultural domain, since they can reflect, reproduce or reinforce social attitudes towards gender roles, sexuality, power relations and violence. Recent large-scale analyses have shown that gender bias and sexist content can be identified in song lyrics across genres and time periods, highlighting the relevance of this domain for computational analysis [3].

Despite its social relevance, misogyny detection in song lyrics remains a challenging task. Lyrics often contain figurative language, slang, irony, ambiguity and highly context-dependent expressions. In addition, misogynistic content may appear through different manifestations, including sexualization, violent references, explicit hate speech or more subtle gender stereotypes. This makes the task more

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
tomas.bernalb@um.es (T. Bernal-Beltrán); ronghao.pan@um.es (R. Pan); jorge.gomezn@um.es (J. Gómez-Navalón); joseantonio.garcia8@um.es (J. A. García-Díaz); Ghassan.Beydoun@uts.edu.au (G. Beydoun); valencia@um.es (R. Valencia-García)
0009-0006-6971-1435 (T. Bernal-Beltrán); 0009-0008-7317-7145 (R. Pan); 0009-0005-7185-6054 (J. Gómez-Navalón); 0000-0002-3651-2660 (J. A. García-Díaz); 0000-0001-8087-5445 (G. Beydoun); 0000-0003-2457-1791 (R. Valencia-García)

complex than a single binary decision, since systems must not only determine whether a song contains misogynistic discourse, but also identify the type of discourse expressed and whether gender-based stereotypes are present. Similar fine-grained and hierarchical formulations have also been adopted in recent sexism detection benchmarks, reflecting the need to model several levels of information within the same task [4]. These difficulties are further increased by the multilabel nature of some categories and the class imbalance commonly found in abusive language datasets, which may bias models towards majority classes and reduce performance on underrepresented labels [5].

The MiSonGyny 2026 shared task [6], organized at IberLEF 2026 [7], addresses this problem by focusing on the automatic detection and characterization of misogynistic content in song lyrics. The task builds on the previous MiSonGyny 2025 shared task [8], which introduced misogyny speech detection in Spanish-language song lyrics as a benchmark within IberLEF. The task is divided into three subtasks. Subtask 1 addresses misogyny detection as a binary classification problem, distinguishing misogynistic from non-misogynistic lyrics at the song level. Subtask 2 focuses on misogyny type classification as a multilabel problem, requiring systems to identify the presence of sexualization, violence and hate. Subtask 3 addresses stereotype identification as a binary classification problem, detecting whether lyrics contain gender-based stereotypes. This formulation is particularly relevant because it combines global detection, fine-grained characterization and stereotype recognition within the same benchmark.

Recent work on misogyny, sexism and hate speech detection has shown the effectiveness of Transformer-based encoders and supervised fine-tuning for abusive language classification. Shared tasks such as AMI [9], EXIST [10] and EDOS [4] have encouraged the development of systems capable of moving beyond binary detection towards more fine-grained and structured formulations. At the same time, multi-task and hierarchical learning approaches have become increasingly attractive for tasks in which several labels describe related aspects of the same phenomenon [11, 12]. In the Spanish domain, Spanish MTLHateCorpus 2023 provides a multi-task hate speech benchmark that jointly models speech type, target, target group and intensity, further supporting the relevance of modeling related abusive language dimensions simultaneously [13].

In this paper, we present the system developed by UMUTeam for the MiSonGyny 2026 shared task at IberLEF 2026. Our approach formulates the problem as a hierarchical multi-task classification task. A shared Transformer encoder is used to learn a common representation of the lyrics, while three task-specific classification heads are optimized for misogyny detection, misogyny type classification and stereotype identification. To explicitly exploit the structure of the task, lower-level predictions are conditioned on the output of the misogyny detection head, allowing the model to incorporate hierarchical dependencies between subtasks. In addition, the system incorporates class-weighted training to mitigate label imbalance, task-specific threshold optimization to improve multilabel decision boundaries and hierarchical prediction constraints during inference.

The official results show that our system achieved the best performance in the three subtasks, ranking first in Subtask 1 with an F1-score of **0.8856**, first in Subtask 2 with an F1-score of **0.7385** and first in Subtask 3 with an F1-score of **0.8283**. These results suggest that modeling misogyny detection in lyrics as a hierarchical multi-task problem is an effective strategy, particularly when combined with class-weighted training, task-specific threshold optimization and consistency constraints between related predictions. The strongest relative improvement was observed in the multilabel misogyny type classification subtask, where the proposed system outperformed the second-ranked submission by more than three F1-score points, indicating that explicitly modeling task dependencies can be especially useful for fine-grained and imbalanced prediction settings.

The remainder of this paper is organized as follows. Section 2 reviews related work on misogyny and sexism detection, song lyrics analysis, Transformer-based classification, multi-task learning, hierarchical classification and threshold optimization. Section 3 describes the proposed system, including the hierarchical multi-task architecture and inference strategy. Section 4 presents the dataset, preprocessing, training configuration and evaluation setup. Section 5 reports and discusses the official results obtained in the shared task. Finally, Section 6 concludes the paper and outlines possible directions for future work.

2. Background Information

The automatic detection of misogyny, sexism and other forms of abusive language has become an important research topic within NLP due to its social impact and the increasing presence of harmful content in online platforms. Early studies mainly focused on identifying hate speech and offensive language in social media through traditional machine learning approaches based on lexical, syntactic and semantic features [1]. As larger annotated corpora became available, research progressively evolved towards more fine-grained formulations capable of distinguishing different manifestations of abusive language and identifying their specific targets [2].

Several shared tasks have significantly contributed to the development of misogyny, sexism and hate speech detection systems. The Automatic Misogyny Identification (AMI) task at EVALITA introduced one of the first benchmark datasets for misogyny detection, including both binary identification and category classification subtasks [9]. Later editions expanded these taxonomies and encouraged the development of systems capable of modeling multiple dimensions of misogynistic content simultaneously. Similarly, the sEXism Identification in Social neTworks (EXIST) task at CLEF proposed multilingual sexism detection benchmarks in Spanish and English, covering both binary and fine-grained categorization settings [10]. In the IberLEF context, related shared tasks have also addressed other forms of abusive language in Spanish, such as HOMO-MEX, which focused on hate speech detection towards the Mexican Spanish-speaking LGBT+ population [14]. More recently, the Explainable Detection of Online Sexism (EDOS) task at SemEval introduced a hierarchical taxonomy of sexism with multiple levels of granularity, highlighting the need for models capable of capturing dependencies between related categories [4].

Beyond social media, the analysis of gender bias and misogynistic content in song lyrics has attracted growing attention. Lyrics constitute an important cultural medium that can reflect and reinforce social stereotypes. Large-scale studies have shown the persistence of gender bias and sexist content across different musical genres and time periods [3]. In addition, the MiSonGyny 2025 shared task specifically addressed misogyny speech detection in Spanish-language song lyrics, providing a direct benchmark for this domain [8]. However, compared with social media, annotated resources for misogyny detection in song lyrics remain scarce, making this domain considerably less explored.

From a methodological perspective, early misogyny and sexism detection systems relied on traditional machine learning algorithms such as Support Vector Machines, Logistic Regression or Random Forests combined with handcrafted linguistic features [15]. The introduction of Transformer-based language models marked a major shift in text classification, since pretrained encoders such as BERT [16] could be fine-tuned end-to-end for downstream tasks with limited architectural changes. This paradigm rapidly became dominant in abusive language detection, where contextual representations allowed models to capture lexical, syntactic and semantic cues more effectively than sparse feature-based approaches. In multilingual and Spanish-language settings, models such as multilingual BERT, XLM-R and Spanish-specific encoders have been widely adopted in shared tasks such as AMI [9] and EXIST [10]. These systems usually follow a supervised fine-tuning strategy, adapting a pretrained encoder to the target dataset with task-specific classification layers.

More recently, Large Language Models (LLMs) have been explored as an alternative to fully supervised fine-tuning. Instead of training a task-specific classifier, these models can be used through prompting, zero-shot and few-shot learning, relying on instruction-following capabilities to produce labels directly from natural language descriptions of the task. This has motivated comparisons between encoder fine-tuning and In-Context Learning (ICL) with generative models such as GPT-style models, LLaMA, Qwen or Gemma. Although LLM-based approaches reduce the need for task-specific training data and can be flexible across tasks, recent studies on Spanish hate speech detection [17] and sentiment analysis [18] show that fine-tuned encoder models still tend to provide more stable and competitive results than zero-shot or few-shot prompting in specialized classification scenarios.

At the same time, researchers have increasingly explored multi-task learning strategies to exploit relationships between correlated labels and subtasks. Multi-task learning has been shown to improve generalization by sharing representations across related objectives [11]. In NLP, this idea is commonly

implemented through shared Transformer encoders combined with task-specific classification heads, allowing a single contextual representation to support several related prediction problems [12]. These approaches are particularly attractive for misogyny and sexism detection, where different labels often describe complementary aspects of the same phenomenon. Recent studies have shown that jointly learning related abusive language tasks can improve robustness and generalization by leveraging shared information across objectives such as hate speech detection, sentiment analysis, emotion recognition or target identification [19]. In the Spanish domain, Spanish MTLHateCorpus 2023 provides a multi-task hate speech benchmark designed to jointly identify speech type, target, target group and intensity, further supporting the relevance of modeling related hate speech dimensions simultaneously instead of treating each classification problem independently [13].

Closely related to multi-task learning, hierarchical classification approaches explicitly model dependencies between labels organized at different levels of abstraction. These methods assume that predictions are not independent and that higher-level decisions can provide useful information for lower-level classification tasks. This idea has become particularly relevant in abusive language, sexism and hate speech detection due to the emergence of datasets with hierarchical annotation schemes, where coarse-grained categories are progressively refined into more specific labels. For example, EDOS introduced a hierarchical taxonomy for sexism detection, covering binary sexism detection, coarse-grained category classification and fine-grained subtype identification [4]. In such settings, treating each label as an isolated classification problem may ignore important relationships between tasks and produce inconsistent predictions across different levels of the taxonomy.

Several strategies have been proposed to incorporate dependencies between labels, including hierarchical prediction pipelines, local classifiers associated with different taxonomy levels and classifier chains, where predictions from previous labels are used as additional inputs for subsequent classifiers [20]. In sexism detection, these ideas have been explored through systems that jointly model different levels of the annotation scheme. For example, EDOS has been addressed using a multi-task framework that learns several sexism-related subtasks while leveraging label descriptions, improving over a fine-tuned DeBERTa-V3 baseline [21]. These works suggest that exploiting hierarchical and label-dependency information can improve the consistency of fine-grained abusive language classification systems.

Research has also highlighted the importance of addressing class imbalance and calibration issues in multilabel classification settings. Class imbalance is a common challenge in abusive language, hate speech and sexism detection datasets, where some categories are substantially less represented than others. This situation may bias models towards majority classes and negatively affect performance on minority labels, particularly when evaluation metrics such as macro F1-score are used. To mitigate this problem, several cost-sensitive learning strategies have been proposed, including class weighting, focal loss and class-balanced loss functions that assign higher importance to underrepresented categories during training [5]. These approaches have become common in hate speech and offensive language detection tasks due to the highly skewed label distributions typically found in real-world datasets.

Beyond class imbalance, recent studies have emphasized the importance of prediction calibration and decision threshold optimization in multilabel classification. Transformer-based classifiers commonly produce confidence scores that must be converted into binary predictions through thresholding. While a default threshold of 0.5 is frequently adopted, several works have shown that label-specific thresholds can substantially improve macro F1-score performance by adapting decision boundaries to the characteristics and prevalence of each category [22]. This is particularly relevant in multilabel hate speech and sexism detection tasks, where labels often exhibit different levels of frequency, ambiguity and separability. Consequently, threshold optimization has become a common post-processing strategy for improving performance without modifying the underlying model architecture.

Inspired by these developments, we formulate misogyny detection in song lyrics as a hierarchical multi-task learning problem. Our approach jointly models three related subtasks: misogyny detection, misogyny type classification and stereotype identification. A shared Transformer encoder is optimized to learn common representations across tasks, while task-specific prediction heads capture the particular characteristics of each classification objective. In addition, hierarchical dependencies between tasks are

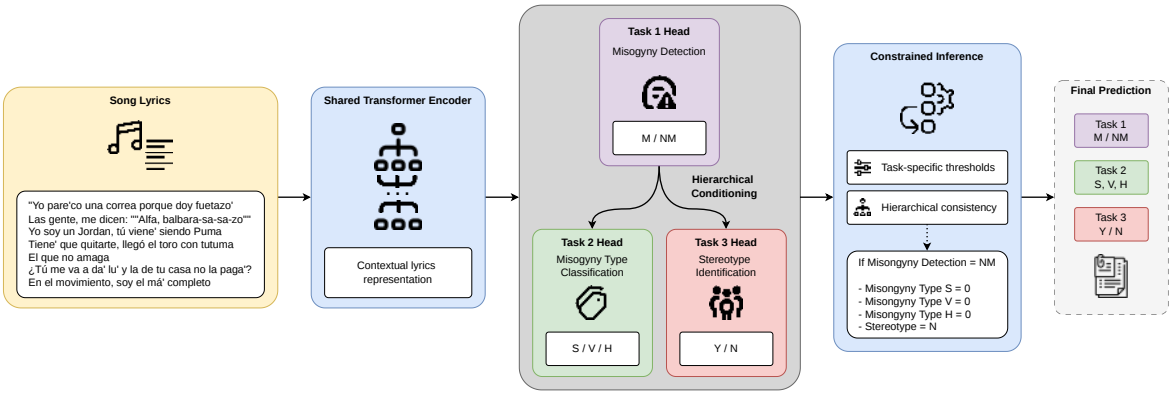

Figure 1: Overview of the proposed hierarchical multi-task system for MiSonGyny 2026. Song lyrics are encoded using a shared Transformer encoder. The resulting contextual representation is processed by three task-specific heads for misogyny detection, misogyny type classification and stereotype identification. The misogyny detection output is used as hierarchical information for the remaining subtasks. During inference, task-specific thresholds and consistency constraints are applied to produce the final predictions.

incorporated by conditioning lower-level predictions on higher-level decisions, reflecting the structure of the annotation scheme. To further address the challenges of imbalanced and multilabel classification, the proposed system combines class-weighted training with task-specific threshold optimization and hierarchical prediction constraints.

3. System Overview

Figure 1 provides an overview of the system developed by UMUTeam for the MiSonGyny 2026 shared task. The proposed pipeline is designed to address the three subtasks jointly by modeling misogyny detection in song lyrics as a hierarchical multi-task classification problem. The system follows a single-encoder architecture in which a shared Transformer model produces a contextual representation of the input lyrics. This representation is then used by three task-specific prediction heads corresponding to misogyny detection, misogyny type classification and stereotype identification. In order to exploit the structure of the task, the predictions for the second and third subtasks are conditioned on the output of the main misogyny detection head.

The first stage of the pipeline consists of preparing the song lyrics and their annotations. Each instance is represented by the lyrics text and the labels associated with the three subtasks. For Subtask 1, the original labels are mapped to a binary format, where misogynistic lyrics are assigned to the positive class and non-misogynistic lyrics to the negative class. For Subtask 2, the labels corresponding to sexualization, violence and hate are represented as a three-dimensional multilabel vector. Missing values in this subtask are converted to zero, since non-misogynistic instances cannot contain any misogyny type according to the task definition. Following the annotation structure of the dataset, missing stereotype labels in Subtask 3 are also converted to the negative class. The lyrics are minimally preprocessed by replacing missing text with empty strings, while preserving the original textual content.

The second stage consists of encoding the lyrics with a shared Transformer encoder. In our implementation, we use `RigoBERTa-large` [23], a Spanish pretrained encoder based on the RoBERTa/XLM-R architecture. The pretrained encoder model was used to extract the contextual representation associated with the first token of the sequence as the sentence-level representation. This representation is passed through dropout and then shared across the three classification heads. The use of a common encoder allows the system to learn a single representation of each song that can support all subtasks, while the task-specific heads specialize this representation for each prediction objective.

The third stage corresponds to the hierarchical multi-task prediction module. The first head predicts whether the lyrics are misogynistic or non-misogynistic. This output acts as the main decision of the hierarchy. The second head performs multilabel misogyny type classification, predicting the presence

of sexualization, violence and hate. The third head predicts whether the lyrics contain gender-based stereotypes. To incorporate the hierarchical dependency between the main task and the remaining subtasks, the logit produced by the misogyny detection head is concatenated with the shared lyrics representation and used as input to the Task 2 and Task 3 heads. This design allows the lower-level heads to take into account the model's confidence about the presence of misogyny when predicting more specific or related labels.

During training, the model is optimized jointly across the three subtasks. Each prediction head is trained using binary cross-entropy with logits, which is suitable for both binary and multilabel classification. To reduce the impact of class imbalance, positive class weights are computed from the training partition and incorporated into the corresponding binary cross-entropy losses. The overall objective is obtained by summing the losses of the three subtasks. Although different task-level loss weights were explored during development, the final configuration assigns equal importance to the three task losses. The model is trained using AdamW and a linear learning-rate scheduler with warmup, and the best checkpoint is selected according to the mean macro F1-score across the three subtasks on the validation set.

After training, the selected model is used to generate predictions for the official test set. The logits produced by each head are converted into probabilities using the sigmoid function. Since the three subtasks have different label distributions and decision requirements, task-specific thresholds are applied instead of using a single default threshold for all outputs. The thresholds are tuned on the validation set to maximize the mean macro F1-score across subtasks. The final thresholds are applied independently to the misogyny detection head, the three misogyny type labels and the stereotype identification head.

The final stage of the pipeline applies hierarchical constraints to ensure consistent predictions across subtasks. In particular, when the system predicts an instance as non-misogynistic in Subtask 1, the three misogyny type labels of Subtask 2 are forced to zero and the stereotype prediction of Subtask 3 is forced to the negative class. This post-processing step reflects the hierarchical structure of the task and prevents incompatible outputs, such as assigning a misogyny type to lyrics predicted as non-misogynistic.

4. Experimental Setup

Our experiments were conducted on the dataset provided by the MiSonGyny 2026 shared task. Each instance consists of the lyrics of a song and the annotations associated with the three subtasks: misogyny detection, misogyny type classification and stereotype identification. In order to monitor model performance during development and select the best configuration, the official training set was divided into internal training, development and test partitions. The split was stratified according to the Subtask 1 label, since misogyny detection defines the main level of the hierarchy. Table 1 reports the main statistics of the dataset and the internal partitions used in our experiments.

As shown in Table 1, the dataset presents a hierarchical and imbalanced label structure. In the official training set, non-misogynistic (NM) lyrics are more frequent than misogynistic (M) ones, with 1,294 NM instances and 1,009 M instances. Moreover, the annotations show a strict dependency between Subtask 1 and the remaining subtasks: instances labeled as NM in Subtask 1 do not contain positive labels in either Subtask 2 or Subtask 3. This confirms that misogyny type classification and stereotype identification are only meaningful when misogynistic discourse is present.

Subtask 2 characterizes misogynistic discourse through the sexualization, violence and hate labels, and its distribution is clearly uneven. Sexualization is the most frequent category, with 804 positive instances in the official training set, followed by hate with 589 instances and violence with only 314 instances. This indicates that the multilabel subtask is not only dependent on the main misogyny detection label, but also internally imbalanced across its three categories. Subtask 3 introduces an additional binary decision related to the presence of gender-based stereotypes, with 687 positive instances in the official training set.

The internal training, development and test partitions preserve similar proportions with respect to

Table 1
Statistics of the MiSonGyny 2026 dataset. Songs denotes the number of instances, W. denotes the average number of words per lyrics text, M and NM denote misogynistic and non-misogynistic instances, respectively. S, V and H correspond to the number of positive labels for sexualization, violence and hate in Subtask 2, and St. denotes the number of positive stereotype labels in Subtask 3. Label counts for the official test set are not available.

Split	Songs	W.	M	NM	S	V	H	St.
Official partitions								
Official Training Set	2303	348.05	1009	1294	804	314	589	687
Official Test Set	576	338.54	–	–	–	–	–	–
Internal split of the official training set								
Internal Training Set	1612	345.69	706	906	565	219	415	484
Internal Development Set	345	348.26	151	194	115	48	88	97
Internal Test Set	346	358.81	152	194	124	47	86	106

the main misogyny label, since the split was stratified according to Subtask 1. Overall, this setting motivates the use of a hierarchical multi-task architecture, class-weighted training, task-specific threshold optimization and hierarchical consistency constraints during inference.

The proposed model was trained as a supervised hierarchical multi-task classifier. We used `RigoBERTa-large` as the shared Transformer encoder of the system. The lyrics were tokenized with the corresponding tokenizer, truncated to a maximum length of 512 tokens and dynamically padded at batch level. The contextual representation associated with the first token of the sequence was used as the input representation for the classification module. This representation was passed through a dropout layer and then fed into the task-specific heads described in Section 3.

The model was optimized jointly across the three subtasks. Subtask 1 and Subtask 3 were treated as binary classification problems, while Subtask 2 was modeled as a multilabel classification problem with three independent outputs corresponding to sexualization, violence and hate. Each output was trained using binary cross-entropy with logits. To mitigate label imbalance, positive class weights were computed from the internal training partition and incorporated into the corresponding loss functions. The final training objective was obtained by summing the three task losses with equal task-level weights.

The official training set was split into internal training, development and test partitions using a fixed random seed of 42. The development partition was used for hyperparameter selection and threshold tuning, while the internal test partition was used to estimate the behavior of the selected configuration before generating the official predictions. Hyperparameter selection was performed through grid search on the internal development set. The explored search space included the following values: batch size $\in [8, 16]$, number of training epochs $\in [5, 10]$, learning rate $\in [2\times10^{-5}, 3\times10^{-5}, 5\times10^{-5}]$, weight decay $\in [0.01, 0.001]$, warmup ratio $\in [0, 0.1]$ and dropout rate $\in [0, 0.1, 0.2, 0.3]$. The final configuration was selected based on the mean macro F1-score across the three subtasks on the internal development set. The best configuration used a batch size of 16, 10 training epochs, a learning rate of 2×10^{-5}, a weight decay of 0.01, a warmup ratio of 0.1 and a dropout rate of 0.2. Optimization was performed using AdamW and a linear learning-rate scheduler with warmup.

After training, the best checkpoint was used to generate predictions for the official test set. For each instance, the model produced one logit for Subtask 1, three logits for Subtask 2 and one logit for Subtask 3. These logits were converted into probabilities using the sigmoid function. Instead of applying the default threshold of 0.5 to all outputs, task-specific thresholds were tuned on the internal development set in order to maximize the mean macro F1-score across subtasks. The explored threshold values were $[0.3, 0.4, 0.5, 0.6, 0.7]$ for Subtask 1, for each of the three labels of Subtask 2, and for Subtask 3. The selected thresholds were 0.7 for Subtask 1, $(0.6, 0.4, 0.4)$ for the sexualization, violence and hate labels of Subtask 2, and 0.5 for Subtask 3.

The inference stage also included a hierarchical post-processing step to enforce consistency between

Table 2
Official MiSonGyny 2026 ranking for the three subtasks according to F1-score.

Subtask 1			Subtask 2			Subtask 3		
Place	**Team**	**F1**	**Place**	**Team**	**F1**	**Place**	**Team**	**F1**
1	**tbernal**	**0.8856**	**1**	**tbernal**	**0.7385**	**1**	**tbernal**	**0.8283**
2	ikanate	0.8712	2	manuelsanz	0.7079	2	ikanate	0.8109
3	aerojash	0.8581	3	vramos	0.6939	3	aerojash	0.8071
4	manuelsanz	0.8559	4	cicese-lcdaa	0.6846	4	franrodrigo	0.8039
4	vramos	0.8559	5	ikanate	0.6813	5	manuelsanz	0.8015
5	cicese-lcdaa	0.8533	6	arjun108	0.6737	5	vramos	0.8015
6	franrodrigo	0.8448	7	aerojash	0.6699	6	cicese-lcdaa	0.7924
7	arjun108	0.8434	8	amisadai_ed	0.6683	7	amisadai_ed	0.7860
8	VerbaNexAI Lab	0.8417	9	franrodrigo	0.6645	8	VerbaNexAI Lab	0.7798
9	sergiomarinnx	0.8334	10	sergiomarinnx	0.6562	9	sergiomarinnx	0.7724
10	amisadai_ed	0.8219	11	misongyny	0.5015	10	misongyny	0.7206
11	misongyny	0.7931	12	VerbaNexAI Lab	0.4022	11	arjun108	0.6366

subtasks. If an instance was predicted as non-misogynistic in Subtask 1, all Subtask 2 labels were set to zero and the Subtask 3 prediction was set to the negative class. This constraint prevents the system from assigning a misogyny type or a stereotype label to instances classified as non-misogynistic.

5. Results

In this section, we present and analyze the official results obtained by the our system in the MiSonGyny 2026 shared task. The evaluation was carried out independently for the three subtasks using the official F1-score metric. Our final submission corresponded to the hierarchical multi-task system described in Section 3, using the selected hyperparameters, task-specific thresholds and hierarchical consistency constraints described in Section 4.

Table 2 reports the official ranking for the three subtasks. Our system (*tbernal*) ranked first in all of them, obtaining an F1-score of **0.8856** in Subtask 1, **0.7385** in Subtask 2 and **0.8283** in Subtask 3.

As shown in Table 2, the proposed system achieved the best result in the three evaluation settings. The highest absolute score was obtained in Subtask 1, where the model reached an F1-score of 0.8856 for binary misogyny detection. This result suggests that the shared Transformer encoder was able to capture the main distinction between misogynistic and non-misogynistic lyrics with high reliability. The system also ranked first in Subtask 3, obtaining an F1-score of 0.8283 for stereotype identification, which indicates that the model was effective not only in detecting explicit misogyny but also in identifying gender-based stereotypical content.

Subtask 2 was the most challenging setting, as reflected by the lower scores obtained by all participating systems. This result is expected, since this subtask requires multilabel classification over three different types of misogynistic discourse: sexualization, violence and hate. In addition, as shown in Table 1, the three labels are unevenly distributed, with violence being the least frequent category. Despite this difficulty, our system achieved an F1-score of 0.7385 and obtained the largest margin over the second-ranked submission among the three subtasks, with an improvement of 0.0306 F1-score points. This suggests that the combination of hierarchical conditioning, class-weighted training and task-specific threshold tuning was particularly useful for the fine-grained multilabel classification setting.

A more detailed interpretation of Subtask 2 suggests that the main difficulty is likely associated with the uneven distribution and semantic heterogeneity of the three misogyny types. As reported in Table 1, violence is the least frequent category in the official training set, with only 314 positive instances, compared with 804 instances for sexualization and 589 for hate. This imbalance may limit the ability of the model to learn robust representations for violent misogynistic discourse. In addition,

violence in song lyrics can be expressed through direct threats, metaphorical language, aggressive slang or narrative descriptions, making its boundaries less explicit than more lexicalized categories such as sexualization. Therefore, although the official ranking only reports the global F1-score for Subtask 2, the distribution of the training data suggests that violence is probably the most challenging label and an important source of errors in the multilabel setting.

Another important aspect concerns the hierarchical consistency rule applied during inference. In the training data, instances labeled as non-misogynistic in Subtask 1 do not contain positive labels in Subtask 2 or Subtask 3, which motivated forcing all misogyny type labels and the stereotype label to the negative class when Subtask 1 was predicted as NM. This rule improves consistency with the annotation structure of the dataset and prevents incompatible outputs. However, it also introduces a strong dependency on the Subtask 1 decision: if the model incorrectly predicts an actually misogynistic song as non-misogynistic, any potentially correct positive prediction in Subtask 2 or Subtask 3 is suppressed. Moreover, this assumption would be problematic in annotation schemes where gender stereotypes are not strictly subordinate to explicit misogyny. For this reason, future systems should consider softer consistency mechanisms or confidence-aware post-processing strategies that preserve hierarchical coherence without fully discarding lower-level evidence.

The results also show that the same submission achieved first place in all three subtasks, suggesting that the proposed architecture provides a consistent solution across the different levels of the task. This is relevant because the subtasks involve different classification scenarios: binary classification in Subtasks 1 and 3, and multilabel classification in Subtask 2. The use of a shared encoder allows the system to exploit common information across tasks, while the task-specific heads and hierarchical constraints help adapt the predictions to the structure of the annotation scheme.

Overall, the official evaluation confirms the effectiveness of modeling MiSonGyny 2026 as a hierarchical multi-task classification problem. The strong performance across the three subtasks indicates that jointly learning the related objectives and enforcing consistency between them is a suitable strategy for misogyny detection in song lyrics. In particular, the results suggest that the proposed system is able to combine robust global misogyny detection with fine-grained characterization of misogynistic discourse and stereotype identification.

6. Conclusion and Future Work

In this work, we presented the system developed by UMUTeam for the MiSonGyny 2026 shared task at IberLEF 2026. Our approach formulates misogyny detection in song lyrics as a hierarchical multi-task classification problem. The proposed system relies on a shared Transformer encoder and three task-specific heads for misogyny detection, misogyny type classification and stereotype identification. In addition, the model explicitly incorporates task dependencies by conditioning the lower-level prediction heads on the output of the main misogyny detection head. The final system also includes class-weighted training, validation-based threshold optimization and hierarchical consistency constraints during inference.

The official results show that the proposed system achieved the best performance in the three subtasks. Our submission ranked first in Subtask 1 with an F1-score of 0.8856, first in Subtask 2 with an F1-score of 0.7385 and first in Subtask 3 with an F1-score of 0.8283. These results indicate that hierarchical multi-task learning is an effective strategy for modeling the different dimensions of misogynistic discourse in song lyrics. In particular, the strong performance obtained in the multilabel misogyny type classification subtask suggests that jointly learning related objectives and applying task-specific thresholds can be especially useful in fine-grained and imbalanced settings.

The results also support the usefulness of enforcing consistency between related subtasks. In the training data, non-misogynistic instances do not contain positive labels for misogyny type classification or stereotype identification. Therefore, incorporating this dependency both in the architecture and in the inference stage helps align the predictions with the annotation structure of the task. Nevertheless, this design choice also constitutes a limitation. The hierarchical rule used in our system is strict,

since an NM prediction in Subtask 1 automatically forces negative predictions in Subtasks 2 and 3. Consequently, errors in the main misogyny detection task may propagate to the remaining subtasks, suppressing potentially correct fine-grained predictions. This issue is particularly relevant for minority or ambiguous categories, such as violence, and for possible annotation settings in which stereotypes are not defined as strictly dependent on misogyny.

Despite these results, several aspects remain open for future work. First, the current system uses a single shared encoder and a simple hierarchical conditioning mechanism based on the Subtask 1 logit. Future work could explore richer forms of task interaction, such as attention-based task conditioning, label dependency modeling or classifier-chain variants that explicitly propagate information between all subtasks. Second, future work should analyze the behavior of the model at the level of individual misogyny types, with special attention to violence, which is the least represented category in the training data and may involve highly diverse linguistic realizations. Third, although the strict hierarchical rule improved consistency with the dataset annotation scheme, alternative soft-constraint strategies could be explored to reduce error propagation from Subtask 1 to the remaining subtasks. In addition, future work could explore the use of larger or domain-adapted language models, as well as the integration of external lexical or cultural knowledge sources related to gender stereotypes and misogynistic discourse in music.

Finally, the task opens interesting possibilities for extending misogyny detection beyond song-level classification. Future systems could incorporate phrase-level or span-level evidence extraction in order to identify which fragments of the lyrics contribute most to each prediction. This would improve interpretability and provide more useful outputs for analyzing how misogyny, sexualization, violence, hate and stereotypes are expressed in musical texts.

Acknowledgments

This work is part of the research project LaTe4PoliticES (PID2022-138099OB-I00) funded by MICIU/AEI/10.13039/501100011033 and the European Regional Development Fund (ERDF/EU - FEDER/UE)-a way of making Europe. Mr. Tomás Bernal-Beltrán is supported by University of Murcia through the predoctoral programme.

Declaration on Generative AI

During the preparation of this work, the authors used DeepL for grammatical and spelling correction. After using this tool, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, Acm Computing Surveys (Csur) 51 (2018) 1–30.

[2] S. Kiritchenko, I. Nejadgholi, K. C. Fraser, Confronting abusive language online: A survey from the ethical and human rights perspective, Journal of Artificial Intelligence Research 71 (2021) 431–478.

[3] L. Betti, C. Abrate, A. Kaltenbrunner, Large scale analysis of gender bias and sexism in song lyrics, EPJ data science 12 (2023) 10.

[4] H. Kirk, W. Yin, B. Vidgen, P. Röttger, Semeval-2023 task 10: Explainable detection of online sexism, in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 2193–2210.

[5] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, S. Belongie, Class-balanced loss based on effective number of samples, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 9268–9277.

[6] T. Alcántara, E. Urios Alacreus, C. Macías, M. Aloy Mayo, O. Garcia-Vazquez, A. Espinosa-Juarez, M. Soto, P. Rosso, J.-E. Valdez-Rodríguez, H. Calvo, Overview of MiSonGyny at IberLEF 2026: Misogyny Speech Detection in Mexican and Spanish Language Song Lyrics, Procesamiento del Lenguaje Natural To be announced (2026).
[7] A. Bonet-Jover, J. Á. González-Barba, L. Chiruzzo, Overview of IberLEF 2026: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR-WS. org, 2026, pp. –.
[8] 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) 441–451.
[9] E. Fersini, P. Rosso, M. Anzovino, et al., Overview of the task on automatic misogyny identification at ibereval 2018., Ibereval@ sepln 2150 (2018) 214–228.
[10] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, T. Donoso, Overview of exist 2021: sexism identification in social networks, Procesamiento del Lenguaje Natural 67 (2021) 195–207.
[11] R. Caruana, Multitask learning, Machine learning 28 (1997) 41–75.
[12] X. Liu, P. He, W. Chen, J. Gao, Multi-task deep neural networks for natural language understanding, in: Proceedings of the 57th annual meeting of the association for computational linguistics, 2019, pp. 4487–4496.
[13] R. Pan, J. A. García-Díaz, R. Valencia-García, Spanish mtlhatecorpus 2023: Multi-task learning for hate speech detection to identify speech type, target, target group and intensity, Computer Standards & Interfaces 94 (2025) 103990.
[14] H. Gómez-Adorno, G. Bel-Enguix, H. Calvo, S. Ojeda-Trueba, S. T. Andersen, J. Vásquez, T. Alcántara, M. Soto, C. Macias, 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.
[15] P. Saha, B. Mathew, P. Goyal, A. Mukherjee, Hateminers: Detecting hate speech against women, arXiv preprint arXiv:1812.06700 (2018).
[16] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 4171–4186.
[17] T. Bernal-Beltrán, R. Pan, J. A. García-Díaz, M. del Pilar Salas-Zárate, M. A. Paredes-Valverde, R. Valencia-García, Detection of maliciously disseminated hate speech in spanish using fine-tuning and in-context learning techniques with large language models, Computers, Materials, & Continua 87 (2026).
[18] T. Bernal-Beltrán, M. A. Paredes-Valverde, M. d. P. Salas-Zárate, J. A. García-Díaz, R. Valencia-García, Sentiment analysis in mexican spanish: A comparison between fine-tuning and in-context learning with large language models, Future Internet 17 (2025) 445.
[19] F. M. Plaza-del Arco, S. Halat, S. Padó, R. Klinger, Multi-task learning with sentiment, emotion, and target detection to recognize hate speech and offensive language, arXiv preprint arXiv:2109.10255 (2021).
[20] J. Read, B. Pfahringer, G. Holmes, E. Frank, Classifier chains: A review and perspectives, Journal of Artificial Intelligence Research 70 (2021) 683–718.
[21] J. Goldzycher, Cl-uzh at semeval-2023 task 10: Sexism detection through incremental fine-tuning and multi-task learning with label descriptions, in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 1562–1572.
[22] Z. C. Lipton, C. Elkan, B. Narayanaswamy, Thresholding classifiers to maximize f1 score, arXiv preprint arXiv:1402.1892 (2014).
[23] I. de Ingeniería del Conocimiento, Rigoberta-2.0, 2025. URL: https://huggingface.co/IIC/RigoBERTa-2.0. doi:10.57967/hf/7048.