# VerbaNex AI at CLEF EXIST 2024: Detection of Online Sexism using Transformer Models and Profiling Techniques*

Notebook for the VerbaNex AI Lab at CLEF 2024

Elizabeth Martinez[1,†], Juan Cuadrado[1,*,†], Juan Carlos Martinez Santos[1,†] and Edwin Puertas[1,†]

[1]Universidad Tecnológica de Bolívar, School of Engineering, Cartagena de Indias 130010, Colombia.

**Abstract**

The integration of social networks into modern life has revolutionized global communication, allowing instantaneous interaction. However, this convenience has also been misused, leading to the proliferation of inappropriate and often sexist remarks on social media. To address this, the field of natural language processing has been developing techniques to identify and mitigate such content. Our research, conducted as part of the CLEF EXIST 2024 competition, introduces a novel approach. We combined features from the 'twitter-roberta-base-sentiment-latest' transformer model with traditional lexical elements and profiling. The profiling involved grouping profiles by gender, age, and education level. Then, we categorized them based on their positive response rate to sexism and trained classifiers accordingly. This method was evaluated using the testing profiles, achieving an F1 score of 0.745. In the evaluation phase, our approach yielded an F1 score of 0.63. The effective combination of linguistic, transformer-based features and profiling was crucial to achieving these results.

**Keywords**

Online Sexism Detection, Profiling Techniques, Natural Language Processing, Social Media Analysis, Binary Classification, Transformer Models

## 1. Introduction

In the modern era, social media has become an integral part of daily life, captivating nearly 80% of individuals through its ubiquitous digital platforms [1]. Social media facilitates communication among citizens, corporations, and governments, making its impact far-reaching and undeniable. However, this digital space has also seen a troubling rise in hate speech, particularly manifesting as gender-based inequities and injustices that disproportionately affect women [2, 3, 4]. The prevalence of sexist content exacerbates feelings of vulnerability and insecurity among women, both in online and offline environments [5, 6, 7].

Addressing this issue, our research is part of the CLEF EXIST 2024 competition [8, 9], focusing on Task 1, which involves identifying sexist expressions and behaviors in tweets and memes. Task 1 goal intent to develop effective techniques to detect and mitigate online sexism. In our approach, we combine features from the 'twitter-roberta-base-sentiment-latest' transformer model with traditional lexical elements and profiling techniques. Profiling involves grouping users based on gender, age, and education level, and further categorizing them according to their positive response rates to sexism. We then train classifiers based on these groups to enhance detection accuracy.

Our method includes rigorous pre-processing, the integration of lexical and transformer-based feature extraction, and the application of profiling techniques. This comprehensive strategy is evaluated using the testing profiles, achieving an F1 score of 0.745. During the evaluation phase, our approach yielded

*CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France*

*Corresponding author.

†These authors contributed equally.

eayala@utb.edu.co (E. Martinez); jflechas@utb.edu.co (J. Cuadrado); jcmartinezs@utb.edu.co (J. C. M. Santos); epuerta@utb.edu.co (E. Puertas)

0000-0001-6592-347X (E. Martinez); 0000-0002-8226-1372 (J. Cuadrado); 0000-0003-2755-0718 (J. C. M. Santos); 0000-0002-0758-1851 (E. Puertas)

an F1 score of 0.63. The integration of these diverse features and profiling techniques demonstrates the potential to significantly improve the detection of online sexism.

The following sections will detail our methodology, including the pre-processing steps, feature extraction techniques, regularization methods, and the evaluation metrics used to assess our system's performance. Through this research, we aim to contribute to the broader efforts of combating online sexism, providing insights and tools that can be used to create a safer and more equitable digital environment.

## 2. Related Work

Online sexism poses a significant problem, impacting women profoundly and creating a sense of insecurity both online and offline [10, 11, 12]. Addressing this issue necessitates the development of robust strategies to foster safer online environments while maintaining freedom of speech. In response to this need, several competitions and initiatives have emerged, focusing on the detection and mitigation of hate speech and sexism on social media platforms.

Competitions such as EVALITA and IberEval have been pivotal in this effort, leveraging diverse datasets from various social media platforms, including Twitter, Reddit, and Gab [13]. These datasets are crucial for developing and evaluating models aimed at detecting online hate speech and sexism. For instance, datasets developed by Wasem and Hovy contain annotations for both sexist and racist content and have served as foundational resources for numerous studies.

Research efforts utilizing these datasets have explored various methodologies. For example, [14] employed word vectors and contextual analysis to detect sexism and racism, using five Long Short-Term Memory (LSTM) networks as classifiers, achieving a precision of 0.9334. Similarly, [15] combined LSTM networks with random embeddings to extract features for Gradient Boosting Decision Trees (GBDT), achieving a precision of 0.930.

The Student Research Workshop (SRW) dataset, highlighted in [16], focuses specifically on sexist hate speech. In this study, a combination of bag-of-words and sequential word features was used with a Support Vector Machine (SVM) classifier [17], resulting in an accuracy of 0.8932. Other studies have experimented with techniques such as sentence embeddings, term frequency-inverse document frequency (TF-IDF) [18], and bag-of-words (BoW) methods, though these approaches generally achieved lower accuracy, with a maximum of 0.704 [19].

The continuous development and release of these datasets through various competitions have facilitated significant advancements in the field. Recently, the integration of transformer models from libraries like Transformers has shown considerable promise in enhancing the detection of online sexism. These sophisticated models represent a critical advancement in improving the accuracy and efficiency of detection systems [20, 21].

## 3. Data

For the CLEF EXIST 2024 competition, we utilized the dataset provided by the organizers, focusing specifically on the identification of sexism in tweets for Task 1. This dataset builds on the EXIST 2023 dataset, incorporating both English and Spanish tweets. The dataset includes a curated lexicon of 250 terms indicative of sexist content. These terms were used to gather a comprehensive collection of over 10,000 annotated tweets, with a balanced representation of English and Spanish content. To achieve a balanced dataset, excessively imbalanced terms were discarded, resulting in approximately 5,000 tweets labeled as sexist and 5,000 tweets labeled as non-sexist, ensuring an even distribution for training and testing. Six annotators from the Prolific app, guided by experts in gender issues, labeled each tweet, considering gender and age to mitigate label bias. Additional demographic details such as education level, ethnicity, and country of residence were also included for the 2023 and 2024 datasets. A learning with disagreements approach was employed, providing all annotations per instance rather than aggregated labels, to capture a diversity of perspectives.

# 4. Architecture

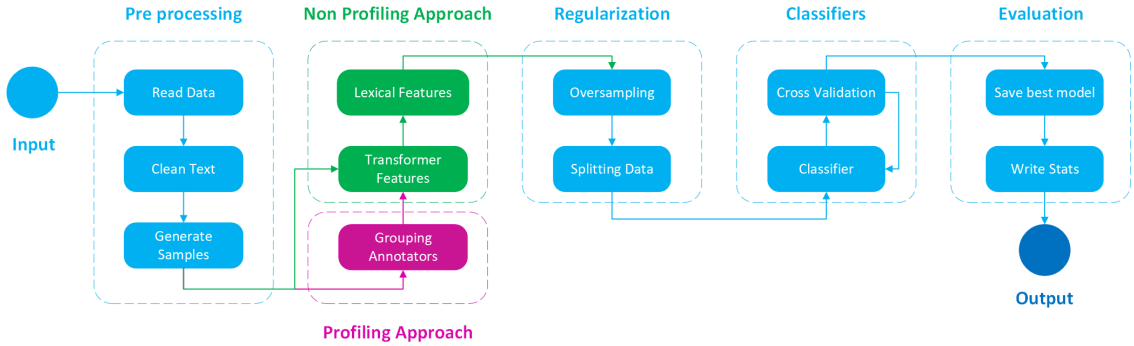


**Figure 1:** System Pipeline.

In this study, we developed a comprehensive system to detect sexism in online text, specifically tweets. Our architecture integrates multiple stages of data processing, feature extraction, and regularization to enhance the detection accuracy. The following subsections provide a detailed overview of each component within our system.

## 4.1. Pre-Processing

In the preprocessing stage, we aimed to standardize and clean the textual data to ensure consistency and clarity. Using the Natural Language Toolkit (NLTK) library [22], we performed a series of transformations on the text data. Hashtags were replaced with the term "hashtag," and user mentions were substituted with "mention." URLs within the text were replaced by the placeholder "URL," and emojis were converted to their corresponding UTF-8 encoded descriptions, labeled as "emoji." Following these substitutions, we further refined the text by removing punctuation, converting all characters to lowercase, and eliminating common stopwords to reduce noise and enhance the quality of the data for subsequent analysis.

## 4.2. Profiling

In the provided dataset each message was annotated by six different individuals. Due to the even number of annotators, we sometimes faced situations where three annotators labeled a message as sexist and the other three labeled it as non-sexist, resulting in a tie. To address this, we implemented a profiling approach based on demographic factors: gender, education level, and age.

We grouped the annotators' responses according to these demographic profiles. For each message, we calculated the total number of responses per profile and the number of times a message was labeled as sexist. This allowed us to categorize the profiles into four groups based on their likelihood of labeling messages as sexist or non-sexist.

We analyzed these profiles to predict the probability of a message being labeled as sexist based on the annotators' demographic tendencies. This profiling approach helped us resolve ties and make more informed decisions regarding the classification of messages.

Following the profiling, we performed feature extraction and trained four distinct systems based on the grouped profiles and their responses. These systems were then evaluated to determine their effectiveness in accurately detecting sexist messages.

## 4.3. Lexical Feature Extraction

To begin our analysis, we focused on extracting traditional lexical features to gain insights into the linguistic patterns present in the data. This process involved identifying various lexical elements as described by Puertas et al. [23]. We categorized these features into 27 distinct groups, including word

usage, hashtags, URLs, emojis, frequently used Part-of-Speech (POS) tags, adverbs, and adjectives. This comprehensive extraction allowed us to conduct a thorough examination of the corpus, providing a solid foundation for understanding the data's linguistic characteristics.

To enhance our approach, we integrated modern techniques by incorporating the Twitter-roBERTa-base model specifically fine-tuned for sentiment analysis[24, 21]. This variant of RoBERTa-base was trained on a large collection of tweets from January 2018 to December 2021 and evaluated using the TweetEval benchmark. By leveraging this model, we were able to extract sentiment-based features alongside traditional lexical features. The combination of these two feature sets—lexical and transformer-based—was achieved through concatenation, resulting in a robust and comprehensive feature representation.

### 4.4. Transformer Integration

To further enhance our feature extraction process, we incorporated the Twitter-roBERTa-base model, specifically fine-tuned for sentiment analysis. This transformer model, a variant of RoBERTa-base, was trained on a large dataset of tweets collected from January 2018 to December 2021 and evaluated using the TweetEval benchmark. The primary advantage of using this model lies in its ability to capture nuanced sentiment features from the text, which are crucial for identifying subtle expressions of sexism.

The integration process involved several key steps. First, we preprocessed the text data as described in the Pre-Processing section, ensuring consistency and clarity. Next, we passed the cleaned text through the Twitter-roBERTa-base model to extract sentiment-based features. These features encapsulate the emotional tone and contextual sentiment of each tweet, providing a deeper understanding of the underlying sentiment patterns.

By combining these sentiment features with the traditional lexical features extracted earlier, we created a comprehensive feature set. This combined feature set was achieved through a concatenation process, where both sets of features were merged to form a unified representation. This approach allowed us to leverage the strengths of both traditional lexical analysis and modern transformer-based sentiment analysis.

The resulting feature representation was then used as input for our classification models. This hybrid approach not only improved the accuracy of sexism detection but also provided a richer and more nuanced understanding of the text data.

### 4.5. Regularization

Regularization was an essential step in our methodology to ensure that our model performed well and was not prone to overfitting. First, we divided the dataset into training and validation sets to facilitate model training and evaluation. To address the issue of class imbalance, we employed techniques to generate synthetic instances, ensuring that both classes were adequately represented in the training data. Specifically, we used the K-Fold Stratified Shuffle-Split technique, as described by Sandoval et al. [25], to create multiple splits of the data. This approach allowed us to maintain the original distribution of classes within each fold, enhancing the robustness and generalizability of our model through effective cross-validation.

## 5. Evaluation

The performance of our system was evaluated using two approaches: one without profiling and one with profiling. We assessed the system's effectiveness based on four key metrics: F1 score, precision, recall, and accuracy. The results for each approach during the training phase are presented in Table 1.

The results from the approach without profiling demonstrate a solid performance across all metrics. The higher precision and recall indicate that the system is capable of effectively identifying sexist messages, while maintaining a reasonable balance between false positives and false negatives.

**Table 1**
Training Phase Evaluation Metrics for Different Approaches

| Metric | Without Profiling | With Profiling |
|---|---|---|
| F1 Score | 75.68 | 74.5 |
| Precision | 75.85 | 74.71 |
| Recall | 75.71 | 74.55 |
| Accuracy | 75.71 | 74.55 |

The results from the approach with profiling show a balanced performance, with precision slightly higher than recall. This suggests that the profiling method contributed to a consistent detection of sexist messages while maintaining a moderate false positive rate.

Overall, the training phase evaluation results from both approaches indicate that our system is capable of reliably detecting sexism in online text. While the metrics are satisfactory, they highlight areas for further improvement and refinement.

### 5.1. Competition Evaluation

The final evaluation of our system was conducted during the CLEF EXIST 2024 competition, where the performance was assessed using the competition's official metrics. The results for both approaches are as follows:

**Table 2**
Competition Evaluation Metrics for Different Approaches

| Approach | Position | ICM-Hard | ICM-Hard Norm | F1_YES |
|---|---|---|---|---|
| Without Profiling | 54 | 0.1064 | 0.5532 | 0.6320 |
| With Profiling | 56 | 0.0390 | 0.5195 | 0.6221 |

The approach without profiling achieved a position of 54 in the competition, with an ICM-Hard score of 0.1064, an ICM-Hard Norm score of 0.5532, and an F1_YES score of 0.6320. This indicates a relatively strong performance in detecting sexist messages.

The approach with profiling, on the other hand, achieved a position of 56, with an ICM-Hard score of 0.0390, an ICM-Hard Norm score of 0.5195, and an F1_YES score of 0.6221. While this approach showed slightly lower performance metrics, it still demonstrated the system's capability in the competition context.

Overall, the competition evaluation results suggest that both approaches have their strengths, with the non-profiling approach slightly outperforming the profiling approach. These results provide valuable insights for further refinement and optimization of our detection system.

## 6. Conclusion

This study presented a novel approach to detecting online sexism in tweets by combining transformer models and profiling techniques. By integrating features from the 'twitter-roberta-base-sentiment-latest' model with traditional lexical elements, and grouping annotator profiles based on demographic factors such as gender, age, and education level, we aimed to improve the accuracy of sexism detection. Our methodology included rigorous pre-processing, comprehensive feature extraction, and robust regularization techniques to ensure the reliability of our model.

The evaluation results demonstrated that both approaches, with and without profiling, performed satisfactorily in the training phase, achieving F1 scores of 74.5 and 75.68, respectively. The competition results showed that the approach without profiling slightly outperformed the profiling approach, with

higher scores in all metrics. While the metrics indicate a solid performance, there is still room for improvement, especially in refining the profiling techniques to better capture demographic biases.

## 7. Future Work

Future work will focus on several areas to enhance the detection of online sexism. First, we plan to refine our profiling techniques by incorporating more detailed demographic data and exploring additional factors that may influence annotator biases. Second, we aim to experiment with other transformer models and fine-tune them specifically for sexism detection to improve the performance further.

Additionally, expanding the dataset to include a wider variety of social media platforms and languages could provide a more comprehensive understanding of online sexism. We also intend to investigate the use of advanced regularization techniques and ensemble methods to increase the robustness of our models.

Finally, collaborating with experts in gender studies and psychology could provide valuable insights into the nuances of sexist language, leading to more accurate and sensitive detection systems. By addressing these areas, we hope to contribute to the development of more effective tools for combating online sexism and promoting safer online environments.

## Acknowledgments

The authors would like to acknowledge the support provided by the master's degree scholarship program in engineering at the Universidad Tecnologica de Bolivar (UTB) in Cartagena, Colombia.

## References

[1] F. Awan, D. Gauntlett, Young people's uses and understandings of online social networks in their everyday lives, Young 21 (2013) 111–132.

[2] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, Automatic classification of sexism in social networks: An empirical study on twitter data, IEEE Access 8 (2020) 219563–219576. doi:10.1109/ACCESS.2020.3042604.

[3] R. Brown, Social identity theory: Past achievements, current problems and future challenges, European journal of social psychology 30 (2000) 745–778.

[4] M. Menegatti, M. Rubini, Gender bias and sexism in language, Oxford Research Encyclopedia of Communication (2018) 451–468. URL: https://cris.unibo.it/handle/11585/623058. doi:10.1093/ACREFORE/9780190228613.013.470.

[5] P. Parikh, H. Abburi, N. Chhaya, M. Gupta, V. Varma, Categorizing sexism and misogyny through neural approaches, ACM Transactions on the Web (TWEB) 15 (2021) 1–31.

[6] 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.

[7] M. KhosraviNik, E. Esposito, Online hate, digital discourse and critique: Exploring digitally-mediated discursive practices of gender-based hostility, Lodz Papers in Pragmatics 14 (2018) 45–68.

[8] L. Plaza, J. Carrillo-de-Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of EXIST 2024 – Learning with Disagreement for Sexism Identification and Characterization in Social Networks and Memes, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), 2024.

[9] L. Plaza, J. Carrillo-de-Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of EXIST 2024 – Learning with Disagreement for Sexism Identification and Characterization in Social Networks and Memes (Extended Overview), in: G. Faggioli,

N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.
[10] A. Jiang, X. Yang, Y. Liu, A. Zubiaga, Swsr: A chinese dataset and lexicon for online sexism detection, Online Social Networks and Media 27 (2022) 100182.
[11] S. Jain, The rising fourth wave: feminist activism on digital platforms in india, ORF Issue Brief 384 (2020) 1–16.
[12] K. Barker, O. Jurasz, Online misogyny, Journal of International Affairs 72 (2019) 95–114.
[13] E. Fersini, D. Nozza, P. Rosso, et al., Overview of the evalita 2018 task on automatic misogyny identification (ami), in: EVALITA Evaluation of NLP and Speech Tools for Italian Proceedings of the Final Workshop 12-13 December 2018, Naples, Accademia University Press, 2018.
[14] G. K. Pitsilis, H. Ramampiaro, H. Langseth, Effective hate-speech detection in twitter data using recurrent neural networks, Applied Intelligence 48 (2018) 4730–4742.
[15] 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.
[16] J. Andreas, E. Choi, A. Lazaridou, Proceedings of the naacl student research workshop, in: Proceedings of the NAACL Student Research Workshop, 2016.
[17] K. Soumya George, S. Joseph, Text classification by augmenting bag of words (bow) representation with co-occurrence feature, IOSR Journal of Computer Engineering 16 (2014) 34–38.
[18] J. Ullman, Mining of massive datasets, Cambridge University Press, 2011.
[19] P. Saha, B. Mathew, P. Goyal, A. Mukherjee, Hateminers: Detecting hate speech against women, arXiv preprint arXiv:1812.06700 (2018).
[20] F. Barbieri, J. Camacho-Collados, L. Neves, L. T. Espinosa-Anke, Unified benchmark and comparative evaluation for tweet classification. arxiv 2020, arXiv preprint arXiv:2010.12421 (2020).
[21] D. Loureiro, F. Barbieri, L. Neves, L. Espinosa Anke, J. Camacho-collados, TimeLMs: Diachronic language models from Twitter, in: V. Basile, Z. Kozareva, S. Stajner (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 251–260. URL: https://aclanthology.org/2022.acl-demo.25. doi:10.18653/v1/2022.acl-demo.25.
[22] S. Bird, Nltk: the natural language toolkit, in: Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, 2006, pp. 69–72.
[23] E. Puertas, L. G. Moreno-Sandoval, F. M. Plaza-del Arco, J. A. Alvarado-Valencia, A. Pomares-Quimbaya, L. Alfonso, Bots and gender profiling on twitter using sociolinguistic features, CLEF (Working Notes) (2019) 1–8.
[24] J. Camacho-collados, K. Rezaee, T. Riahi, A. Ushio, D. Loureiro, D. Antypas, J. Boisson, L. Espinosa Anke, F. Liu, E. Martínez Cámara, TweetNLP: Cutting-edge natural language processing for social media, in: W. Che, E. Shutova (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Abu Dhabi, UAE, 2022, pp. 38–49. URL: https://aclanthology.org/2022.emnlp-demos.5. doi:10.18653/v1/2022.emnlp-demos.5.
[25] L. Sandoval, E. Puertas, A. Quimbaya, J. Valencia, Assembly of polarity, emotion and user statistics for detection of fake profiles, in: CLEF, 2020.