Multilingual Sexism Detection through Domain Adaptation and Label-Augmented Translation

Notebook for the EXIST Lab at CLEF 2025

Sahrish Khan[1], Arshad Jhumka[2] and Gabriele Pergola[1]

[1]Department of Computer Science, University of Warwick, Coventry CV4 7AL, UK
[2]School of Computing, University of Leeds, Leeds LS2 9JT, UK

Abstract

Detecting sexism in online discourse presents persistent challenges for multilingual NLP systems, especially when dealing with culturally specific expressions and subtle contextual cues. In our submission to the EXIST 2025 Task 1.1, we propose a pipeline that combines label-aware translation, domain-adaptive pre-training, and ensemble learning to address these challenges. A central component of our system is a prompt-based Spanish-to-English translation step, designed to preserve the tone and task-relevant semantics of the original message, selectively incorporating label cues during training. This enables the use of high-performance monolingual models while maintaining semantic fidelity across languages. We further adapt DeBERTa-v3-Large and RoBERTa-Large using 2 million unlabeled posts from the EDOS dataset and fine-tune them individually and in a fused configuration (DTFN). Final predictions are generated via majority voting, with a tie-handling rule that improves robustness. In the experimental assessment, our system achieved 4th place in English and ranked in the top 10 across all evaluation tracks.

Keywords

Sexism Detection, Multilingual NLP, Domain-adaptive pre-training, Prompt-based translation, Ensemble learning, EXIST Task 1.1

1. Introduction

Sexism or discrimination based on gender continues to manifest in diverse ways across online discourse. In social media, sexist remarks can spread rapidly and take many forms, from casual stereotyping to overt hostility [1]. Detecting such content reliably across languages and platforms remains a key challenge in natural language processing (NLP), particularly when subtle contextual cues or culturally-specific expressions are involved. Multilingual sexism detection presents an added layer of complexity: language-specific expressions of bias can differ in tone, terminology, and structure. While large multilingual models offer a direct solution, they often struggle to balance cross-lingual generalization with semantic precision. Our system, submitted to EXIST 2025 Task 1.1, takes a different approach. We combine domain-adaptive pre-training, label-aware translation, and ensemble learning to construct a pipeline that preserves the semantics of bias-related content while maximizing classification robustness.

A central component of our pipeline is a prompt-based Spanish-to-English translation step. The Large Language Model's prompt for translation is carefully designed to preserve the tone and intent of the original message and, when training labels are available, selectively incorporates task-relevant cues aligned with the sexism taxonomy. This translation strategy therefore acts as augmentation strategy and allows the models to operate fully within an English-language modeling space, taking advantage of the wider development of monolingual neural models, while mitigating semantic drift. We complement this approach with model-level adaptation and aggregation. Specifically, we domain-adapt

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
sahrish.khan@warwick.ac.uk (S. Khan); h.a.jhumka@leeds.ac.uk (A. Jhumka); gabriele.pergola.1@warwick.ac.uk (G. Pergola)
https://www.dcs.warwick.ac.uk/~u2149613/ (S. Khan);
https://eps.leeds.ac.uk/computing/staff/9540/professor-arshad-jhumka (A. Jhumka);
https://www.dcs.warwick.ac.uk/~u1898418/ (G. Pergola)
0000-0003-0540-2845 (A. Jhumka); 0000-0002-7347-2522 (G. Pergola)

DeBERTa-v3-Large and RoBERTa-Large using 2 million unlabeled tweets drawn from the *Explainable Detection of Online Sexism* (EDOS) dataset [2], and fine-tune these models—individually and in a fused configuration (DTFN) [3]. Final predictions are produced via ensemble learning, using majority voting and a tie-handling rule to increase consistency across prediction variants. Together, the label-informed translation design and ensemble-based decision mechanism enable strong cross-lingual performance while maintaining conceptual and architectural clarity. Our system ranked 4th in English and achieved top-10 scores across all tracks in the EXIST 2025 shared task.

The contributions of this work can be summarized as follows:

- We introduce a label-aware, prompt-based translation strategy that preserves semantics relevant to sexism detection across languages.
- We combine domain-adaptive pre-training with a majority-vote ensembling for robust multilingual classification.
- We provide detailed empirical results and analysis on the EXIST 2025 benchmark, achieving high rankings in all evaluation tracks.

2. Related Work

Detecting sexism in online content has become a growing focus in the NLP community, driven by the widespread use of social media and the urgent need to mitigate harmful language. As interest in the task has grown, a series of shared tasks have emerged to benchmark progress and foster collaboration across the research community [4, 5, 6, 7, 8, 9, 10, 11]. A key initiative in this space is the *EXIST shared task*, launched in 2021 as part of IberLEF. Running annually through 2024, EXIST has established itself as a central benchmark for multilingual sexism detection, focusing on English and Spanish social media content. The datasets span diverse platforms such as Gab and X (formerly Twitter), reflecting real-world linguistic variation and user behavior [12, 13, 14].

Building on this foundation, other shared tasks have expanded the scope of the problem. SemEval-2022 introduced Task 5, *Multimedia Automatic Misogyny Identification (MAMI)*, which explored multimodal misogyny detection in memes, combining image and text analysis [15]. This was followed by SemEval-2023's Task 10, *Explainable Detection of Online Sexism (EDOS)*, which emphasized explainability by introducing fine-grained sexist categories, ranging from broad groupings to eleven specific subtypes. EDOS also emphasized source diversity, drawing data from Reddit and Gab to challenge generalization across platforms [2].

Although these tasks contributed valuable insights into modality and taxonomy expansion, the EXIST 2024 shared task returned its focus to multilingual detection, this time with greater methodological variety and evaluation depth. In this edition, top-performing systems explored a diverse range of strategies to address the challenges of multilingual sexism detection [16]. Several teams focused on leveraging ensemble learning, combining outputs from transformer-based models such as DeBERTa, RoBERTa, Mistral, and XLM-R, often enhanced through domain-adaptive pre-training [3, 17, 16].

In parallel, a growing number of studies explored translation-based approaches to better manage multilingual input. By translating all content into a single language, researchers could apply monolingual models across languages, thus simplifying the learning process [18, 19, 20]. Recent ensemble methods include adaptive model selection, where systems dynamically choose which models to include based on the input and confidence weighted fusion, which assigns weights to each model's output based on its predicted certainty. This kind of aggregation brings complementary perspectives together, helping capture the varied and nuanced expressions of online sexism. Together, these strategies reflect a wider trend in the field: towards flexible layered systems that integrate translation, ensemble learning, and model tuning to better address the complexity of online sexism detection [5, 17, 21, 22, 3, 4, 23, 24, 25].

Our system adopts a focused and reproducible strategy for multilingual sexism detection. We unify Spanish and English inputs through prompt-guided translation using GPT-4o, carefully preserving tone and domain-specific language. This allowed us to operate entirely within an English-language setup, leveraging powerful monolingual models.

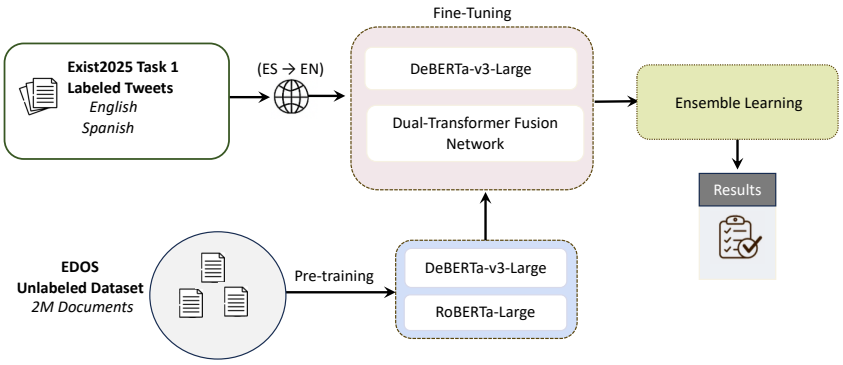

Figure 1: Overview of our pipeline for sexism detection in the EXIST2025 Task 1.1. Spanish tweets are translated to English using GPT-4o before fine-tuning. Models are initialized with domain-adaptive pre-training on 2M unlabeled EDOS documents. Final predictions are generated via ensemble learning.

3. Pipeline Overview

To tackle multilingual sexism detection effectively, we designed a simple yet powerful pipeline built around four key stages (Figure 1). First, we adapt transformer models to the particular style of online messages through large-scale pre-training. Second, we unify the input space by translating all Spanish tweets into English using GPT-4o and a task-aware prompt. Third, we fine-tune two high-performing models—DeBERTa-v3-Large and our Dual-Transformer Fusion Network (DTFN) [3], on the labeled dataset, Finally, we apply ensemble learning using majority voting to combine the predictions from models.

These four stages are detailed in the following sections, covering pre-training, translation, model design, and ensemble learning.

3.1. Domain-Adaptive Pre-training

As the initial step in our pipeline, we performed domain-adaptive pre-training on DeBERTa-v3-Large and RoBERTa-Large using a large corpus of 2 million unlabeled posts from the EDOS dataset [2], which includes user-generated content from platforms such as Gab and Reddit. These pre-trained models were then used in the next stages of the pipeline: or as a standalone DeBERTa-v3-Large, or by combining them, as in the case of DeBERTa-v3-Large and RoBERTa-Large combined into the Dual-Transformer Fusion Network (DTFN), a dual-encoder architecture introduced and discussed in earlier work [3].

3.2. Label-Augmented Prompt-Based ES→EN Translation

To keep the the multilingual processing pipeline simple, rather than introducing dedicated components, we translated all Spanish tweets into English using GPT-4o [26] and a carefully designed prompt. This allowed us to work entirely within an English-language setup, enabling the use of monolingual transformer models, such as DeBERTa-v3-Large and DTFN.This strategy was guided by both practical considerations, as English transformer models are typically more robust and better optimized due to pre-training on larger, more diverse corpora.

To ensure consistency, we applied the same translation approach and prompt uniformly across the training, development, and test sets. For the training data, the label was selectively used to guide the translation process, particularly to preserve task-relevant language and semantic cues associated with sexism. However, for fair comparison and evaluation, at test time no label information was used during translation.

The prompt was specifically crafted to preserve the tone and intent of each tweet. If the content was sexist, GPT-4o was guided to retain that tone and, when appropriate, use terminology aligned with the sexism typology (e.g., misogyny, objectification, mansplaining), aligned with the EDOS [2] taxonomy (see Figure 2).

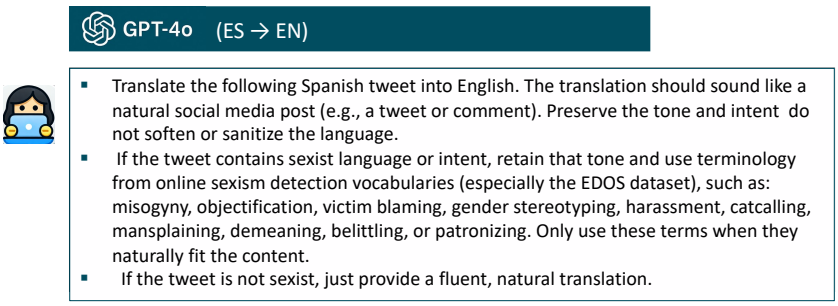

Figure 2: Prompt used for GPT-4o-based Spanish-to-English translation. The design encourages tone preservation and selective incorporation of sexism-related terminology, aligned with the EDOS taxonomy [2].

3.3. Model Fine-tuning

Following domain-adaptive pre-training and translation, we fine-tuned DeBERTa-v3-Large and our Dual-Transformer Fusion Network (DTFN) on the labeled dataset provided in EXIST 2025. The training objective was to minimize the binary cross-entropy loss, formulated as:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

where N is the number of training samples, $y_i \in \{0, 1\}$ is the ground-truth label for the i-th sample, and $\hat{y}_i \in [0, 1]$ is the predicted probability of the sample being classified as sexist.

3.4. Ensemble Learning with Majority Voting

To further improve the consistency and variety of predictions, we employed a majority voting ensemble over the best fine-tuned models. Specifically, we combined outputs from DeBERTa-v3-Large and DTFN. As in our prior work [3], we adopted the same tie-handling rule, defaulting to the "sexist" label in ambiguous cases.

4. Experimental Assessment

Dataset

Table 1
Distribution of Data for Task 1.1: Sexism Detection in the EXIST 2025 Dataset

Language	Training Set	Development Set	Test Set
English	3,260	489	978
Spanish	3,460	549	1,098
Both Languages	6,920	1,038	2,076

The experimental assessment was conducted on the EXIST 2025 Tweets Dataset, designed for Task 1.1 on sexism detection. The dataset comprises over 10,000 tweets in English and Spanish, each annotated for binary classification, labeling posts as either sexist ("YES") or not sexist ("NO").

The dataset is split into three parts: training, development, and test sets. All models were trained using gold-standard (hard) labels from both English and Spanish subsets in the training. The full distribution of the dataset is summarized in Table 1.

Baselines

In the following, we briefly describe the baselines evaluated:

- **RoBERTa-Large [27]:** An optimized variant of BERT that removes the Next Sentence Prediction (NSP) objective and leverages larger training corpora and longer training durations.
- **DeBERTa-V3-Large [28]:** An advancement over BERT and RoBERTa that introduces a disentangled attention mechanism and an enhanced masked language modeling objective, with strong performance in low-resource or noisy data settings.
- **DTFN (Dual-Transformer Fusion Network) [3]:** Our previously proposed architecture that integrates representations from both RoBERTa-Large and DeBERTa-V3-Large using a dual-transformer fusion strategy.

We adopted the publicly available pre-trained versions of all models from the HuggingFace Transformers library[1], and further performed domain-adaptive pre-training on 2 million unlabeled samples from the EDOS dataset. These adapted models were then fine-tuned on the EXIST 2025 Task 1.1 dataset under both baseline and translation-based setups.

Parameter Settings

For each model used in our experiments, we identified a set of optimal hyperparameters through preliminary testing. The selected hyperparameters include the number of training epochs, learning rate (η), batch size, and weight decay (λ). Table 2 presents the optimal hyperparameters for each model.

Table 2
Best Hyperparameters per Model

Hyperparameter	RoBERTa-Large	DeBERTa-V3-Large	DTFN
Number of Epochs	30	30	30
Learning Rate	6×10^{-6}	6×10^{-6}	6×10^{-6}
Batch Size	16	16	4
Weight Decay	5×10^{-3}	5×10^{-3}	5×10^{-3}

Evaluation Metrics

In Task 1.1 of EXIST 2025, system performance was evaluated using three official metrics: ICM-Hard, ICM-Hard Norm, and F1-score. The ICM metric, introduced by Amigo and Delgado [29], is grounded in information theory and quantifies the similarity between a system's predictions and the gold-standard labels. To account for class imbalance in the dataset, the organizers also reported a normalized variant, ICM-Hard Norm, enabling fairer comparisons across systems. For all three metrics, higher scores indicate better alignment with the reference annotations, with ICM-Hard Norm offering particular robustness in uneven label distributions.

5. Results

5.1. Experimental Results on the Development set

We evaluated both individual model variants and ensemble configurations on the development set. As shown in Table 3, our experiments involved DeBERTa-v3-Large and DTFN, each domain-adaptively pre-trained on 2 million unlabeled tweets from the EDOS dataset (for 7 or 9 epochs) and subsequently fine-tuned for 30 epochs on the labeled EXIST 2025 training set. These evaluations were conducted

[1] https://huggingface.co/

under two setups: the *baseline setting*, which used the original bilingual data (English and Spanish tweets), and the *translation-based* setting, where all Spanish tweets were first translated into English using our GPT-4o prompt.

In the baseline setup (Index 1–4), models were fine-tuned on both language input. DTFN slightly outperformed DeBERTa-v3-Large at 9 epochs (ICM-Hard: 0.6019 vs. 0.6012), while F1 scores remained closely aligned. Ensembling these models resulted in a noticeable improvement (ICM-Hard: 0.6533, F1: 0.8800), illustrating the value of combining model representations even in a multilingual context. The translation-based setup (Index 6–9) led to further performance gains. Notably, DTFN—pre-trained for 7 epochs and fine-tuned on translated Spanish tweets achieved strong standalone results (ICM-Hard: 0.6214, F1: 0.8736; see Index 8). The best overall scores were obtained by ensembling all translated variants via majority voting (Index 10), reaching an ICM-Hard of 0.6864 and an F1 score of 0.8950. These findings highlight the effectiveness of our lightweight multilingual strategy, achieving strong performance without resorting to complex multilingual systems.

Table 3
Performance of individual models and ensemble variants on the development set under different pre-training configurations (7 and 9 epochs on EDOS) and input settings (baseline vs. translated). All models were fine-tuned for 30 epochs on the labeled EXIST 2025 training set. The best results were achieved by ensembling DeBERTa-v3-Large and DTFN on translated inputs.

Index	Model	Pre-training	ICM-Hard	F1 Score
Baseline Models (English + Spanish)				
1	DeBERTa-v3-large	7	0.6012	0.8671
2	DeBERTa-v3-large	9	0.5977	0.8660
3	DTFN	7	0.5910	0.8638
4	DTFN	9	0.6019	0.8672
5	**Ensemble Learning**	-	0.6533	0.8800
Translation Models (ES → EN)				
6	DeBERTa-v3-large (TR)	7	0.6108	0.8703
7	DeBERTa-v3-large (TR)	9	0.5864	0.8618
8	DTFN (TR)	7	0.6214	0.8736
9	DTFN (TR)	9	0.6078	0.8693
10	**Ensemble Learning**	-	**0.6864**	**0.8950**

5.2. Results of the Official Leaderboard

Table 4 presents our official results on the EXIST 2025 Task 1.1 test set, as submitted under the team name *EquityExplorer-2.0*. We submitted two system variants for each language-specific and multilingual evaluation track.

In Run 1, we used our translation-based setup: all Spanish tweets were first translated into English using our GPT-4o prompt, and models were trained on this unified English dataset. In Run 2, we submitted results from our baseline configuration, where models were trained on the original bilingual data (English + Spanish).

Across both runs, our system achieved strong, consistent performance. In the English-only track, Run 1 ranked 4^{th} out of 158 teams with an ICM-Hard score of 0.6018 and an F1 score of 0.7685. Even in Run 2, without translation, our model remained in the top 10 (8^{th}/158), demonstrating the strength of our pre-training and ensemble components. In the Spanish track, performance was slightly lower but still competitive. Run 1 achieved an ICM-Hard of 0.5490 and ranked 24^{th}/153, while Run 2 slightly improved in F1 score (0.7967) and ranked 21^{th}.

For the combined multilingual track, our translation-based pipeline (Run 1) outperformed the baseline, placing 10^{th} out of 160 teams with an ICM-Hard of 0.5806 and an F1 score of 0.7837. The baseline configuration (Run 2), which trained directly on English and Spanish tweets without translation, also

Table 4
Official results on the test set for EXIST 2025 Task 1.1 (Sexism Detection). Performance across English, Spanish, and combined multilingual tracks under two configurations: Run 1 (translation-based ES→EN pipeline) and Run 2 (baseline bilingual input). The best overall results were achieved using the translation-based setup in Run 1.

Evaluation	Language	Team (*EquityExplorer-2.0*)	ICM-Hard	ICM-Hard Norm	F1	Rank
Hard-Hard	English	Run 1	**0.6018**	**0.8071**	**0.7685**	**4th/158**
Hard-Hard	English	Run 2	0.5884	0.8003	0.7638	**8th/158**
Hard-Hard	Spanish	Run 1	0.5490	0.7746	0.7949	24^{th}/153
Hard-Hard	Spanish	Run 2	**0.5553**	**0.7777**	**0.7967**	21^{th}/153
Hard-Hard	Both	Run 1	**0.5806**	**0.7918**	**0.7837**	10^{th}/160
Hard-Hard	Both	Run 2	0.5779	0.7904	0.7827	13^{th}/160

performed competitively—ranking 13th with an ICM-Hard of 0.5779 and F1 score of 0.7827. This small performance gap further reinforces the effectiveness of our prompt-based ES→EN translation strategy in multilingual settings.

6. Conclusion

In this paper, we introduced a compact yet effective system for multilingual sexism detection in the EXIST 2025 shared task 1.1. Our approach combines domain adaptive pre-training, prompt-based GPT-4o translation, and lightweight ensemble learning, all designed to maximize performance while keeping the pipeline simple and reproducible.

By translating Spanish tweets into English using a carefully crafted prompt, we were able to work within an English-only setup and apply monolingual transformers effectively. We fine-tuned DeBERTa-v3-Large and our previously proposed DTFN, and combined their predictions through majority voting with a simple tie-handling rule.

This minimal design achieved competitive results across all language tracks, including 4th place in English and top-10 rankings overall. Our findings reinforce that thoughtful reuse of existing components combined with strategic translation and fine-tuning can match or even exceed the performance of more complex architectures in real world tasks like online sexism detection.

In future work, we aim to explore more adaptive translation strategies, and investigate the model's behavior on emerging forms of online bias.

Acknowledgements

Sahrish Khan, Gabriele Pergola, and Arshad Jhumka, were partially supported by the Police Science, Technology, Analysis, and Research (STAR) Fund 2022–23 and 2025-26, funded by the National Police Chiefs' Council (NPCC), in collaboration with the Forensic Capability Network (FCN). Gabriele Pergola was partially supported by the ESRC-funded project *Digitising Identity: Navigating the Digital Immigration System and Migrant Experiences*, as part of the Digital Good Network. This work was conducted on the Sulis Tier-2 HPC platform hosted by the Scientific Computing Research Technology Platform at the University of Warwick. Sulis is funded by EPSRC Grant EP/T022108/1 and the HPC Midlands+ consortium.

Declaration on Generative AI

The authors used generative AI tools (GPT-4o) in two ways: (1) as part of the experimental pipeline, to translate Spanish tweets into English while preserving task-relevant semantics, and (2) for minor

language editing, including grammar correction and paraphrasing, during manuscript preparation. All research ideas, analyses, results, and conclusions are the authors' own.

References

[1] X. Luo, B. Liang, Q. Wang, J. Li, E. Cambria, X. Zhang, Y. He, M. Yang, R. Xu, A literature survey on multimodal and multilingual sexism detection, IEEE Transactions on Computational Social Systems (2025) 1–19. doi:10.1109/TCSS.2025.3561921.

[2] H. Kirk, W. Yin, B. Vidgen, P. Röttger, SemEval-2023 task 10: Explainable detection of online sexism, in: A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2193–2210.

[3] S. Khan, G. Pergola, A. Jhumka, Multilingual sexism identification via fusion of large language models, in: E. names (Ed.), Joint Proceedings of the ACM IUI 2024 Workshops, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, Grenoble, France, 2024. URL: https://ceur-ws.org/Vol-3740/paper-99.pdf.

[4] X. Tan, C. Lyu, H. M. Umer, S. Khan, M. Parvatham, L. Arthurs, S. Cullen, S. Wilson, A. Jhumka, G. Pergola, SafeSpeech: A comprehensive and interactive tool for analysing sexist and abusive language in conversations, in: N. Dziri, S. X. Ren, S. Diao (Eds.), Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations), Association for Computational Linguistics, Albuquerque, New Mexico, 2025.

[5] S. Khan, A. Jhumka, G. Pergola, Explaining matters: Leveraging definitions and semantic expansion for sexism detection, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2025.

[6] L. Zhu, G. Pergola, L. Gui, D. Zhou, Y. He, Topic-driven and knowledge-aware transformer for dialogue emotion detection, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), 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, Online, 2021, pp. 1571–1582.

[7] J. Lu, X. Tan, G. Pergola, L. Gui, Y. He, Event-centric question answering via contrastive learning and invertible event transformation, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2022, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 2377–2389.

[8] Z. Sun, G. Pergola, B. Wallace, Y. He, Leveraging ChatGPT in pharmacovigilance event extraction: An empirical study, in: Y. Graham, M. Purver (Eds.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 344–357.

[9] C. Lyu, , G. Pergola, SciGisPy: a novel metric for biomedical text simplification via gist inference score, in: M. Shardlow, H. Saggion, F. Alva-Manchego, M. Zampieri, K. North, S. Štajner, R. Stodden (Eds.), Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024), Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 95–106.

[10] X. Tan, Y. Zhou, G. Pergola, Y. He, Set-aligning framework for auto-regressive event temporal graph generation, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 3872–3892.

[11] X. Tan, Y. Zhou, G. Pergola, Y. He, Cascading large language models for salient event graph generation, in: L. Chiruzzo, A. Ritter, L. Wang (Eds.), Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human

Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 2223–2245. URL: https://aclanthology.org/2025.naacl-long.112/.

[12] F. J. Rodríguez-Sanchez, J. C. de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, T. Donoso, Overview of exist 2021: sexism identification in social networks, Proces. del Leng. Natural 67 (2021) 195–207.

[13] L. Plaza, J. C. 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: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025, pp. –. To appear.

[14] L. Plaza, J. C. 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, in: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. G. S. de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, Springer, 2025, pp. –. To appear.

[15] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, J. Sorensen, SemEval-2022 task 5: Multimedia automatic misogyny identification, in: G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, S. Ratan (Eds.), Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 533–549.

[16] 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, in: CLEF (Working Notes), 2024, pp. 1003–1011. URL: https://ceur-ws.org/Vol-3740/paper-93.pdf.

[17] J. Tavarez-Rodríguez, F. Sánchez-Vega, A. Rosales-Pérez, A. P. López-Monroy, Better together: Llm and neural classification transformers to detect sexism, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[18] F. Manzi, L. Weber-Genzel, B. Plank, Fine-grained sexism detection in Italian newspapers, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 556–583. URL: https://aclanthology.org/2024.clicit-1.66/.

[19] M. Chen, P.-A. Duquenne, P. Andrews, J. Kao, A. Mourachko, H. Schwenk, M. R. Costa-jussà, BLASER: A text-free speech-to-speech translation evaluation metric, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 9064–9079. URL: https://aclanthology.org/2023.acl-long.504/. doi:10.18653/v1/2023.acl-long.504.

[20] B. Ding, C. Qin, R. Zhao, T. Luo, X. Li, G. Chen, W. Xia, J. Hu, A. T. Luu, S. Joty, Data augmentation using large language models: Data perspectives, learning paradigms and challenges, 2024. URL: https://arxiv.org/abs/2403.02990. arXiv:2403.02990.

[21] M. Zhou, PingAnLifeInsurance at SemEval-2023 task 10: Using multi-task learning to better detect online sexism, in: A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2188–2192. URL: https://aclanthology.org/2023.semeval-1.304. doi:10.18653/v1/2023.semeval-1.304.

[22] F. Aldabbas, S. Ashraf, R. Sifa, L. Flek, MultiProp framework: Ensemble models for enhanced cross-lingual propaganda detection in social media and news using data augmentation, text segmentation, and meta-learning, in: M. El-Haj (Ed.), Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script, Association for Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 7–22. URL: https://aclanthology.org/2025.abjadnlp-1.2/.

[23] H. Yan, L. Gui, G. Pergola, Y. He, Position bias mitigation: A knowledge-aware graph model for emotion cause extraction, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), 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, Online, 2021, pp. 3364–3375.

[24] G. Pergola, L. Gui, Y. He, A disentangled adversarial neural topic model for separating opinions from plots in user reviews, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 2870–2883.

[25] J. Lu, J. Li, B. Wallace, Y. He, G. Pergola, NapSS: Paragraph-level medical text simplification via narrative prompting and sentence-matching summarization, in: A. Vlachos, I. Augenstein (Eds.), Findings of the Association for Computational Linguistics: EACL 2023, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 1079–1091.

[26] OpenAI, Gpt-4o model card, https://openai.com/index/gpt-4o-system-card/, 2024. Accessed: today.

[27] 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, 2019. arXiv:1907.11692.

[28] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, 2021. arXiv:2006.03654.

[29] E. Amigo, A. Delgado, Evaluating extreme hierarchical multi-label classification, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 5809–5819.