NetGuardAI at EXIST2025: Sexism Detection using mDeBERTa

Notebook for the EXIST2025 Lab at CLEF 2025

Maria-Diana Cotelin[1], Ciprian-Octavian Truică[1,2] and Elena-Simona Apostol[1,2,*]

[1] *National University of Science and Technology Politehnica University Bucharest, Splaiul Independenței 313, București 060042, Romania*

[2] *Academy of Romanian Scientists, Ilfov 3, Bucharest, 050044, Romania*

Abstract

In this paper, we present our system for the EXIST2025 shared task on sexist content detection in English and Spanish. We experimented with several transformer-based models, including DeBERTa, mDeBERTa, XLM-RoBERTa, Detoxify, and HateBERT, alongside three levels of text preprocessing: Light, Classic, and Aggressive Cleaning. Although various data augmentation strategies were tested, such as translation-based augmentation using Meta AI's NLLB model and pseudo-labeling with the EDOS dataset, the final submitted system did not include these enhancements. The system achieved competitive rankings in the soft evaluation setting, placing in the top 15 for several subtasks.

Keywords

Sexism Detection, Transformers, Text Classification, Learning with Disagreements

1. Introduction

With the proliferation of internet access, social media has woven itself into the fabric of everyday life. Platforms like X (formerly Twitter), TikTok, and Instagram provide users with the means to express themselves freely and connect with others across the globe. Yet, these same spaces have also become breeding grounds for harassment and abuse, especially targeting women [1, 2, 3, 4, 5, 6].

This article presents the working notes for our submission to EXIST 2025 [7, 8, 9], representing the effort of team NetGuardAI. EXIST (sEXism Identification in Social neTworks) is a series of scientific events and shared tasks focused on identifying sexism in social networks. Its goal is to promote the automatic detection of sexism in all its forms, ranging from overt misogyny to more subtle, implicit sexist behaviors.

For this edition, we tackled one of the three tasks, focusing on the Natural Language Processing (NLP) challenges.

- Subtask 1.1: Sexism Identification - binary classification
- Subtask 1.2: Source Intention - multi-class (4) classification
- Subtask 1.3: Sexism Categorization - multi-class (6) classification

Our system utilized multilingual DeBERTa (mDeBERTa) [10, 11] and XLM-RoBERTa [12] models fine-tuned for the detection of sexist content. Before using model inference, we applied pre-processing techniques such as lowercasing, punctuation normalization, and removal of special characters to clean the text data. To enhance the diversity and robustness of the training data, we augmented the existing dataset by translating it using NLLB (No Language Left Behind) [13]. The processed and augmented

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

[*] Corresponding author.

maria_diana.cotelin@stud.acs.upb.ro (M. Cotelin); ciprian.truica@upb.ro (C. Truică); elena.apostol@upb.ro (E. Apostol)
https://sites.google.com/view/ciprian-octavian-truica (C. Truică); https://sites.google.com/view/elena-simona-apostol (E. Apostol)
0009-0006-6414-9834 (M. Cotelin); 0000-0001-7292-4462 (C. Truică); 0000-0001-6397-4951 (E. Apostol)

inputs were then fed into the mDeBERTa and XLM-RoBERTa models, which were trained to classify and identify sexist language based on contextual understanding.

The rest of the paper is structured as follows. Section 2 presents the current studies for sexism identification. Section 3 describes the experimental setup. Section 4 presents and discusses the results. Finally, Section 5 concludes our work and hints at future directions.

2. Related Work

Harmful content detection has been tackled in recent years from three different perspectives. The most common approaches use word and transformer embeddings together with deep learning models to classify the textual content [14, 15, 16, 17, 18]. Another approach is to enhance the context using metadata such as social context [19] or the information diffusion within a network [20]. Moreover, novel approaches also propose the use of network immunization to stop the spread of harmful content online [21, 22, 23, 24, 25, 26]. Furthermore, full-fledged architectures that analyze social media in real-time have been developed [27, 28].

When analysing the EXIST 2024 competition, we find that a variety of learning strategies were employed for sexism detection across multiple modalities and languages. Traditional machine learning models, including Logistic Regression, Support Vector Machines, Random Forests, and ensemble methods like stacking and boosting, were used effectively, especially when combined with careful feature engineering and text pre-processing [29, 30]. However, transformer-based architectures dominated the competition.

Transformer models such as BERT, RoBERTa, and especially XLM-RoBERTa were widely used and consistently ranked among the top-performing models across tasks [29, 31, 32, 33, 18, 34, 35, 36, 37]. BERT variants (including BERT multilingual, BETO for Spanish, and DeBERTa-v3) were employed in several configurations, often fine-tuned on competition-specific data [29, 31, 32, 18, 34, 35, 36]. XLM-RoBERTa stood out as a strong baseline across multiple tasks, being used both as a standalone model and in ensembles [31, 32, 33, 18, 34, 35, 37].

Data augmentation techniques were used to enhance model generalization. These included synonym replacement, contextual augmentation with transformers, back-translation, and methods like AEDA [29, 34]. Some teams enriched data through few-shot learning using large language models such as GPT-3.5 and GPT-4, particularly effective in multimodal tasks involving memes [33, 38, 36].

Domain-specific adaptation was another notable trend. Models pretrained on sentiment, hate speech, or Twitter-specific data showed competitive results when fine-tuned for EXIST tasks [18, 34, 35]. Additionally, participant teams used demographic metadata (e.g., annotator gender, age, ethnicity) to enrich input representations and tailor predictions [34, 37].

3. Method

3.1. Task Description

The EXIST 2025 shared task is divided into three components: 1) tweet classification (Task 1), 2) meme classification (Task 2), and 3) video classification (Task 3). Our participation focused exclusively on tweet classification. Within this domain, Task 1.1 is a binary classification task that aims to detect whether a tweet expresses sexist intent. Task 1.2 involves multiclass classification, where tweets identified as sexist are further assigned to one of three predefined categories based on the perceived intent of the author. Lastly, Task 1.3 is a multi-label classification task, allowing sexist tweets to be associated with one or more relevant categories.

3.2. Dataset

The 2025 edition includes nine subtasks in English and Spanish, covering three classification tasks: sexism identification, source intention detection, and sexism categorization applied to three different types of media: text (tweets), images (memes), and videos (TikToks). We present each subtask for Task 1 as follows:

- Subtask 1.1 - Binary Sexism Detection: A binary classification where models predict whether a post is sexist or not. (6 920 samples: 3 367 non-sexist, 2 697 sexist)
- Subtask 1.2 - Category of Sexism: For sexist posts, a four-class classification to determine the type of sexism: Direct, Judgemental, Reported, Unknown (Total: 2 697 samples).
- Subtask 1.3 - Fine-grained Vector of Sexism: A 6-class classification further detailing sexist content: Ideological and inequality, Stereotyping and Dominance, Objectification, Sexual Violence, Misogyny and Non-Sexual Violence (Total: 2 697 samples)

The tweets dataset is divided into three distinct sets: train, development, and test (Table 1).

Table 1
EXIST 2025 Tweets Dataset for Task 1

Set	English Tweets	Spanish Tweets	Total
Train	3 260	3 660	6 920
Dev	489	549	1 038
Test	978	1 098	2 076

3.3. Pre-processing & Data Augmentation

Multiple pre-processing techniques were explored to prepare the textual data for analysis [39]. The initial **Light Cleaning** phase involved removing emojis, URLs, and mentions, as well as normalizing whitespace to reduce noise. Building on this, the **Classic Cleaning** step extended the process by removing non-ASCII characters, punctuation, and digits, and converting all text to lowercase for consistency. The third method, **Aggressive Cleaning**, included the steps from Classic Cleaning, followed by the removal of stopwords and the application of stemming to reduce words to their root forms.

Given the bilingual nature of the data, an augmentation step is included to add translated versions of the tweets. This step ensures that both Spanish and English content are represented more uniformly across the dataset, improving the generalization of the model. Two translation solutions were explored: the first used the Google Translate API, and the second leveraged the NLLB [13] (No Language Left Behind) model developed by Facebook. Both approaches aimed to maintain translation quality while handling domain-specific language effectively.

3.4. Model Architecture

For our experiments, we began by evaluating a range of transformer-based models (Table 2). The initial evaluation was conducted using the test data from Task 1. Our classification approach prioritized the most frequently occurring label per tweet; in cases of a tie, the tweet was classified as sexist. We also tested the impact of light pre-processing and used accuracy and F1-Score [40] metrics to determine the best-performing models (Table 3).

In the second phase, we focused on assessing the effectiveness of the three pre-processing strategies using the mDeBERTa model. Among Light, Classic, and Aggressive Cleaning, the Classic Cleaning approach yielded the best results (Table 4). Based on these initial findings, we selected mDeBERTa and XLM-RoBERTa-base for further training and evaluation across the three tasks, as both models demonstrated strong performance.

To enhance the training data, we implemented two data augmentation strategies. The first involved translation-based augmentation using Meta AI's NLLB model to generate high-quality translations and promote a more balanced bilingual dataset. The second strategy utilized external data augmentation

Table 2
The Proposed models for our experiments

Model Name	Multilingual
DeBERTa [11, 10]	No
mDeBERTa [11, 10]	Yes
XLM-RoBERTa [12]	Yes
HateBERT [41]	No
Detoxify [42]	No
RoBERTa-hate-speech [43]	No

Table 3
Transformers Performance Comparison Across Misogyny Detection Tasks 1

Model	Pre-processing	Accuracy	F1-No	F1-Yes
DeBERTa	Yes	0.80	0.82	0.77
DeBERTa	No	0.81	0.82	**0.79**
mDeBERTa-large	Yes	0.76	0.81	0.69
mDeBERTa-large	No	0.79	0.82	0.75
mDeBERTa-base	Yes	0.80	0.83	0.76
XLM-RoBERTa	Yes	0.80	0.83	0.76
XLM-RoBERTa	No	0.80	0.83	0.76
HateBERT	Yes	0.75	0.73	0.74
HateBERT	No	0.78	0.81	0.75
Detoxify	Yes	0.78	0.80	0.77
Detoxify	No	0.77	0.82	0.71
RoBERTa-hate-speech	Yes	0.76	0.80	0.68
RoBERTa-hate-speech	No	0.78	0.81	0.74

Table 4
Model Performance Comparison Based on pre-processing

Model	Pre-processing	Accuracy	F1-No	F1-Yes
mDeBERTa	Light	0.81	0.82	0.80
mDeBERTa	Classic	0.82	0.83	0.80
mDeBERTa	Aggressive	0.80	0.80	0.80

with the EDOS (Evaluation Dataset for Online Hate Speech) dataset [44]. For this, we applied a pseudo-labeling technique: a model trained on the original labeled data was used to predict labels for the unlabeled EDOS examples. To maintain label quality, we filtered predictions based on confidence scores, retaining only those above 80% for Task 1 and above 60% for Tasks 2 and 3, due to their increased complexity and subjectivity.

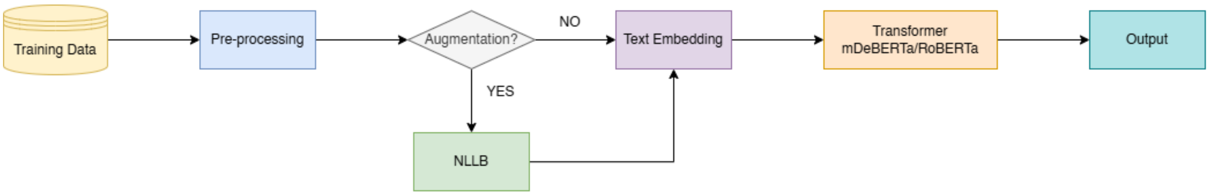

Figure 1: Proposed Architecture

This combined augmentation strategy aimed to increase both the volume and diversity of the training data while preserving label reliability, ultimately boosting the robustness and generalization of the model. As shown in Tables 5, 6, and 7, the best-performing configuration across all three tasks was achieved using the mDeBERTa model with Classic Cleaning and translation-based augmentation via the NLLB model, while excluding additional augmentation from the EDOS dataset as we can see in Figure 1.

Table 5
Task 1.1 - Results

Model	Pre-processing	Augmented Translation	Augmented EDOS	Accuracy	F1-Score
mDeBERTa	Classic	No	No	0.82	0.81
mDeBERTa	Classic	Yes	No	0.82	0.81
XLM-RoBERTa-base	Classic	Yes	No	0.81	0.80
mDeBERTa	Classic	Yes	Yes	0.81	0.81

Table 6
Task 1.2 - Results

Model	Pre-processing	Augmented Translation	Augmented EDOS	Accuracy	F1-Score
mDeBERTa	Classic	No	No	0.64	0.46
mDeBERTa	Classic	Yes	No	0.66	0.48
XLM-RoBERTa-base	Classic	Yes	No	0.65	0.36
mDeBERTa	Classic	Yes	Yes	0.65	0.37

Table 7
Task 1.3 - Results

Model	Pre-processing	Augmented Translation	Augmented EDOS	Accuracy	F1-Score
mDeBERTa	Classic	No	No	0.64	0.47
mDeBERTa	Classic	Yes	No	0.65	0.50
XLM-RoBERTa-base	Classic	Yes	No	0.64	0.45
mDeBERTa	Classic	Yes	Yes	0.65	0.48

4. Results

Tables 8 and 9 show the official leaderboard results. The evaluated model was mDeBERTa, using classic pre-processing without any data augmentation. It is hypothesized that incorporating translations could have led to improved performance.

The submitted system delivered competitive results in the EXIST2025 competition, particularly in the soft evaluation setting. As shown in Table 8, the system ranked within the top 15 for Task 1.1 across languages, reaching 11th place out of 66 for English and 14th overall. Performance was even stronger in Task 1.2, with the system placing 8th out of 56, demonstrating robustness across both English and Spanish subtasks. However, results were less competitive for Task 1.3, with rankings around the middle of the leaderboard. In the hard evaluation (Table 9), rankings were generally lower, with a peak performance of 38th place for Task 1.1 in Spanish, while the English results lagged behind. These outcomes suggest that while the model generalized well under soft evaluation metrics, it struggled to maintain top-tier performance under strict classification thresholds. The absence of augmentation and translation strategies might have limited the model's adaptability, particularly in handling language-specific nuances and subtler forms of sexist content.

Table 8
Ranking and Results in EXIST2025 competition - Soft

Task	Evaluation	Rank	Total systems	ICM-Soft	ICM-Soft-Norm	Cross Entropy
1.1	Soft-Soft-All	14	67	0.6707	0.6076	0.8774
1.1	Soft-Soft-ES	21	64	0.6737	0.6080	0.8558
1.1	Soft-Soft-EN	11	66	0.6295	0.6011	0.9015
1.2	Soft-Soft-All	8	56	-1.3444	0.3917	1.5681
1.2	Soft-Soft-ES	9	54	-1.1955	0.4043	1.5688
1.2	Soft-Soft-EN	8	55	-1.6199	0.3676	1.5674
1.3	Soft-Soft-All	17	53	-3.9061	0.2937	-
1.3	Soft-Soft-ES	16	49	-3.8908	0.2975	-
1.3	Soft-Soft-EN	17	52	-3.8994	0.2863	-

Table 9
Ranking and Results in EXIST2025 competition - Hard

Task	Evaluation	Rank	Total systems	ICM-Hard	ICM-Hard-Norm	F1-Score
1.1	Hard-Hard-All	55	160	0.4966	0.7496	0.7553
1.1	Hard-Hard-ES	38	153	0.5112	0.7557	0.7775
1.1	Hard-Hard-EN	102	158	0.4664	0.7380	0.7261
1.2	Hard-Hard-All	54	140	0.0969	0.5315	0.4567
1.2	Hard-Hard-ES	56	136	0.1045	0.5327	0.4733
1.2	Hard-Hard-EN	59	139	0.0663	0.5229	0.4334
1.3	Hard-Hard-All	48	132	-0.1516	0.4648	0.4357
1.3	Hard-Hard-ES	47	128	-0.1454	0.4675	0.4599
1.3	Hard-Hard-EN	61	131	-0.1692	0.4585	0.3968

5. Conclusion and future directions

The experiments conducted in this study explored various transformer models, pre-processing techniques, and data augmentation strategies to address the EXIST2025 shared task on sexist content detection. Among the evaluated approaches, the combination of the mDeBERTa model with Classic Cleaning pre-processing consistently demonstrated strong performance, particularly in the soft evaluation setting. Despite not incorporating data augmentation or translation-based enhancements in the final submission, the system achieved competitive results, thus ranking within the top 15 for multiple subtasks.

However, the lower rankings in the hard evaluation suggest that the model's robustness could be further improved by integrating additional strategies, such as translation-based augmentation and external datasets like EDOS. These findings underscore the importance of both model selection and pre-processing, while also highlighting the potential impact of thoughtful data augmentation. Future work will focus on refining augmentation techniques and exploring ensemble approaches to enhance classification performance, especially under stricter evaluation conditions and across multiple languages.

Acknowledgments

The research presented in this paper was supported in part by (1) The Academy of Romanian Scientists through the funding of the project "SCAN-NEWS: Smart system for deteCting And mitigatiNg misinformation and fake news in social media" (AOȘR-TEAMS-III); (2) The Academy of Romanian Scientists through the funding of the project "NetGuardAI: Intelligent system for harmful content detection and immunization on social networks" (AOȘR-TEAMS-IV); and (3) the National University of Science and Technology, POLITEHNICA Bucharest, through the PubArt program.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] E. A. Vogels, The State of Online Harassment, 2021. URL: https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/, accessed: 26 March 2025.

[2] A. Petrosyan, U.S. online harassment environments 2020, by gender, 2023. URL: https://www.statista.com/statistics/334074/online-harassment-environments-gender/, accessed: 26 March 2025.

[3] S. J. Dixon, Effects of harmful online contact worldwide 2022, by gender, 2022. URL: https://www.statista.com/statistics/1301606/effects-of-harmful-online-contact/, accessed: 26 March 2025.

[4] The Female Lead, Young women are leaving social media - and it's not hard to guess why, 2025. URL: https://community.thefemalelead.com/p/young-women-are-leaving-social-media, accessed: 2025-05-15.

[5] S. Galdi, F. Guizzo, Media-induced sexual harassment: The routes from sexually objectifying media to sexual harassment, Sex Roles 84 (2021) 645–669.

[6] L. D. Meco, Monetizing Misogyny: Gendered Disinformation and the Undermining of Women's Rights and Democracy Globally, Technical Report, #ShePersisted, 2023. URL: https://she-persisted.org/wp-content/uploads/2023/02/ShePersisted_MonetizingMisogyny.pdf.

[7] L. Plaza, J. Carrillo-de Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, EXIST 2025: Learning with Disagreement for Sexism Identification and Characterization in Tweets, Memes, and TikTok Videos, in: European Conference on Information Retrieval, Springer, 2025, pp. 442–449.

[8] L. Plaza, J. Carrillo-de Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, Overview of exist 2025: Learning with disagreement for sexism identification and characterization in tweets, memes, and tiktok videos. experimental ir meets multilinguality, multimodality, and interaction., in: Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

[9] L. Plaza, J. Carrillo-de Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, Overview of exist 2025: Learning with disagreement for sexism identification and characterization in tweets, memes, and tiktok videos (extended overview)., in: CLEF 2025 Working Notes, 2025.

[10] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, in: International Conference on Learning Representations, 2023.

[11] P. He, X. Liu, J. Gao, W. Chen, DeBERTa: Decoding-enhanced BERT with Disentangled Attention, in: International Conference on Learning Representations, 2021.

[12] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised Cross-lingual Representation Learning at Scale (2020) 8440–8451. doi:10.18653/v1/2020.acl-main.747.

[13] M. R. Costa-Jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, et al., No language left behind: Scaling human-centered machine translation, arXiv preprint arXiv:2207.04672 (2022).

[14] V.-I. Ilie, C.-O. Truică, E.-S. Apostol, A. Paschke, Context-Aware Misinformation Detection: A Benchmark of Deep Learning Architectures Using Word Embeddings, IEEE Access 9 (2021) 162122–162146. doi:10.1109/ACCESS.2021.3132502.

[15] C.-O. Truică, E.-S. Apostol, MisRoBÆRTa: Transformers versus Misinformation, Mathematics 10 (2022) 1–25(569). doi:10.3390/math10040569.

[16] C.-O. Truică, E.-S. Apostol, A. Paschke, Awakened at CheckThat! 2022: Fake News Detection using BiLSTM and sentence transformer, in: Working Notes of the Conference and Labs of the Evaluation Forum, 2022, pp. 749–757.

[17] C.-O. Truică, E.-S. Apostol, It's all in the Embedding! Fake News Detection using Document Embeddings, Mathematics 11 (2023) 1–29(508). doi:10.3390/math11030508.

[18] A. Petrescu, C.-O. Truică, E.-S. Apostol, Language-based Mixture of Transformers for EXIST2024, in: Working Notes of the Conference and Labs of the Evaluation Forum, volume 3740 of *CEUR Workshop Proceedings*, 2024, pp. 1157–1164.

[19] C.-O. Truică, E.-S. Apostol, P. Karras, DANES: Deep Neural Network Ensemble Architecture for Social and Textual Context-aware Fake News Detection, Knowledge-Based Systems 294 (2024) 1–13(111715). doi:https://doi.org/10.1016/j.knosys.2024.111715.

[20] C.-O. Truică, E.-S. Apostol, M. Marogel, A. Paschke, GETAE: Graph Information Enhanced Deep Neural NeTwork Ensemble ArchitecturE for fake news detection, Expert Systems with Applications 275 (2025) 126984. doi:10.1016/j.eswa.2025.126984.

[21] A. Petrescu, C.-O. Truică, E.-S. Apostol, Sentiment Analysis of Events in Social Media, in: 2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP), IEEE, 2019, pp. 143–149. doi:10.1109/iccp48234.2019.8959677.

[22] A. Petrescu, C.-O. Truică, E.-S. Apostol, A. Paschke, EDSA-Ensemble: an Event Detection Sentiment Analysis Ensemble Architecture, IEEE Transactions on Affective Computing 16 (2024) 555–572. doi:10.1109/TAFFC.2024.3434355.

[23] C.-O. Truică, E.-S. Apostol, T. Ștefu, P. Karras, A Deep Learning Architecture for Audience Interest Prediction of News Topic on Social Media, in: International Conference on Extending Database Technology (EDBT2021), 2021, pp. 588–599. doi:10.5441/002/EDBT.2021.69.

[24] A. Petrescu, C.-O. Truică, E.-S. Apostol, P. Karras, Sparse Shield: Social Network Immunization vs. Harmful Speech, in: ACM International Conference on Information and Knowledge Management (CIKM2021), ACM, 2021, pp. 1426–1436. doi:10.1145/3459637.3482481.

[25] C.-O. Truică, E.-S. Apostol, R.-C. Nicolescu, P. Karras, MCWDST: A Minimum-Cost Weighted Directed Spanning Tree Algorithm for Real-Time Fake News Mitigation in Social Media, IEEE Access 11 (2023) 125861–125873. doi:10.1109/ACCESS.2023.3331220.

[26] E.-S. Apostol, Özgur Coban, C.-O. Truică, CONTAIN: A community-based algorithm for network immunization, Engineering Science and Technology, an International Journal 55 (2024) 1–10(101728). doi:https://doi.org/10.1016/j.jestch.2024.101728.

[27] E.-S. Apostol, C.-O. Truică, A. Paschke, ContCommRTD: A Distributed Content-Based Misinformation-Aware Community Detection System for Real-Time Disaster Reporting, IEEE Transactions on Knowledge and Data Engineering (2024) 1–12. doi:10.1109/tkde.2024.3417232.

[28] C.-O. Truică, A.-T. Constantinescu, E.-S. Apostol, StopHC: A Harmful Content Detection and Mitigation Architecture for Social Media Platforms, in: IEEE International Conference on Intelligent Computer Communication and Processing (ICCP 2024), 2024, pp. 1–5. doi:10.1109/ICCP63557.2024.10793051.

[29] S. Fan, R. A. Frick, M. Steinebach, FraunhoferSIT@ EXIST2024: leveraging stacking ensemble learning for sexism detection, Working Notes of CLEF (2024).

[30] S. K. Murari Sreekumar, D. Thenmozhi, S. Gopalakrishnan, K. Swaminathan, Sexism Identification In Tweets Using Machine Learning Approaches, Working Notes of CLEF (2024).

[31] M. Usmani, R. Siddiqui, S. Rizwan, F. Khan, F. Alvi, A. Samad, Sexism identification in tweets using BERT and XLM–Roberta, Working Notes of CLEF (2024).

[32] A. Shah, A. Gokhale, Team Aditya at EXIST 2024–detecting sexism in multilingual tweets using contrastive learning approach, Working Notes of CLEF (2024).

[33] A. Azadi, B. Ansari, S. Zamani, S. Eetemadi, Bilingual sexism classification: fine-tuned XLM-RoBERTa and GPT-3.5 few-shot learning, arXiv preprint arXiv:2406.07287 (2024).

[34] Y.-Z. Fang, L.-H. Lee, J.-D. Huang, NYCU-NLP at EXIST 2024–leveraging transformers with diverse annotations for sexism identification in social networks, Working Notes of CLEF (2024).

[35] R. Pan, J. A. García Díaz, T. Bernal Beltrán, R. Valencia-García, UMUTeam at EXIST 2024: multimodal identification and categorization of sexism by feature integration, Working Notes of CLEF (2024).

[36] J. Ma, R. Li, RoJiNG-CL at EXIST 2024: sexism identification in memes by integrating prompting and fine-tuning, Working Notes of CLEF (2024).

[37] L. Quan, D. Thin, Sexism identification in social networks with generation-based approach, Working Notes of CLEF (2024).

[38] M. Siino, I. Tinnirello, Prompt engineering for identifying sexism using GPT mistral 7B, Working Notes of CLEF (2024).

[39] C.-O. Truică, J. Darmont, J. Velcin, A Scalable Document-Based Architecture for Text Analysis, in: International Conference on Advanced Data Mining and Applications, Springer, 2016, pp. 481–494. doi:10.1007/978-3-319-49586-6_33.

[40] C.-O. Truică, C. A. Leordeanu, Classification of an imbalanced data set using decision tree algorithms, Univiversity Politechnica of Bucharest Scientific Bulletin - Series C Electrical Engineering and Computer Science 79 (2017) 69–84.

[41] T. Caselli, V. Basile, J. Mitrović, M. Granitzer, HateBERT: Retraining BERT for Abusive Language Detection in English, in: Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021), 2021, pp. 17–25. doi:10.18653/v1/2021.woah-1.3.

[42] L. Hanu, Unitary team, Detoxify, Github. https://github.com/unitaryai/detoxify, 2020.

[43] B. Vidgen, T. Thrush, Z. Waseem, D. Kiela, Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, 2021, pp. 1667–1682. doi:10.18653/v1/2021.acl-long.132.

[44] H. R. Kirk, W. Yin, B. Vidgen, P. Röttger, Semeval-2023 task 10: Explainable detection of online sexism, in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), ACL, 2023, pp. 2193–2210. doi:10.18653/v1/2023.semeval-1.305.