Sintax Squad at MiSonGyny 2025: Transformer Models for Misogyny Detection in Spanish Lyrics

Luis Ramos[1,*,†], Irari Jiménez-López[1,†], Yoqsan Angeles[1,†] and Olga Kolesnikova[1]

[1]Centro de Investigación en Computación, Instituto Politécnico Nacional, Mexico City, Mexico

Abstract

The rapid growth of digital streaming services has made music more accessible than ever, but that same reach has allowed damaging lyrics reinforcing gender stereotypes to travel far beyond local scenes. Many of these messages do not arise solely from individual prejudice; they stem from a structured system of misogyny that punishes women who refuse narrow definitions of femininity and blames them for any disruption to the status quo. Identifying and analysing such harmful content is essential if researchers, educators, and audiences are to minimise its impact and move toward a fairer media environment. To that end, this paper outlines an experimental pipeline that applies state-of-the-art transformer models for the task of detecting misogynistic language in song lyrics. For the MiSonGyny 2025 shared task at IberLEF 2025, two clear tasks were established: first, a binary classification for lyrics as Misogynist or Not Misogynist; and second, a fine-grained, multiclass scheme that sorts lyrics into Sexualization, Violence, Hate, or Not Related. In pursuit of this goal, different transformer models were evaluated, which resulted in a Macro F1 Score of 0.7762 for binary classification and 0.4947 for multiclass classification.

Keywords

Misogyny, Lyrics, Spanish, Transformers, Fine-Tuning, NLP, Classification

1. Introduction

With the increasing reach of the internet, the spread of harmful content has become a significant societal concern [1]. One of the most widely consumed forms of media is music, whose accessibility has grown through digital streaming platforms. Music serves as a powerful medium for expressing emotions, opinions, and experiences. While many songs focus on themes such as love, sadness, or joy, others include lyrics that convey harmful or discriminatory messages. Of particular concern is the presence of gender based stereotypes and social expectations, which may be transmitted either intentionally or unintentionally through song lyrics [2]. Detecting such content is essential to promoting healthier media environments.

As argued by Manne et al. [3], misogyny should not be understood merely as personal hostility toward women, but rather as a structural system that enforces gendered expectations within patriarchal societies. It manifests through the sanctions and social hostility that women experience when deviating from prescribed roles. In this context, misogynistic lyrics contribute to reinforcing these norms by disseminating hostile or demeaning messages toward women.

Recent studies have explored sentiment analysis in both song lyrics and social media texts. For instance, Barat et al. [4] addressed the classification of objectionable and suitable song lyrics using weakly supervised learning. They compared six different models, including Deep Learning (DL), Random Forest (RF), and Support Vector Machine (SVM). Logistic regression combined with DistilBERT embeddings achieved the highest F-score of 67.8. Similarly, Chen et al.(2025) [2] conducted a large scale analysis of gender bias in English 0. language song lyrics. Using topic modelling and bias measurement

IberLEF 2025, September 2025, Zaragoza, Spain

*Corresponding author.

†These authors contributed equally.

lramos2020@cic.ipn.mx (L. Ramos); ijimenezl2024@cic.ipn.mx (I. Jiménez-López); yangelesg2020@cic.ipn.mx (Y. Angeles); kolesnikova@cic.ipn.mx (O. Kolesnikova)

0009-0008-5586-6668 (L. Ramos); 0009-0007-1171-3435 (I. Jiménez-López); 0009-0004-0886-5540 (Y. Angeles); 000-0002-1307-1647 (O. Kolesnikova)

techniques such as BERTopic, they categorized over half a million songs into six themes, including Pleasant, Strength, and Weakness. Their findings revealed an increasing trend of sexualization and an implicit gender bias—where terms related to weakness and appearance were associated with women, while intelligence and strength were more often linked to men.

Some researchers have explored the sentiment analysis in lyrics and text in social media. Barat et al. [4] classified objectionable and suitable song lyrics using weakly supervised learning and compared 6 different models including DL, RF and SVM. They observed that Logistic Regression with DistilBert embedding obtained the best result with 67.8 in F-score. Furthermore, Chen et al. [2] presented an analysis of gender bias in English song lyrics. They used topic modelling and bias measurement techniques. They used BERTopic in a dataset of more than half a million songs into six different classes including Pleasant, Strength, or Weakness. They found an increase of sexualization over time. Their results also revealed an implicit gender bias in song lyrics. For example, with Weakness and Appearance words showing a female bias, while Intelligence and Strength words exhibit a male bias. Calderón-Suarez et al. [5] labelled songs based on lyrics containing two kind of words: misogynistic words (according to lexicons in English and Spanish) and words related to women. These song lyrics helped to enhance the models classifying misogynic content in social networks. The best accuracy with the Iber-Sp dataset was 0.824, obtained using Linear Regression and Twitter texts with single phrases in songs.

This research explores the use of transformers to identify misogynic lyrics. Transformers are a type of neural network architecture specifically designed to handle sequential data [6]. They are particularly effective in Natural Language Processing tasks, but have also proven useful in other domains. What sets them apart is their use of attention mechanisms instead of recurrences. This allows the network to process the entire sequence simultaneously, assigning different attention weights to each element based on its relevance. In the original formulation, the architecture is composed of an encoder-decoder structure. Additionally, since the architecture lacks intrinsic order awareness, positional encodings are added to provide information about the position of elements in the sequence.

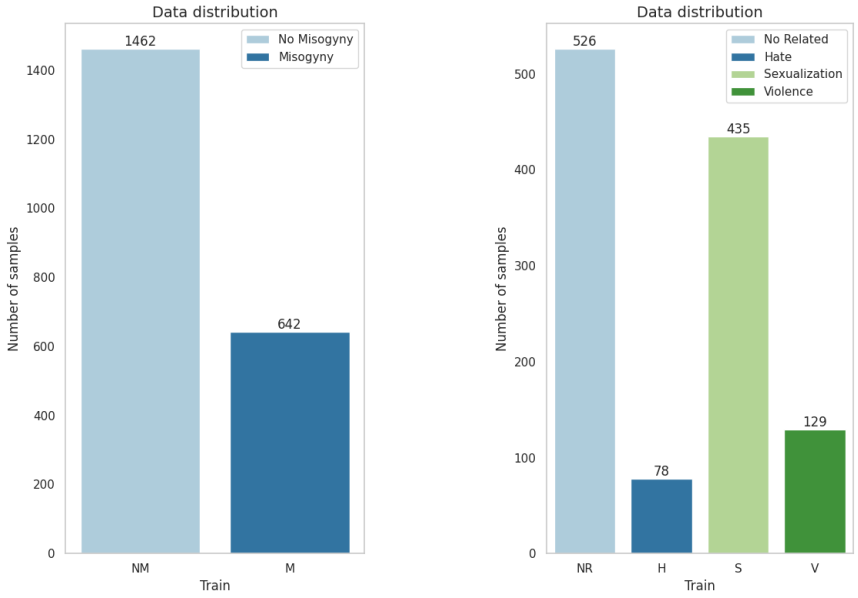
This proposal aims to solve two tasks for MiSonGyny 2025 [7] in the IberLEF 2025 [8]:

- Task1: The aim of the task is to develop a classifier that identifies song lyrics as Misogynist (M) or Not Misogynist (NM). A phrase is marked as M if it is an incursion of misogynistic hate speech; that is, an utterance which includes insipid insults or derogatory language aimed at women, or claims which perpetuate harmful gender stereotypes that subordinate or objectify women. Those marked NM may refer to women and attributes no negative or stereotypical content
- Task2: The aim of the task is to develop a multiclass classifier for the identification of types of speech within song lyrics, demarcating each song as Sexualization (S), Violence (V), Hate (H), or Not Related (NR). Sexualization (S) covers suggestive or descriptive sexually explicit phrases. Violence (V) marks physical and verbal acts of aggression or threat. Hate (H) involves pejorative or bigoted phrases aimed at a person or group. NR incorporates everything else, that is, phrases devoid of sexual, violent, or hateful content.

2. Data description

The dataset contains Spanish song lyrics, its statistics for the both task are presented in Figure 1. Here, we can notice that there is an imbalance between classes; the imbalance ratio (IR) of the first database is 2.27 and of the second dataset is 6.74. An increased imbalance ratio IR results in a greater extent imbalance of the dataset, making it harder to classify datasets with higher IR [9]. Table 1 shows some examples from the first task. The first row corresponds to a Not Misogynist song from the singer *Thalía*. The second row corresponds to a Misogynist song from the singer *Frankie Ruíz*. Table 2 shows some examples of the second task. The first row corresponds to a Not Related song from *Camilo Sesto*. The second row corresponds to a Hate song from *Cartel de Santa*. The second row corresponds to a Sexualization song from *Chris Jedi*. Finally, the fourth row corresponds to a Violence song from *Los Rieleros del Norte*. Tables 3 and 4 are the English translations of tables 1 and 2 respectively.

The predictions of the first task have 527 records, while the second task has 293 examples.



(a) (b)

Figure 1: Data Statistics

	Lyrics	Label
1	No me enseñaste cómo estar sin ti ¿Y qué le digo yo a este corazón? Si tu te has ido y todo ...	NM
2	Por vivir, a tu manera. Entregandole a cualquiera. Tus caricias pasajeras. Sin sentir ...	M

Table 1
Original lyric song samples for Task

	Lyrics	Label
1	Que no me falte tu cuerpo jamás, jamás. Ni el calor de tu forma de amar, jamás...	NR
2	Por que alguien mas la quiera. Enterrada muchos metros bajo de la tierra. Sin una pinche moneda	H
3	Dile a ese infeliz que te la coma como yo te la comí. Y que te haga mojarte como yo te hice venirte	S
4	Hoy te encuentras perdida, vendiendo tu vida. Y yo muy contento. Quiera Dios que tu cuerpo se seque	V

Table 2
Original lyric song samples for Task 2

	Lyrics	Label
1	You didn't show me how to be without you. And what do I say to this heart? If you have gone and all...	NM
2	To live your way. Giving yourself to anyone. Your fleeting caresses. Without feeling...	M

Table 3
Lyrics samples translated in Task 1

3. Methodology

This section outlines the specifics of the approach taken with different transformer models. Only one data cleaning phase is needed for this method before the tweets are input into transformer models. Cleaning phase and transformer model details are described in detail in the following subsections as each phase of the proposed methodology showed in Figure 2.

3.1. Data Cleaning

This phase encompasses lowercase and the removal of emojis, URLs, numerals, special symbols, words framed by parentheses or number symbols (as exemplified in Table 1), as well as stop words (using

	Lyrics	Label
1	May I never lack your body, ever, ever. Nor the warmth of your way of loving, ever...	NR
2	Because someone else want her. Buried many meters underground. Without a damn coin.	H
3	Tell that bastard to eat you out the way I ate you out. And to make you wet the way I made you come	S
4	Today you find yourself lost, selling your life. And I'm very happy. May God wants your body dry up.	V

Table 4
Lyrics samples translated in Task 2

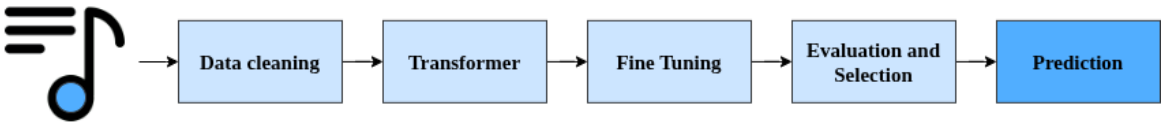


Figure 2: Methodology overview diagram

NLTK library). Additionally, lemmatization of words was applied utilizing Spacy library. This phase prepares the text for further analysis and modelling [10].

3.2. Transformer Models

A transformer model consists of encoders and decoders with softmax normalization applied to them. Inputs are embedded with positional encodings to capture contextual meaning. Contextual meaning is processed by multi-head attention mechanisms and feed-forward networks in parallel. The decoder uses masked multi-head attention that incorporates encoder-derived attention vectors, subsequently applying feed-forward and linear layers to output probabilities [11]. Transformer architectures utilize self-attention mechanisms to dynamically capture context at different levels [12]. Due to the more complicated architecture and the larger size of transformer parameters, numerous modern fine-tuning methods are proposed to adapt to downstream different tasks efficiently and effectively[13], such as hate speech [14], toxic speech [15], homophobic language [16], hope speech [17], social support [18]. Diverse transformers models from Hugging Face [1] were selected for this proposal. The corresponding models with an ID is described in Table 5. Every model was trained using the following specific parameter values and the rest in default: 5 epochs, train and eval batch size of 16, *load_best_model_at_end = True* and *metric_for_best_model='f1'*. The last two parameters selects automatically the best model based on the best Macro F1 score.

ID	Model
M1	nlptown/bert-base-multilingual-uncased-sentiment
M2	papluca/xlm-roberta-base-language-detection
M3	pysentimiento/robertuito-sentiment-analysis
M4	lxyuan/distilbert-base-multilingual-cased-sentiments-student
M5	finiteautomata/beto-sentiment-analysis
M6	citizenlab/distilbert-base-multilingual-cased-toxicity
M7	textdetox/xlmr-large-toxicity-classifier-v2
M8	glombardo/misogynistic-statements-classification-model
M9	JonatanGk/roberta-base-bne-finetuned-cyberbullying-spanish
M10	LaProfeClaudis/LGBeTO_detection_Model

Table 5
Models

[1] https://huggingface.co/

4. Results

For the development set, we offer only macro metrics since the primary comparative performance metric is based on macro F1 score. However, for the best models performance on the test set, we only report macro F1 score. The two best models were trained with ten epochs to see if the performance improved. The results of these models are identified as "M# - 10". For Task 1, the results obtained with the ten models on training set is reported in Table 6. The results with testing set is in Table 7. For Task 2, the results with the training set is shown in Table 8. The results with testing set is in Table 9. In the tables, there is an M or W next to the metrics. M is for Macro and W is for Weight metrics.

Model	Acc.	Prec. M.	Rec. M.	F1 M.	Prec. W.	Rec. W.	F1 W.
M1	0.9810	0.9837	0.9715	0.9773	0.9812	0.9810	**0.9809**
M1 - 10	0.9886	0.9900	0.9831	0.9865	0.9886	0.9886	0.9886
M2	0.6949	0.3474	0.5000	0.4100	0.4828	0.6949	0.5698
M3	0.8902	0.8868	0.8485	0.8642	0.8895	0.8902	0.8874
M4	0.9705	0.9769	0.9539	0.9644	0.9713	0.9705	0.9702
M5	0.9933	0.9953	0.9891	0.9921	0.9934	0.9933	**0.9933**
M5 - 10	0.9905	0.9928	0.9849	0.9887	0.9906	0.9905	**0.9905**
M6	0.9876	0.9889	0.9819	0.9853	0.9877	0.9876	0.9876
M7	0.6949	0.3474	0.5000	0.4100	0.4828	0.6949	0.5698
M8	0.9121	0.8967	0.8957	0.8962	0.9120	0.9121	0.9120
M9	0.7029	0.7706	0.5150	0.4432	0.7436	0.7029	0.5914
M10	0.9923	0.9941	0.9892	0.9916	0.9930	0.9929	0.9929
M10 - 10	0.9986	0.9990	0.9977	0.9984	0.9986	0.9986	**0.9986**

Table 6
Results with training set for Task 1

Model	F1 Macro
M10 - 10	0.7762
M5	0.7691
M1	0.7610
M5 - 10	0.7435

Table 7
Results with testing set for Task 1

Model	Acc.	Prec. M.	Rec. M.	F1 M.	Prec. W.	Rec. W.	F1 W.
M1	0.469	0.7989	0.5811	0.5802	0.8296	0.8305	0.7987
M2	0.789	0.3677	0.4440	0.4020	0.6038	0.7329	0.6616
M3	0.200	0.9315	0.8667	0.8937	0.9460	0.9469	0.9439
M4	0.354	0.8588	0.7676	0.8005	0.8949	0.9007	0.8921
M5	0.159	0.9265	0.8867	0.9045	0.9582	0.9598	**0.9583**
M5 - 10	0.005	0.9995	0.9981	0.9988	0.9991	0.9991	0.9991
M6	0.272	0.8677	0.7639	0.7794	0.9154	0.9135	0.9044
M7	1.159	0.1126	0.2500	0.1553	0.2028	0.4503	0.2797
M8	0.669	0.3861	0.4683	0.4232	0.6348	0.7714	0.6964
M9	0.900	0.3526	0.4220	0.3828	0.5784	0.7003	0.6313
M10	0.094	0.9562	0.9425	0.9492	0.9773	0.9777	**0.9774**
M10 - 10	0.004	0.9995	0.9981	0.9988	0.9991	0.9991	**0.9991**

Table 8
Results with training set for Task 2

Model	F1 Macro
M10 - 10	0.4947
M5 - 10	0.4546
M5	0.4723

Table 9
Results with testing set for Task 2

5. Discussion

The best performing models in this study were based on BERT and BETO architectures. Each model was fine tuned for a specific target task. Models M1 and M5 were adapted for sentiment analysis, while M10, which achieved the highest performance on both training and testing sets, was configured specifically for hate speech detection against the LGBT community.

In contrast, the models that yielded the poorest results in the binary classification task were M2 and M9. M9 was trained for cyberbullying detection, and M2 was designed for multilingual language identification across 20 languages, including English and Spanish. The low performance of these models can be attributed to their original training objectives, which are less aligned with the detection of misogynistic content.

Regarding the multiclass classification task, the lowest performance was observed with models M7 and M9. M7 was trained for multilingual toxicity detection and, in its original training, exhibited low F1 scores on the Spanish subset of the dataset. This underperformance may be due to an imbalanced dataset and a possible domain mismatch between the pretraining data and the misogyny classification task.

6. Conclusion

This study contributes to the growing of research on misogynistic and sexually explicit content in music lyrics. Given the widespread dissemination of such messages through social media; it is worth developing tools to identify their impact.

The results highlight the importance of task-specific training, particularly when the source and target tasks share semantic and contextual similarities. In this case, models originally trained on hate speech detection tasks outperformed those trained on broader objectives such as general toxicity or cyberbullying. This suggests that misogyny is more closely aligned with hate speech rather than with other forms of harmful content.

Furthermore, models trained on fewer languages or specifically tailored for Spanish consistently achieved better performance than their multilingual counterparts. This indicates that language specific optimization captures cultural and linguistic nuances in misogynistic expression.

Declaration on Generative AI

During the preparation of this work, the authors used GPT in order to: Grammar and spelling check.

References

[1] J. H. Park, P. Fung, One-step and two-step classification for abusive language detection on twitter, arXiv preprint arXiv:1706.01206 (2017).

[2] D. Chen, A. Satish, R. Khanbayov, C. Schuster, G. Groh, Tuning into bias: A computational study of gender bias in song lyrics, in: Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025), 2025, pp. 117–129.

[3] K. Manne, Down girl: The logic of misogyny, Oxford University Press, 2017.
[4] B. K. Bolla, S. R. Pattnaik, S. Patra, Detection of objectionable song lyrics using weakly supervised learning and natural language processing techniques, Procedia Computer Science 235 (2024) 1929–1942.
[5] R. Calderón-Suarez, R. M. Ortega-Mendoza, M. Montes-Y-Gómez, C. Toxqui-Quitl, M. A. Márquez-Vera, Enhancing the detection of misogynistic content in social media by transferring knowledge from song phrases, IEEE Access 11 (2023) 13179–13190.
[6] N. Patwardhan, S. Marrone, C. Sansone, Transformers in the real world: A survey on nlp applications, Information 14 (2023) 242.
[7] 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).
[8] 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.
[9] R. Zhu, Y. Guo, J.-H. Xue, Adjusting the imbalance ratio by the dimensionality of imbalanced data, Pattern Recognition Letters 133 (2020) 217–223.
[10] S. Z. Ridoy, J. Sultana, Z. F. Ria, M. A. Uddin, M. H. Rahman, R. M. Rahman, An efficient text cleaning pipeline for clinical text for transformer encoder models, in: 2024 IEEE 12th International Conference on Intelligent Systems (IS), IEEE, 2024, pp. 1–9.
[11] F. A. Acheampong, H. Nunoo-Mensah, W. Chen, Transformer models for text-based emotion detection: a review of bert-based approaches, Artificial Intelligence Review 54 (2021) 5789–5829.
[12] A. Qasim, G. Mehak, N. Hussain, A. Gelbukh, G. Sidorov, Detection of depression severity in social media text using transformer-based models, Information 16 (2025) 114.
[13] H. Li, M. Wang, S. Zhang, S. Liu, P.-Y. Chen, Learning on transformers is provable low-rank and sparse: A one-layer analysis, in: 2024 IEEE 13rd Sensor Array and Multichannel Signal Processing Workshop (SAM), IEEE, 2024, pp. 1–5.
[14] S. Mukherjee, S. Das, Application of transformer-based language models to detect hate speech in social media, Journal of Computational and Cognitive Engineering 2 (2023) 278–286.
[15] G. Damas, R. Torres Anchiêta, R. Santos Moura, V. Ponte Machado, A transformer-based tabular approach to detect toxic comments, in: Brazilian Conference on Intelligent Systems, Springer, 2024, pp. 18–30.
[16] L. Ramos, C. Palma-Preciado, O. Kolesnikova, M. Saldana-Perez, G. Sidorov, M. Shahiki-Tash, Intellileksika at homo-mex 2024: Detection of homophobic content in spanish lyrics with machine learning, in: Journal Name or Conference Proceedings Here, 2024.
[17] G. Sidorov, F. Balouchzahi, L. Ramos, H. Gómez-Adorno, A. Gelbukh, Mind-hope: Multilingual identification of nuanced dimensions of hope (2024).
[18] M. S. Tash, L. Ramos, Z. Ahani, R. Monroy, H. Calvo, G. Sidorov, et al., Online social support detection in spanish social media texts, arXiv preprint arXiv:2502.09640 (2025).