Avahi at MiSonGyny 2025: A BETO-Based Embedding Approach with KNN for Spanish Music Misogyny Detection

Diana Jimenez[1,†], Marco Cardoso-Moreno[1,†] and Luis Moreno-Mendieta[1,†]

[1] *Avahi,1390 Market Street, Suite 200 San Francisco, CA 94102 United States, https://www.avahitech.com*

Abstract

This paper addresses the detection of misogynistic content in Spanish song lyrics through a novel hybrid approach that combines traditional Natural Language Processing techniques with transformer-based embeddings. We propose a K-Nearest Neighbors classification system inspired by Retrieval-Augmented Generation (RAG) architectures, utilizing BETO embeddings stored in a vector database for similarity-based classification. Our methodology aims to identify subtle forms of gender-based discrimination in musical content, including micro-machismo expressions that often go unnoticed in everyday consumption. The proposed approach achieved competitive results while maintaining interpretability and avoiding direct reliance on Large Language Model outputs for classification decisions.

Keywords

natural language processing, transformer architectures, vector databases, BETO, Spanish NLP, retrieval augmented classification, embedding models

1. Introduction

Misogyny remains a pervasive issue in contemporary society, manifesting across various forms of media and cultural expressions. We can observe, hear, and experience numerous instances of this type of discrimination in our daily lives. [1] Within this context, there exists a concept known as "micro-machismo", which encompasses seemingly minor acts that are nevertheless charged with misogynistic undertones. Society has become so accustomed to these expressions that we often fail to recognize when such discrimination is occurring around us.

Music, as one of the most influential cultural mediums, serves as both a reflection of societal attitudes and a vehicle for perpetuating certain ideologies. We frequently listen to music without consciously processing the lyrical content, potentially normalizing misogynistic messages embedded within songs. This unconscious consumption of discriminatory content contributes to the broader societal acceptance of gender-based violence and inequality. [2].

It is crucial for our society to cease normalizing this form of violence, and the first step toward this goal is simply recognizing its presence. The identification of misogynistic content in popular music represents an important step in raising awareness and promoting more conscious media consumption. In this paper, we approach this problem through the lens of Natural Language Processing, specifically addressing the Task 1 proposed by the MiSonGyny 2025, for Spanish lyrics mysogyny detection. [3], part of the IberLEF tasks [4].

Our methodology combines traditional preprocessing techniques with the contextual understanding capabilities of Large Language Models (LLMs), while avoiding direct dependence on generative AI outputs for classification. This approach ensures both the reliability of our results and the interpretability

IberLEF 2025, September 2025, Zaragoza, Spain

[*] Corresponding author.

[†] These authors contributed equally.

diana.lopez@avahitech.com (D. Jimenez); marco.cardoso@avahitech.com (M. Cardoso-Moreno); enrique.moreno@avahitech.comLuisMoreno-Mendieta (L. Moreno-Mendieta)
https://www.linkedin.com/in/diana-jimenez-aab038201 (D. Jimenez); https://cardoso1994.github.io/ (M. Cardoso-Moreno)
000000023326557X (D. Jimenez); 0009-0001-1072-2985 (M. Cardoso-Moreno); 0009-0007-7198-6849 (L. Moreno-Mendieta)

of our system, making it suitable for practical applications in content moderation and social awareness initiatives.

2. Literature Review

The detection of misogynistic content in text has garnered significant attention in recent years, particularly in the context of social media analysis and hate speech detection. As highlighted by [5], the challenge of identifying various types of hate speech, including misogyny, in social media is particularly complicated due to the context-dependent nature of offensive language and the diverse factors that influence its interpretation.

Previous research has explored various approaches ranging from traditional machine learning techniques to sophisticated deep learning models. The literature reveals two principal approaches for misogyny detection in social media platforms. The first approach is based on classical machine learning models, which have demonstrated competitive performance across multiple studies. For instance, [6] presented a model based on Support Vector Machines with K-fold cross-validation that achieved substantial results in misogyny identification tasks on Twitter data.

Ensemble approaches have also shown promising results in this domain. [7] created an ensemble combining Logistic Regression, Naive Bayes, and Support Vector Machines, achieving notable performance on offensive language detection tasks including misogyny, with accuracy scores reaching 0.81 for female-targeted content when using Logistic Regression classifiers.

The second principal approach involves neural network architectures, which have demonstrated competitive results in misogyny detection tasks. [8] employed Convolutional Neural Networks enhanced with GRU layers, taking advantage of GRU's simpler gate structure compared to LSTM networks, which allows for better training and generalization on smaller datasets. Their approach achieved high performance in classifying Twitter data across multiple categories including sexist posts.

More sophisticated neural architectures have been explored as well. [9] presented HybridCNN, which combined character-level and word-level convolutional networks, achieving an F1-score of 0.827 in the classification of sexist, racist, and neutral tweets. Similarly, [10] demonstrated the effectiveness of Long Short-Term Memory Networks, achieving an impressive F1-score of 0.930 on similar classification tasks.

Bidirectional approaches have also proven effective in this domain. [11] utilized Bidirectional LSTM networks for misogyny language identification, leveraging the advantage of processing contextual information from both directions to create more complete word representations compared to traditional unidirectional approaches.

The field has been significantly advanced through shared tasks and evaluation campaigns. The Automatic Misogyny Identification (AMI) shared tasks, held at IberEval 2018 and Evalita 2018, provided standardized evaluation frameworks for comparing different approaches. These competitions focused on two main subtasks: binary misogyny identification and misogynistic behavior classification with target identification.

Results from these shared tasks reveal interesting patterns in model performance. In the IberEval 2018 competition, the best results for English misogyny detection were achieved by [12] using SVM models with different kernels (RBF for English, linear for Spanish), reaching 0.913 accuracy. The competition also highlighted the importance of lexical features, including swear word counts, sexist slur presence, and hashtag analysis.

For more complex classification tasks involving misogynistic behavior categorization, ensemble approaches showed superior performance. [13] achieved the best results in Subtask B with 0.44 average F-Measure using SVM with linear kernel combined with ensemble models that integrated SVM, Random Forest, and Gradient Boosting classifiers.

Interestingly, the evaluation campaigns revealed that classical machine learning approaches, particularly ensemble methods, often outperformed neural network-based models on the available datasets. As noted by the survey authors, this observation might be attributed to the relatively small size of the training datasets, since neural networks typically require larger amounts of data to achieve their full potential.

The Spanish language context presents unique challenges and opportunities for misogyny detection. [14] demonstrated language-specific considerations by achieving top results for Spanish datasets while showing moderate performance on English data, highlighting the importance of language-specific feature engineering and the cultural context embedded in misogynistic expressions.

Recent work in Spanish misogyny detection has demonstrated the effectiveness of combining linguistic features with word embeddings. [15] achieved 85.175% accuracy by combining classification based on average word embeddings with linguistic features, highlighting the importance of understanding which linguistic phenomena contribute to misogyny identification. Their work revealed that offensive language, grammatical gender, and grammatical errors with misspellings serve as discerning linguistic features, while cultural and dialectal differences across Spanish-speaking regions complicate the identification process.

Multi-task learning approaches have shown particular promise in misogyny and aggression detection. [16] presented a BERT-based multi-task architecture that simultaneously addresses aggression identification and misogynistic aggression identification. Their approach leveraged the relationship between these tasks, demonstrating that aggression and misogyny detection are inherently related problems. The system achieved competitive results across multiple languages (English, Hindi, and Bengali), with their best performance reaching 0.8579 weighted F1-measure on English misogyny detection, securing 3rd place out of 15 teams in the TRAC-2 shared task.

The multi-task framework proposed by [16] utilized attention mechanisms over BERT representations, followed by fully-connected layers and separate classification heads for each subtask. Their analysis revealed that Covertly Aggressive (CAG) content represents the most challenging category to detect, often being misclassified as Non-Aggressive due to its indirect and sarcastic nature. This finding highlights the importance of capturing subtle linguistic patterns in discriminatory language detection.

Recent developments in Spanish-language hate speech detection have expanded beyond general misogyny to address specific vulnerable populations. [17] presented the HOMO-MEX shared task focusing on hate speech detection towards the Mexican Spanish-speaking LGBT+ population at IberLEF 2024. This work represents an important evolution in the field, addressing the need for culturally and linguistically specific approaches to hate speech detection, particularly considering the unique challenges posed by Mexican Spanish variants and LGBT+-targeted discrimination.

3. Proposal

We aim to avoid direct reliance on LLM-generated outputs for classification decisions. Instead, our proposal combines traditional Natural Language Processing techniques with the powerful attention mechanisms inherent in transformer architectures. Our approach implements a K-Nearest Neighbors classification system inspired by RAG architectures, utilizing embeddings created through transformer-based models to perform vector similarity searches within a specialized vector database.

This methodology offers several advantages over direct LLM classification approaches. First, it maintains transparency in the decision-making process, allowing for better interpretability of results. Second, it reduces the computational overhead associated with running inference on large language models for each classification task. Finally, it provides a more stable and predictable classification framework that is less susceptible to the variability often observed in generative model outputs.

3.1. Preprocessing

The dataset consisted of lyrics from Spanish songs, which presented several data quality challenges typical of web-scraped content. The raw data contained numerous artifacts commonly found in web pages, including HTML tags, special characters, and content fragments native to web page structures rather than actual lyrical content. Additionally, we observed repeated verses within songs, which could potentially bias our classification system.

To address these challenges, we developed a comprehensive preprocessing pipeline designed to clean and standardize the textual data. The pipeline encompasses several stages: HTML tag removal, special character normalization, elimination of n-grams without semantic content that are characteristic of web page artifacts, and deduplication of repeated verses within individual songs.

The preprocessing stage proved crucial for ensuring the quality of our embeddings and, consequently, the accuracy of our classification system. By removing non-lyrical content and normalizing the text representation, we created a cleaner dataset that better represents the actual musical content intended for analysis.

3.2. Models for Embeddings

For vectorial representation of the textual content, we employed BETO (Spanish BERT), a transformer-based model specifically trained on Spanish corpora. BETO provides contextualized embeddings that capture both semantic meaning and linguistic nuances specific to the Spanish language, making it particularly well-suited for analyzing Spanish song lyrics.

The choice of BETO over multilingual models was motivated by the need to capture culture-specific expressions and linguistic patterns that may be particularly relevant to the identification of misogynistic content in Spanish-speaking contexts. These embeddings serve as the foundation for our similarity-based classification approach, encoding each lyrical segment into a dense vector representation that preserves semantic relationships.

The embedding generation process involves tokenizing the preprocessed lyrics and obtaining contextualized representations from BETO's final hidden layers. These embeddings are then stored in our vector database, creating a searchable repository of lyrical content representations.

3.3. Classification

Our classification approach operates on two levels. As a baseline, we implemented a fine-tuning approach using BETO for direct classification. This involved adapting the pre-trained BETO model for the binary classification task of detecting misogynistic content, utilizing standard fine-tuning procedures with appropriate hyperparameters optimized for our specific dataset.

However, our main approach leverages Amazon Web Services' OpenSearch as a vector database solution. We store the BETO-generated embeddings (BETO was previously fine-tuned) in this database, enabling efficient similarity-based queries. For classification, we employ OpenSearch's native K-Nearest Neighbors implementation to identify the most similar lyrical content to a given query.

The classification decision is made based on the labels of the nearest neighbors, implementing a voting mechanism that considers both the similarity scores and the class distribution among the retrieved neighbors. This approach allows us to leverage the semantic relationships captured by the BETO embeddings while maintaining interpretability through the similarity-based decision process. The hyperparameters for our KNN approach include the number of nearest neighbors to consider, the similarity metric used for distance calculation, and the weighting scheme for neighbor votes.

4. Results

Approach	F1
BETO	0.7865
BETO + OpenSearch	0.8051

Table 1
Performance comparison of different approaches for misogyny detection in Spanish song lyrics.

5. Discussion

Our experimental results demonstrate that combining BETO embeddings with vector database queries yields superior performance compared to traditional classification approaches. The hybrid methodology successfully leverages the contextual understanding capabilities of transformer models while maintaining the interpretability and efficiency of similarity-based classification.

The vector database approach offers several practical advantages for real-world deployment. The system can efficiently handle large-scale lyrical databases while providing transparent decision-making processes. This transparency is particularly important in applications involving content moderation, where understanding the reasoning behind classification decisions is crucial for both system operators and content creators. Furthermore, the approach demonstrates robustness in detecting subtle forms of misogyny, including "micro-machismo" expressions that traditional keyword-based systems might miss. The contextual embeddings capture semantic relationships that enable identification of discriminatory content even when expressed through indirect or culturally-specific language patterns.

6. Conclusions

This work presents a novel approach to detecting misogynistic content in Spanish song lyrics, combining the strengths of transformer-based language models with the efficiency and interpretability of vector similarity search. Our methodology successfully addresses the challenge of identifying both explicit and subtle forms of gender-based discrimination in musical content.

The results demonstrate the effectiveness of hybrid approaches that leverage modern NLP techniques while maintaining practical considerations such as interpretability and computational efficiency. This work contributes to the broader effort of raising awareness about discriminatory content in popular media and provides a foundation for developing tools that can assist in creating more conscious media consumption practices.

Future work could explore the extension of this approach to other languages and cultural contexts, as well as the development of more sophisticated similarity metrics that better capture the nuances of discriminatory language. Additionally, investigating the temporal evolution of misogynistic expressions in music could provide valuable insights into changing societal attitudes toward gender equality.

7. Declaration on Generative AI

The author(s) used ChatGPT and Grammarly in order to: grammar and spelling check, paraphrase and reword. After using these tools/services, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

Acknowledgments

The authors gratefully acknowledge the financial support provided by Avahi Inc, which made this research possible.

References

[1] K. Manne, Down girl: The logic of misogyny, Oxford University Press, 2017.

[2] E. G. Armstrong, Sexism and misogyny in music land, Journal of Criminal Justice and Popular Culture 8 (2001) 96–126.

[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] E. Shushkevich, J. Cardiff, Automatic misogyny detection in social media: A survey, [Journal Name] ([Year]).

[6] V. Nina-Alcocer, Ami at ibereval2018 automatic misogyny identification in spanish and english tweets, in: CEUR Workshop Proceedings, 2018, pp. 274–279.

[7] H. Saleem, K. Dillon, S. Benesch, D. Ruths, A web of hate: Tackling hateful speech in online social spaces, CoRR abs/1709.10159 (2017).

[8] Z. Ziki, L. Lei, Hate speech detection: A solved problem? the challenging case of long tail on twitter, arXiv preprint arXiv.1803.03662 10 (2018) 925–945. doi:10.3233/SW-180338.

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

[10] P. Badjatiya, S. Gupta, M. Gupta, V. Varma, Deep learning for hate speech detection in tweets, in: Proceedings of the 26th International Conference on World Wide Web Companion, 2017, pp. 759–760. doi:10.1145/3041021.3054223.

[11] M. Anzovino, E. Fersini, P. Rosso, Automatic identification and classification of misogynistic language on twitter, in: Proc. 23rd Int. Conf. on Applications of Natural Language to Information Systems, NLDB-2018, LNCS 10859, 2018, pp. 57–64. doi:10.1007/978-3-319-91947-8_6.

[12] E. Pamungkas, A. Cignarella, V. Basile, V. Patti, 14-exlab@unito for ami at ibereval2018: Exploiting lexical knowledge for detecting misogyny in english and spanish tweets, in: CEUR Workshop Proceedings, 2018, pp. 234–241.

[13] S. Frenda, B. Ghanem, M. Montes-y Gomez, Exploration of misogyny in spanish and english tweets, in: CEUR Workshop Proceedings, volume 2150, 2018, pp. 260–267.

[14] J. Canós, Misogyny identification through svm at ibereval 2018, in: CEUR Workshop Proceedings, 2018, pp. 229–233.

[15] J. A. García-Díaz, M. Cánovas-García, R. Colomo-Palacios, R. Valencia-García, Detecting misogyny in spanish tweets. an approach based on linguistics features and word embeddings, Future Generation Computer Systems (2020). doi:10.1016/j.future.2020.08.032.

[16] N. S. Samghabadi, P. Patwa, S. PYKL, P. Mukherjee, A. Das, T. Solorio, Aggression and misogyny detection using bert: A multi-task approach, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), 2020, pp. 126–131.

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