# DACKERS at MiSonGyny 2025: A Transformer Ensemble Approach for Misogyny Detection in Spanish

Carlos Díez-Fenoy[1,†], Ariel López-González[1,†] and Jorge Daniel Valle-Díaz[1,†]

[1]*Computer Science and Engineering Department, Universidad Carlos III de Madrid, Madrid, 28911, Spain*

## Abstract

This work presents a two-stage classification approach for detecting misogynistic content in Spanish song lyrics. In the first task, we addressed the binary classification of misogynistic vs. non-misogynistic lyrics. In the second, we further classified misogynistic content into four subcategories: Not Related, Violence, Sexual, and Harassment. We experimented with several transformer-based models, including BETO, RoBERTa, BERT, and LLaMA, as well as multiple ensemble configurations. In Task 1, the best performance was achieved by the Transformer Ensemble 2, with an F1-Score of 0.828, followed by RoBERTa (0.812) and BERT (0.798). In Task 2, where the classification problem was more fine-grained, the best-performing model was again an ensemble (Transformer Ensemble 2), with an F1-Score of 0.434, followed by LLaMA (0.488) and BERT (0.415). These results highlight the robustness of transformer ensembles for detecting subtle forms of misogyny in song lyrics and demonstrate the challenges of fine-grained categorization in this domain.

## Keywords

Misogyny detection, Spanish lyrics Classification, Transformer Ensemble

## 1. Introduction

Misogynistic content in song lyrics has become a growing social concern, particularly in music genres where degrading or violent references to women are normalized or even glamorized. In Spanish-speaking contexts, this issue is of particular pertinence due to the global reach and influence of certain musical styles, which are often consumed extensively by younger demographics [1]. From a societal perspective, the normalization of misogyny through music has the potential to reinforce gender stereotypes and contribute to broader patterns of symbolic violence and discrimination [2].

From a computational perspective, the automatic detection of misogynistic speech in lyrics poses several challenges [3, 4]. These include the presence of figurative language, slang, irony, and ambiguous phrasing, all of which are common in creative domains such as music. Furthermore, there is a notable scarcity of annotated datasets focused on song lyrics, especially in the Spanish language, and the few available resources tend to be highly imbalanced, which hinders the development of robust and generalisable models [1, 2].

In this study, we propose a two-stage classification pipeline for the detection of misogynistic content in Spanish-language song lyrics [5]. In the initial phase, we undertake binary classification to differentiate between misogynistic and non-misogynistic content. In the second stage of the analysis, the misogynistic lyrics are further classified into four categories: The following categories are not relevant: violence, sexual, and harassment [2].

In this study, we undertake a thorough evaluation of multiple pre-trained transformer models, namely BERT, RoBERTa, and LLaMA, in addition to a classical Vector Space Model (VSM) approach. This investigation involves a meticulous fine-tuning and evaluation process, ensuring a comprehensive assessment of the models' capabilities and limitations [6]. Furthermore, we have developed multiple ensemble strategies that combine the predictions of individual transformer models to improve robustness.

The experimental results demonstrate that transformer-based ensembles consistently outperform individual models in terms of F1-Score, particularly in the context of binary classification tasks [7, 8].

The dataset under consideration consists of manually annotated Spanish-language song lyrics and represents a novel, socially impactful resource in a low-resource language [5]. This study underscores the promise of integrating conventional and deep learning methodologies, encompassing transformer ensembles and established baselines such as VSM, to confront the intricate and culturally contingent classification dilemmas prevalent in natural language processing [9, 10, 11, 12].

## 2. Task Description

The **MiSonGyny 2025** [13] shared task focuses on detecting misogynistic content in Spanish song lyrics. It is structured into two classification tasks of increasing complexity: binary detection of misogyny and fine-grained classification of misogynistic speech types. Below, we summarize the objectives and label definitions for each task, as well as the evaluation metrics used.

### 2.1. Task 1: Misogynistic Speech Detection

The first task consists of a **binary classification** problem where the goal is to determine whether a given phrase from a song lyric contains misogynistic speech.

- **Misogynist (M):** The phrase expresses hate speech or contempt directed at women, or reinforces harmful gender stereotypes that promote subordination, objectification, or marginalization of women.
- **Not Misogynist (NM):** The phrase does not contain hate speech or contempt against women. It may refer to women without reinforcing gender stereotypes or expressing negative attitudes.

*Example:*

```
ID_Track1, "M" – this song is defined as misogynistic
```

### 2.2. Task 2: Fine-grained Misogynistic Speech Detection

The second task aims to **classify the specific type of misogynistic content** present in a phrase. This is a **multi-class classification** task, where each misogynistic phrase is assigned to one of the following categories:

- **Sexualization (S):** Phrases that describe or imply sexual acts, use sexual language, or make sexual insinuations.
- **Violence (V):** Phrases that refer to physical or verbal aggression, threats, or violent behavior.
- **Hate (H):** Phrases that include offensive, hostile, or discriminatory language, targeting women or groups through expressions of contempt or dehumanization.
- **Not Related (NR):** Phrases that do not belong to the above categories and contain no sexual, violent, or hateful content.

### 2.3. Evaluation Metrics

Both tasks are evaluated using **macro-averaged** metrics to address potential class imbalance and ensure that each class contributes equally to the final scores, regardless of its frequency in the dataset [14].

The following metrics are reported for each task:

- **Macro F1-Score**
- **Macro Precision**
- **Macro Recall**

Let $B(tp, fp, tn, fn)$ be any binary evaluation measure (e.g., precision, recall, or F1-score), where:

- $tp$: true positives
- $fp$: false positives
- $tn$: true negatives
- $fn$: false negatives

For each class label $\lambda$ in a set of $q$ labels, the metric $B$ is computed using the binary evaluation for that specific class (one-vs-rest). Then, the macro-averaged version of the metric is calculated as:

$$B_{\text{macro}} = \frac{1}{q} \sum_{\lambda=1}^{q} B(tp_\lambda, fp_\lambda, tn_\lambda, fn_\lambda)$$

This formulation ensures that all classes contribute equally, which is particularly important for datasets where some classes may be underrepresented.

### 2.4. Data Description

The datasets provided for the MiSonGyny 2025 [13] shared task include annotated Spanish song lyrics, separated by task. Each dataset consists of short textual segments (phrases or lines) labeled according to the task definitions.

For **Task 1** (binary misogyny detection), the training set contains **2,104 instances**, with a noticeable class imbalance: **642 misogynistic (M)** and **1,462 non-misogynistic (NM)** examples.

For **Task 2** (fine-grained classification), the training set includes **1,168 instances** labeled across four categories. The distribution is as follows:

- **Sexualization (S):** 435 instances
- **Violence (V):** 129 instances
- **Hate (H):** 78 instances
- **Not Related (NR):** 526 instances

These imbalances reinforce the need to evaluate models using macro-averaged metrics and highlight the importance of robust classification methods, especially for the underrepresented categories.

## 3. Models and Approaches

For both subtasks, we explore a range of models suited for natural language processing (NLP) tasks. These include transformer-based architectures as well as a traditional machine learning baseline:

- **BERT** [15, 16]: A widely used transformer model pre-trained on large corpora using a masked language modeling objective. It provides strong baseline performance for sentence-level classification tasks.
- **RoBERTa** [17, 16]: An optimized version of BERT trained with larger data and no next-sentence prediction objective, offering improved results on various NLP benchmarks.
- **BETO** [18]: A Spanish version of BERT, pre-trained on a large Spanish corpus, making it more suitable for tasks in this language.
- **LLaMA** [19]: A recent multilingual transformer-based model designed for efficiency and strong generalization across languages. We use it to assess its ability to handle subtle linguistic phenomena in Spanish lyrics.
- **SVM (Support Vector Machine) [20, 21]**: As a classical baseline, we include an SVM classifier using TF-IDF representations. This provides a useful point of comparison for transformer-based methods.

- **Random Forest (RF)** [22, 23, 24]: A tree-based ensemble method that builds multiple decision trees and aggregates their predictions. It is included as a second traditional baseline due to its robustness and interpretability in high-dimensional text classification tasks.

These models allow us to compare traditional and modern approaches to the detection and classification of misogynistic speech in Spanish song lyrics.

## 4. Methodology

This section describes the methodology used for Task 1: binary classification of misogynistic vs. non-misogynistic lyrics.

### 4.1. Task 1 Overview

The process for Task 1 (binary misogyny detection) consisted of the following main steps. Figure 1 illustrates the complete pipeline:

1. **Initial data analysis:** The original training set for Task 1 contained 2,104 examples, with a highly imbalanced class distribution: 642 labeled as Misogynistic (M) and 1,462 as Not Misogynistic (NM). The imbalance ratio was calculated and confirmed to be significant. The imbalance ratio (IR) is calculated as follows:

$$\text{IR} = \frac{\text{NM}}{\text{M}} = \frac{1462}{642} \approx 2.28 \tag{1}$$

2. **Data balancing:** To reduce bias in model training, we augmented the dataset using two external resources with similar annotation schemes:

   - *Sexism in the Lyrics of the Most Listened to Songs in Spain*, with 15,836 NM and 4,758 M examples [5, 25].
   - *Corpus of Song Lyrics in Spanish Labeled for Gender-Based Violence Against Women*, with 778 NM and 222 M examples [26, 2].

   Using a sampling strategy, we balanced the final dataset to contain **6,000 NM** and **5,622 M** instances.

3. **Preprocessing and tokenization:** All lyrics were tokenized using the default tokenizer of each transformer model (BERT, RoBERTa, BETO, and LLaMA). No additional preprocessing (e.g., lemmatization or stopword removal) was applied, to preserve the original semantics and language style of the lyrics.

4. **Model training:** Each model was trained independently using the same data split: 70% for training, 20% for validation, and 10% for testing. We used 3 to 5 epochs per model depending on the validation loss convergence. The following models were used:

   - **BERT**
   - **RoBERTa**
   - **BETO**
   - **LLaMA**
   - **Vector Space Model (VSM)** as a classical baseline

5. **Ensemble construction:** Two ensemble models were created by combining the predictions from BERT, RoBERTa, and BETO. A majority voting scheme was applied, where the final prediction corresponds to the label chosen by at least two of the three models.

This multi-step pipeline allowed us to integrate data augmentation, robust modeling with pre-trained transformers, and ensemble techniques to address the challenges of Task 1.
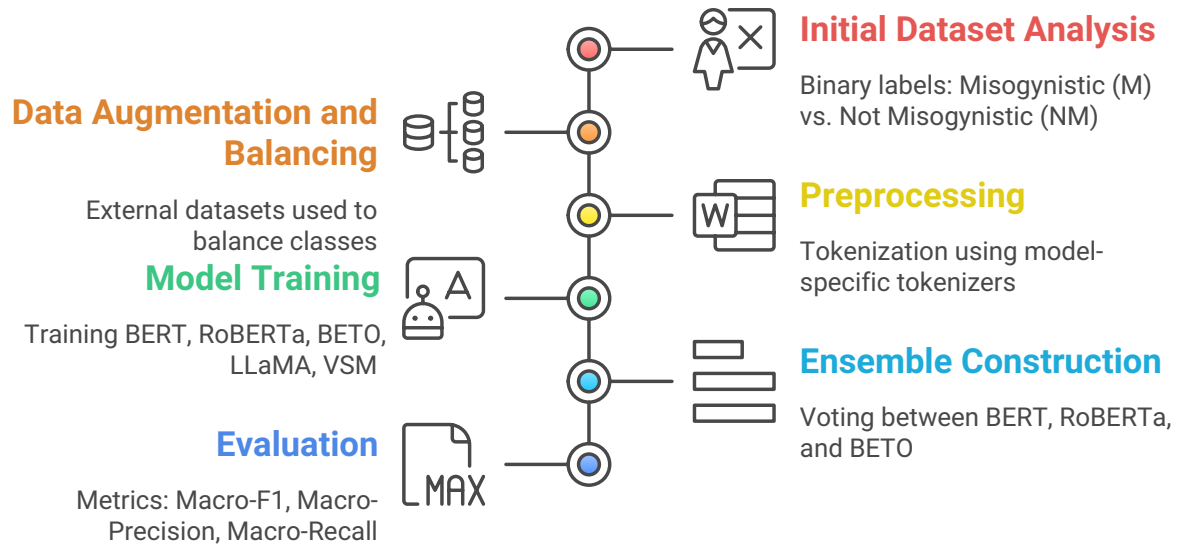
# From Dataset Analysis to Model Evaluation



**Figure 1:** Training process for task 1. *Figure constructed using napkin* (https://app.napkin.ai/).

## 4.2. Task 2 Overview

The process for Task 2 (fine-grained misogyny classification) consisted of the following main steps. Figure 2 shows the complete pipeline:

1. **Initial Dataset Analysis:** The original dataset contained the following class distribution:

   - Sexualization (S): 435 instances
   - Violence (V): 129 instances
   - Hate (H): 78 instances
   - Not Related (NR): 526 instances

   The most common class (S) had 435 examples, while the rarest (H) had 78. The resulting imbalance ratio (IR) was:

   $$\text{IR} = \frac{435}{78} \approx 5.58$$

2. **Data Augmentation with Generative AI:** To mitigate the imbalance, we used the GROK generative AI tool to synthesize additional Spanish song lyrics for the minority classes (Violence and Hate). This resulted in a more balanced dataset with the following class distribution:

   - NR: 526 instances
   - S: 435 instances
   - V: 324 instances
   - H: 273 instances

3. **Preprocessing and Tokenization:** As in Task 1, tokenization was performed using the respective tokenizers of the transformer models. No additional text cleaning or normalization was applied.

4. **Model Training:** The following models were trained on the augmented dataset:

   - BERT
   - RoBERTa
   - BETO
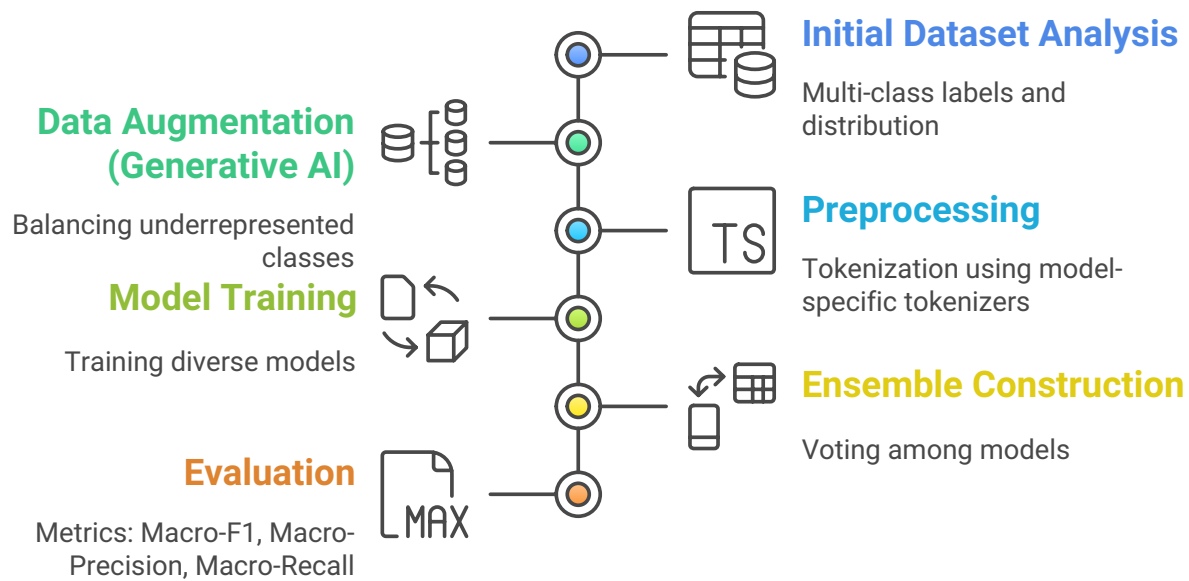   - LLaMA

# Enhancing Text Classification with AI



**Figure 2:** Training process for task 2. *Figure constructed using napkin* (https://app.napkin.ai/).

- Random Forest (RF)

We used a 70% training, 20% validation, and 10% testing split. Each model was trained for 3 to 5 epochs depending on convergence.

5. **Ensemble Construction:** Two ensembles were created using the predictions from BERT, RoBERTa, and BETO. As in Task 1, majority voting was applied for the final prediction.

6. **Evaluation:** Performance was assessed using macro-averaged F1-score, precision, and recall, following the evaluation guidelines of the MiSonGyny 2025 shared task.

## 5. Experiments and Results

This section presents the evaluation results obtained for both tasks using the macro-averaged metrics required by the MiSonGyny 2025 shared task: F1-score, Precision, and Recall. We report test set performance for each model, highlighting the benefits of transformer-based ensembles compared to individual models and traditional baselines.

### 5.1. Task 1: Binary Classification of Misogynistic Speech

Table 1 summarizes the performance of each model on Task 1.
The best result was obtained by the **Ensemble Transformers 2** model (F1 = 0.828), which outperformed all individual models, including RoBERTa (0.812) and BERT (0.798). This confirms the benefit of combining transformer outputs through majority voting. Traditional models such as SVM showed substantially lower performance (F1 = 0.479), highlighting the advantage of using transformer-based architectures for this task.

### 5.2. Task 2: Fine-Grained Classification of Misogynistic Content

Table 2 shows the results for Task 2.

**Table 1**
Task 1 – Test Results (Binary Classification)

| Model | F1 Score | Precision | Recall |
|---|---|---|---|
| Ensemble Transformers 2 | **0.828** | 0.824 | 0.833 |
| Ensemble Transformers 1 | 0.814 | 0.819 | 0.810 |
| RoBERTa | 0.812 | 0.804 | 0.824 |
| BERT (v2) | 0.798 | 0.796 | 0.799 |
| LLaMA | 0.784 | 0.798 | 0.774 |
| BERT (v1) | 0.790 | 0.793 | 0.788 |
| SVM | 0.479 | 0.493 | 0.492 |

**Table 2**
Task 2 – Test Results (Multi-Class Classification)

| Model | F1 Score | Precision | Recall |
|---|---|---|---|
| LLaMA | **0.488** | 0.514 | 0.490 |
| Ensemble Transformers 2 | 0.434 | 0.429 | 0.455 |
| BERT | 0.415 | 0.406 | 0.431 |
| Ensemble Transformers 3 | 0.402 | 0.387 | 0.423 |
| RoBERTa | 0.376 | 0.348 | 0.408 |
| BETO | 0.317 | 0.351 | 0.344 |
| Random Forest | 0.261 | 0.264 | 0.265 |

For Task 2, the best result was achieved by the **LLaMA** model (F1 = 0.488), followed by the ensemble Transformer model (F1 = 0.434) and BERT (F1 = 0.415). Again, ensemble methods improved over most base models. However, the overall F1 scores are lower than in Task 1, which reflects the increased difficulty of the fine-grained classification problem. Random Forest and traditional models performed poorly on this task, with an F1 score of only 0.261.

## 5.3. Results on the Official Competition Test Sets

In addition to local test evaluation, the official MiSonGyny 2025 organizers provided two blind test sets for external evaluation: 527 unlabeled instances for Task 1 and 393 for Task 2. Our team submitted a total of seven prediction files for Task 1 and eight for Task 2, each corresponding to a different model or ensemble configuration. The organizers computed macro-averaged F1-scores for each submission.

### 5.3.1. Task 1 – Official Test Set Results

Table 3 summarizes the results for our submissions to Task 1.

**Table 3**
Task 1 – Official Evaluation Results (Blind Test Set, 527 samples)

| Submission # | Submission ID | F1-Score |
|---|---|---|
| 1 | 280264 | 0.8140 |
| 2 | 281225 | 0.7838 |
| 3 | 282378 | 0.4792 |
| 4 | 283395 | 0.8120 |
| 5 | 289199 | 0.7902 |
| 6 | 289232 | **0.8280** |
| 7 | 289321 | 0.7976 |

The best result was obtained in **Submission 6** (ID: 289232), with an F1-score of **0.8280**. This submission secured **3rd place overall** in Task 1 of the competition.

### 5.3.2. Task 2 – Official Test Set Results

Table 4 shows the macro F1-scores obtained from the official evaluation of our submissions for Task 2.

**Table 4**
Task 2 – Official Evaluation Results (Blind Test Set, 393 samples)

| Submission # | Submission ID | F1-Score |
|---|---|---|
| 1 | 280264 | 0.4150 |
| 2 | 281225 | **0.4883** |
| 3 | 282378 | 0.2606 |
| 4 | 283257 | 0.3172 |
| 5 | 283395 | 0.3755 |
| 6 | 289199 | 0.4150 |
| 7 | 289232 | 0.4343 |
| 8 | 289321 | 0.4025 |

The best result was achieved in **Submission 2** (ID: 281225), with an F1-score of **0.4883**, which placed our team in **6th position overall** for Task 2.

## 6. Conclusions

The detection of misogynistic speech in Spanish-language song lyrics is both a socially urgent and technically challenging task. Misogynistic and violent lyrics contribute to the normalization of harmful gender-based stereotypes and symbolic violence, particularly in musical genres with massive reach and influence among young audiences.

This study highlights the difficulty of developing robust classification systems in domains with limited annotated data. One of the main challenges faced in both tasks was the scarcity of publicly available labeled datasets, especially for fine-grained categories such as hate or violence. Moreover, the original datasets provided for the MiSonGyny 2025 shared task were significantly imbalanced. To address this issue, we adopted a dual strategy: leveraging external Spanish corpora with related annotations for Task 1, and generating new examples using generative AI (GROK) for Task 2.

Our experiments confirm the effectiveness of transformer-based models, especially when used in ensemble configurations. In Task 1, a transformer ensemble reached an F1-score of 0.828 and secured third place in the competition. In Task 2, which required multi-class classification, the best-performing system achieved an F1-score of 0.488 and ranked sixth. These results validate the value of ensemble learning in low-resource and high-variance classification problems such as misogyny detection in lyrics.

Future work may focus on improving class-specific performance, leveraging multilingual models, or incorporating contextual metadata such as genre or artist information to enhance classification robustness.

## Declaration on Generative AI

Generative AI tools were used in this work in two specific contexts:

- **Visualization:** Figures 1 and 2 were created using *Napkin AI* (https://app.napkin.ai/), a document editing platform that transforms structured text into visual diagrams.
- **Data Augmentation:** For Task 2, we employed the generative platform *GROK* (https://grok.com) to synthesize additional text samples for the underrepresented classes (*Hate* and *Violence*). This was necessary to mitigate class imbalance and support model training.

The use of generative AI was limited to these tasks and did not involve the generation of article content, code, or evaluation results.

# References

[1] E. Aldana-Bobadilla, A. Molina-Villegas, Y. Montelongo-Padilla, I. Lopez-Arevalo, O. S. Sordia, A language model for misogyny detection in latin american spanish driven by multisource feature extraction and transformers, Applied Sciences 2021, Vol. 11, Page 10467 11 (2021) 10467. doi:10.3390/APP112110467.

[2] R. C. Viluñir, A. S. Navarrete, C. Vidal-Castro, C. Martínez-Araneda, Improving automatic detection of gender-based violence in spanish song lyrics using deep learning, data augmentation and undersampling, Lecture Notes in Networks and Systems 1284 LNNS (2025) 189–209. URL: https://link.springer.com/chapter/10.1007/978-3-031-85363-0_12. doi:10.1007/978-3-031-85363-0_12.

[3] 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).

[4] 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.

[5] L. Casanovas-Buliart, P. Alvarez-Cueva, C. Castillo, Evolution over 62 years: an analysis of sexism in the lyrics of the most-listened-to songs in spain, Cogent Arts and Humanities 11 (2024) 2436723. URL: https://www.tandfonline.com/doi/pdf/10.1080/23311983.2024.2436723. doi:10.1080/23311983.2024.2436723;JOURNAL:JOURNAL:OAAH20;WGROUP:STRING:PUBLICATION.

[6] R. Calderon-Suarez, R. M. Ortega-Mendoza, M. Montes-Y-Gomez, C. Toxqui-Quitl, M. A. Marquez-Vera, Enhancing the detection of misogynistic content in social media by transferring knowledge from song phrases, IEEE Access 11 (2023) 13179–13190. doi:10.1109/ACCESS.2023.3242965.

[7] E. Hashmi, M. M. Yamin, S. Imran, S. Y. Yayilgan, M. Ullah, Enhancing misogyny detection in bilingual texts using fasttext and explainable ai, Proceedings - 2024 International Conference on Engineering and Computing, ICECT 2024 (2024). doi:10.1109/ICECT61618.2024.10581058.

[8] E. Hashmi, S. Y. Yayilgan, M. M. Yamin, M. Ullah, Enhancing misogyny detection in bilingual texts using explainable ai and multilingual fine-tuned transformers, Complex and Intelligent Systems 11 (2025) 1–19. URL: https://link.springer.com/article/10.1007/s40747-024-01655-1. doi:10.1007/S40747-024-01655-1/FIGURES/14.

[9] A. Jiang, Tackling Sexist Hate Speech: Cross-Lingual Detection and Classification on Social Media, Phd thesis, Queen Mary University of London, 2025. URL: https://qmro.qmul.ac.uk/xmlui/handle/123456789/98199.

[10] N. Jindal, P. K. Kumaresan, R. Ponnusamy, S. Thavareesan, S. Rajiakodi, B. R. Chakravarthi, Mistra: Misogyny detection through text–image fusion and representation analysis, Natural Language Processing Journal 7 (2024) 100073. URL: https://www.sciencedirect.com/science/article/pii/S2949719124000219. doi:10.1016/J.NLP.2024.100073.

[11] J. A. García-Díaz, S. M. Jiménez-Zafra, M. A. García-Cumbreras, R. Valencia-García, Evaluating feature combination strategies for hate-speech detection in spanish using linguistic features and transformers, Complex and Intelligent Systems 9 (2023) 2893–2914. URL: https://link.springer.com/article/10.1007/s40747-022-00693-x. doi:10.1007/S40747-022-00693-X/TABLES/13.

[12] F. B. M. P. del Arco SUPERVISED, M. L. T. M.-V. A. U.-L. JAÉN, DETECTING OFFENSIVE LANGUAGE BY INTEGRATING MULTIPLE LINGUISTIC PHENOMENA, volume 52, Jaén : Universidad de Jaén, 2023. URL: https://hdl.handle.net/10953/2400.

[13] Codabench, MiSonGyny 2025: Misogyny in Song Lyrics, https://www.codabench.org/competitions/5914/, 2025. Accessed: 2025-05-22.

[14] G. Naidu, T. Zuva, E. M. Sibanda, A review of evaluation metrics in machine learning algorithms, Lecture Notes in Networks and Systems 724 LNNS (2023) 15–25. URL: https://link.springer.com/chapter/10.1007/978-3-031-35314-7_2. doi:10.1007/978-3-031-35314-7_2.

[15] 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 (NAACL), 2019.

[16] S. Chopra, P. Agarwal, J. Ahmed, S. S. Biswas, Ahmed, J. Obaid, Roberta and bert: Revolutionizing mental healthcare through natural language, SN Computer Science 2024 5:7 5 (2024) 1–12. URL: https://link.springer.com/article/10.1007/s42979-024-03202-8. doi:10.1007/S42979-024-03202-8.

[17] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[18] J. Cañete, G. Chaperon, R. Fuentes, J. Pérez, D. Parra, Spanish pre-trained bert model and evaluation data, arXiv preprint arXiv:2002.02340 (2020).

[19] I. Almubark, Exploring the impact of large language models on disease diagnosis, IEEE Access (2025). doi:10.1109/ACCESS.2025.3527025.

[20] B. E. Boser, I. M. Guyon, V. N. Vapnik, A training algorithm for optimal margin classifiers, in: Proceedings of the fifth annual workshop on Computational learning theory, ACM, 1992, pp. 144–152.

[21] M. E. Hassan, M. Hussain, I. Maab, U. Habib, M. A. Khan, A. Masood, Detection of sarcasm in urdu tweets using deep learning and transformer based hybrid approaches, IEEE Access 12 (2024) 61542 − 61555. doi:10.1109/ACCESS.2024.3393856.

[22] L. Breiman, Random forests, Machine learning 45 (2001) 5–32.

[23] W. Li, M. Zhang, A comparative study of classical and transformer-based models for social bias detection, Journal of Computational Linguistics and Applications (2024).

[24] M. H. Shohan, K. R. Ahmed, N. F. Kahar, N. Jahan, M. M. Hassan, R. B. Ahmad, N. Yaakob, O. B. Lynn, N. Islam, Use of natural language processing for the detection of hate speech on social media, Journal of Advanced Research in Applied Sciences and Engineering Technology 51 (2025) 86 − 96. doi:10.37934/araset.51.2.8696.

[25] L. Casanovas-Buliart, C. Castillo, P. Alvarez-Cueva, Sexism in the lyrics of the most listened to songs in spain, 2023. URL: https://doi.org/10.5281/zenodo.8134122. doi:10.5281/zenodo.8134122.

[26] R. Calbullanca Viluñir, A. Segura Navarrete, C. Vidal-Castro, C. Martínez-Araneda, Corpus of song lyrics in spanish labeled for gender-based violence against women, 2024. URL: https://doi.org/10.5281/zenodo.13370289. doi:10.5281/zenodo.13370289.