Leveraging Model Confidence and Diversity: A Multi-Stage Framework for Sexism Detection*

Anwar Alajmi[1,2,*], Gabriele Pergola[1]

[1]Department of Computer Science, University of Warwick, Coventry CV4 7AL, UK
[2]Department of Computer Science, College of Business Studies, Public Authority of Applied Education and Training (PAAET), Kuwait

Abstract

Sexism on social media often manifests in implicit, sarcastic, or context-dependent forms, making it difficult to detect reliably with existing systems. What is perceived as sexist language can at times vary across individuals, cultures, and platforms, highlighting its inherently multifaceted nature. Inspired by this diversity of human interpretation, we propose a hybrid detection framework that integrates the outputs of multiple neural language models, each encoding different perspectives on the task. Our system combines fine-tuned monolingual transformers (e.g., BERTweet for English, RoBERTuito for Spanish) with instruction-tuned large language models (LLMs) such as Claude 3 Sonnet and LLaMA3-70B-Instruct. These models are combined within a confidence-based multi-stage pipeline: high-confidence predictions from task-specialized models are preserved, while uncertain instances are routed to general-purpose LLMs for zero-shot classification. This dynamic strategy combines high-confidence predictions from specialized models with broader judgments from instruction-tuned LLMs, enabling the system to better manage expressions of sexist language across different linguistic contexts. Evaluated on the EXIST 2025 shared task, our approach ranked 2nd on the Spanish test set, 4th overall, and 9th on English, demonstrating the effectiveness of confidence-guided ensemble learning in multilingual sexism detection.

Keywords

Sexism Detection, Tweets Classification, Multilingual Models, Large Language Models, Ensemble, EXIST 2025

1. Introduction

Sexism can be described as gender-based stereotyping or discrimination. Sexist language can appear in both explicit and implicit forms, ranging from overtly hateful comments to subtle sexist jokes. With the widespread use of social media, such expressions increasingly occur online, posing serious psychological and societal risks. As a result, automatic detection of sexist language has become a vital task in Natural Language Processing (NLP), aimed at fostering safer and more inclusive digital spaces. Despite notable progress in sexism detection using deep learning methods, several core challenges remain. Sexist content can be highly implicit, relying on contextual cues, sarcasm, or coded language, which makes it difficult for models to detect reliably. Moreover, due to differences in cultural, social, and individual norms, the interpretation of what constitutes sexist content can vary widely. These challenges underscore the need for detection systems that are not only accurate but also robust to linguistic variation and ambiguity.

To address these issues, this study proposes a hybrid sexism detection framework that integrates the outputs of diverse NLP models. Different neural language models, whether simply pre-trained, instruction-tuned, or trained with reinforcement learning from human feedback (RLHF), may already encode different assumptions and interpretations about what constitutes sexist language. These differences can arise from the datasets and training objectives used during their development, which implicitly shape the models' sensitivity to linguistic nuances and social cues. By aggregating the outputs of such diverse systems, our approach seeks to synthesize these pre-encoded interpretive frames to enhance detection robustness. Therefore, rather than relying on a single classifier, we introduce a new approach that leverages a confidence-based ensemble of fine-tuned monolingual transformers

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
anwar.alajmi@warwick.ac.uk (A. Alajmi); gabriele.pergola.1@warwick.ac.uk (G. Pergola)
0000-0003-4795-5915 (A. Alajmi); 0000-0002-7347-2522 (G. Pergola)
© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

CEUR-WS.org/Vol-4038/paper_137.pdf

(such as BERTweet [1] and RoBERTuito [2]) and large language models (LLMs) including Llama-3-70B-Instruct [3] and Claude 3 Sonnet [4]. We posit that each model is able to offer a distinct perspective on the task, contributing to a more comprehensive understanding of the data, and that combining high-confidence predictions from these models, the system improves reliability and generalization. Our main contributions are as follows:

- We introduce a robust framework for sexism detection that combines fine-tuned language-specific models and general-purpose large language models through a confidence-based ensemble method.
- We conduct a thorough experimental evaluation, including ablation studies and multiple ensemble combinations, to quantify the individual and combined impact of each model component.

This work was conducted as part of the EXIST (sEXism Identification in Social Networks) 2025 shared task [5], which evaluates systems for detecting sexist content in multilingual social media data under realistic conditions. Our system demonstrates competitive performance on the EXIST 2025 shared task, ranking 2nd on the Spanish test set, 4th overall, and 9th on the English portion, highlighting the value of synthesizing multiple model perspectives for detecting multi faceted classification tasks such as online sexism.

The rest of the paper is structured as follows: Section 2 reviews and summarizes related works in the field of sexism detection. Sections 3 explores the proposed methodology, while Section 4 discusses the experimental results. Finally, the conclusion and future work are presented in Section 5.

2. Related Work

Hateful and toxic speech detection has been an ongoing challenge in NLP. This area has received considerable attention from machine learning (ML) researchers, with early work employing traditional supervised ML techniques [6, 7, 8, 9, 10, 11, 12, 13]. More recent approaches have leveraged Transformer models [14], such as BERT, which has significantly improved toxicity classification performance by enabling deeper contextual understanding as demonstrated by the works in [15, 16, 17, 18, 19, 20, 21]. More recently, several methodologies focused on employing LLMs, and in particular prompt-based approaches, for various hate speech detection tasks, including sexism classification [22, 23]. For example, the study by [24] demonstrates that zero-shot sexism detection can be enhanced through expert-guided prompting, even when experts are not familiar with LLMs. Their findings emphasize the advantage of combining human knowledge with model capabilities to detect subtle sexist language.

An ongoing challenge in the detection of sexism and other forms of hateful speech is the presence of annotation disagreements, which often originate from social and cultural biases [25]. The EXIST shared task aims to tackle this issue by promoting systems capable of accurately identifying sexist content regardless of inconsistent labeling. Submissions to EXIST 2024 included a range of approaches, such as the utilization of BERT-based models (DistilBERT [26], DeBERTa [27],and RoBERTa [28]), LLMs [29], ensemble methods [30] and knowledge-based techniques [31]. One of the best performing systems was proposed by [32], where the authors used data augmentation to train DeBERTa with hard parameter sharing [33]. Their approach also incorporated annotators information and Round to Closed Value [34].

This work leverages a variety of classification systems to emulate annotators and models their disagreements. Moreover, it proposes a combination of monolingual and multilingual NLP models within a confidence-based ensemble technique to produce robust and accurate final predictions.

3. Methodology

In this section, we describe the classification framework in detail. The system follows a multi-stage ensemble pipeline that combines predictions from fine-tuned discriminative transformers and instruction-following large language models (LLMs). Each input tweet is first processed by one or more fine-tuned transformer models. These include language-specific models, such as BERTweet [1] for English, RoBERTuito [2]) for Spanish, as well as a multilingual model, LLaMA-3.2-1B [3], fine-tuned jointly on both

languages. If the model's confidence in its prediction exceeds a predefined threshold, the prediction is accepted directly and included in the final output. If confidence falls below this threshold, the input is rerouted to a secondary stage involving zero-shot classification by general-purpose LLMs (Claude 3 Sonnet [4] and Llama-3-70B-Instruct [3]). This dynamic routing mechanism allows the system to retain high-confidence transformer outputs while deferring uncertain cases to more general, instruction-tuned models. As shown in Figure 1, the final predictions consist of both retained transformer outputs and LLM-based decisions, combined to enhance robustness and coverage across a wide range of linguistic inputs.

3.1. Confidence-based Routing and Decision Integration

The central mechanism driving our ensemble strategy is the use of model confidence to control routing and the final predictions. Rather than aggregating predictions uniformly, we adopt a selective approach that promotes reliable outputs from specilised models, while using LLMs as fallback experts in cases of uncertainty.

For each fine-tuned model, we compute confidence scores based on the logit margin: the difference between the top two predicted class logits. A threshold τ is determined individually for each model using its training set distribution of logit margins. In our experiments, the threshold typically falls between the 25th and 35th percentiles, providing a conservative cutoff for what constitutes a confident prediction. Predictions with logit margins above this threshold are retained as-is. Instances that fall below the threshold are rerouted to the LLM layer. These tweets are classified in a zero-shot manner using both Claude 3 Sonnet and Llama-3-70B-Instruct, prompted with the following instruction: "`You are an accurate sexism classification system. Classify the given tweets as sexist (YES) or not sexist (NO).`". The final decision for each instance is determined as follows: if both LLMs return the same label, that label is used. If they disagree, we select the output of Claude 3 Sonnet, which demonstrated higher reliability in our preliminary experiments on the development set.

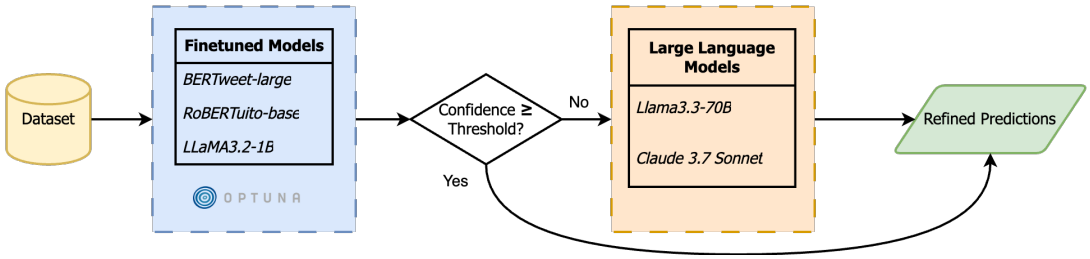

Figure 1: Confidence-based ensemble strategy for combining LLM and fine-tuned model predictions based on confidence.

4. Experimental Results & Analysis

This section discusses the dataset, the experimental set up, the evaluation metrics, and the results of the proposed methods.

4.1. Dataset

The EXIST 2025 Tweets dataset was used to classify sexist content. It contains 10,034 samples labeled 'YES' or 'NO' in both English and Spanish, distributed across training, development, and test sets, as shown in Table 1.

Table 1
EXIST 2025 Tweets Dataset information.

Set	English	Spanish	All
Training	3,260	3,660	6,920
Development	489	549	1,038
Test	978	1,098	2,076

4.2. Baselines

We compare our proposed approach against several strong baselines commonly used in multilingual and monolingual text classification tasks:

- **XLM-RoBERTa-large**: A multilingual transformer model introduced by [35], that is pre-trained on 100 languages and acts as a strong fine-tuned baseline for multilingual text classification tasks.
- **LLaMA3.2-1B**: A 1B-parameter model that is fine-tuned for sexism classification of tweets.
- **Merged Monolingual Transformers (MMT)**: This setup combines BERTweet-large solely fine-tuned on English data, and RoBERTuito-base exclusively trained on Spanish samples for the detection of sexist tweets.

4.3. Experimental Setting

The baseline models are fine-tuned on the entire training set of 6,920 samples. The hyperparameter settings were 0.1 for the `warmup_ratio` , 4 for both `per_device_train_batch_size` and `per_device_eval_batch_size`, 3 for `num_train_epochs`, and `2e-5` for learning rate. The pre-trained versions of the baseline models were obtained from Huggingface [36]. For LLM-related experiments, LangChain [37] and Ollama [38] were used to download and run the models locally. Hyperparameter optimization was conducted during the training phase using Optuna [39], a framework for efficiently searching optimal training configurations for each model. We additionally employed a data augmentation strategy, where Claude 3.7 Sonnet [4] was used to generate additional balanced synthetic examples in English and Spanish. Using the data augmentation prompt: `"Generate a paraphrased (augmented) version of the tweet that preserves its original meaning, tone, and any sexist or non-sexist elements"`, additional 1,000 English and 1,000 Spanish tweets (for a total of 2,000 augmented examples) were appended to the original training dataset.

4.4. Evaluation Metrics

The following are the official metrics used to evaluate the system under the EXIST 2025 **hard–hard** evaluation scheme:

1. *ICM*: The Inter-Consistency Measure (ICM) [40] as shown in Equation 1 measures the alignment between two sets of predictions, A and B.

$$\mathrm{ICM}(A, B) = \alpha_1 \,\mathrm{IC}(A) + \alpha_2 \,\mathrm{IC}(B) - \beta \,\mathrm{IC}(A \cup B) \tag{1}$$

2. *ICM-Norm*: A normalized ICM score that is rescaled to the $[0, 1]$ range.
3. *F1*: The harmonic mean of precision and recall as shown in Equation 2.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{2}$$

4.5. Results

4.5.1. Development Set

Table 2 demonstrates the performance of the proposed techniques using the development set of EXIST 2025 Tweets dataset for both languages on Task 1.1. The baseline MMT achieves an ICM score of 0.5439 and a normalized ICM score of 0.7720, which outperforms LLaMA3.2-1B and XLM-RoBERTa-large models. This shows that language-specific models are more efficient than multilingual ones on this kind of classification tasks. However, LLaMA3.2-1B obtains the highest F1 score among the baseline models. Instruction-tuned large language models (LLMs) outperform fine-tuned baselines even without task-specific adaptation. Claude 3 Sonnet and LLaMA-3-70B-Instruct show strong zero-shot performance, with Claude achieving the best standalone results across all metrics (ICM: 0.6098, ICM Norm: 0.8050, F1: 0.8701). This suggests that LLMs encode broadly useful social and pragmatic knowledge that supports robust generalization, particularly for subtle or implicit forms of sexism.

The next set of experiments in Table 2 illustrates the impact of confidence-based ensemble combinations on enhancing the performance. We label predictions as *strong* or *weak* depending on whether the model's confidence exceeds a model-specific threshold. Strong predictions are retained; weak predictions are rerouted to a fallback model. In particular, predictions from MMT are divided into strong (above the confidence threshold) and weak (below the threshold). In the MMT (strong) + LLaMA3.2-1B (weak) setup, replacing weak predictions with those from LLaMA3.2-1B leads to improvements of 9.93% in ICM, 3.51% in ICM-Norm, and 2.13% in F1. When LLaMA3.3-70B is used for replacement, the gains increase to 14.73% for ICM, 5.19% for ICM-Norm, and 3.15% for F1. Similarly, combining MMT with Claude 3.7 yields improvements of 14.47%, 5.1%, and 3.11% in ICM, ICM-Norm, and F1, respectively.

This approach is also applied to LLaMA3.2-1B, where strong predictions are retained and weak ones are replaced by those from LLaMA3.3-70B, resulting in a 14.03% increase in ICM and a 4.95% increase in ICM-Norm. When Claude 3.7 is used instead, the improvements reach 20.44% in ICM and 7.21% in ICM-Norm.

The last three results in Table 2 reflect the performance of the approach where the predictions of the LLMs and the fine-tuned models are merged. The ensemble of Claude 3.7, LLaMA3.3-70B, and MMT improves the performance of the baseline MMT by 13.99% in ICM, 4.94% in normalized ICM, and 3.01% in F1 score. Similarly, the ensemble of Claude 3.7, LLaMA3.3-70B, and LLaMA3.2-1B outperforms the standalone LLaMA3.2-1B by 14.45% in ICM and 5.11% in normalized ICM. The full confidence-based ensemble, combining Claude 3.7, LLaMA3.3-70B, MMT, and LLaMA3.2-1B, achieves an F1 score of 0.8799, representing a 3.92% improvement over the baseline LLaMA3.2-1B, along with gains of 18.65% in ICM and 6.57% in normalized ICM.

4.5.2. Test Set (Leaderboard)

The three submitted runs for the EXIST 2025 competition Subtask 1.1 (hard setting) were as follows:

- *warwick_1*: Represents the full confidence-based ensembling approach combining high-confidence predictions from Claude 3 Sonnet and LLaMA-3-70B-Instruct with the predictions of MMT and LLaMA-3-1B (*Claude 3 Sonnet and LLaMA-3-70B-Instruct (s) + MMT and LLaMA-3-1B (w)*).
- *warwick_2*: Combines the high-confidence predictions of MMT with fallback predictions from Claude 3 Sonnet (*MMT (s) + Claude 3 Sonnet (w)*).
- *warwick_3*: Uses confident predictions from LLaMA-3-1B and replaces low-confidence outputs with those from Claude 3 Sonnet (*LLaMA-3-1B (s) + Claude 3 Sonnet (w)*).

Across all instances (English and Spanish), *warwick_1* ranked 4th, achieving an ICM score of 0.6249, a normalized ICM of 0.8141, and an F1 score of 0.7991. *warwick_2* ranked 7th (ICM: 0.5834, ICM Norm: 0.7932, F1: 0.7892), while *warwick_3* ranked 12th out of 160 submissions, scoring 0.5793 on ICM, 0.7912 on ICM Norm, and 0.7888 on F1.

For Spanish tweets, *warwick_1* ranked 2nd, with an ICM of 0.6441. *warwick_2* followed in 5th, scoring 0.6312 on ICM. *warwick_3* ranked 6th, with ICM: 0.6126.

Baseline Models	**ICM**	**ICM Norm**	**F1**
XLM-RoBERTa-large	0.5161	0.7581	0.8387
Llama3.2-1B	0.5390	0.7693	0.8467
MMT	0.5439	0.7720	0.8032
LLMs	**ICM**	**ICM Norm**	**F1**
LLaMA-3-70B-Instruct	0.5983	0.7993	0.8661
Claude 3 Sonnet	0.6098	0.8050	0.8701
Confidence-based Ensemble Combinations	**ICM**	**ICM Norm**	**F1**
MMT (s) + LLaMA-3-1B (w)	0.5979	0.7991	0.8203
MMT (s) + LLaMA-3-70B-Instruct (w)	0.6240	0.8121	0.8285
MMT (s) + Claude 3 Sonnet (w)	0.6226	0.8114	0.8282
LLaMA-3-1B (s) + LLaMA-3-70B-Instruct (w)	0.6146	0.8074	0.8255
LLaMA-3-1B (s) + Claude 3 Sonnet (w)	**0.6491**	**0.8247**	0.8364
Claude 3 Sonnet + LLaMA-3-70B-Instruct (s) + MMT (w)	0.6200	0.8101	0.8274
Claude 3 Sonnet + LLaMA-3-70B-Instruct (s) + LLaMA-3-1B (w)	0.6169	0.8086	0.8263
Claude 3 Sonnet + LLaMA-3-70B-Instruct (s) + MMT + LLaMA-3-1B (w)	0.6394	0.8198	**0.8799**

Table 2
Experimental results on the EXIST 2025 Tweets Dataset (development set). MMT refers to the Merged Monolingual Transformers configuration, which combines BERTweet for English and RoBERTuito for Spanish. "*Strong*" predictions, indicated with '(s)' refer to instances where the model's confidence exceeds the predefined threshold and are retained. "*Weak*" predictions, indicated with '(w)', are those below the threshold and are rerouted to another model for classification.

Submitted Runs	**ICM**	**ICM Norm**	**F1**	**Rank**
warwick_1 (All instances)	**0.6249**	**0.8141**	**0.7991**	**4th/160**
warwick_2 (All instances)	0.5834	0.7932	0.7892	7th/160
warwick_3 (All instances)	0.5793	0.7912	0.7888	12th/160
warwick_1 (Spanish)	**0.6441**	**0.8221**	**0.8255**	**2nd/153**
warwick_2 (Spanish)	0.6312	0.8157	0.8206	5th/153
warwick_3 (Spanish)	0.6126	0.8064	0.8179	6th/153
warwick_1 (English)	**0.5869**	**0.7995**	**0.7634**	**9th/158**
warwick_2 (English)	0.5078	0.7591	0.7439	55th/158
warwick_3 (English)	0.5261	0.7685	0.7497	43rd/158

Table 3
Experimental results on the EXIST 2025 Subtask 1.1 Leaderboard (test set).

In contrast, performance on English samples was lower, with *warwick_1* ranking 9th, *warwick_2* ranking 55th, and *warwick_3* ranking 43rd. Nevertheless, the top performing submissions on the English test set achieved lower scores compared to their Spanish counterparts. This performance gap may be influenced by several factors. One possibility is the proportion and frequency of forms of sarcasm, humor, or subtle implication, which can be harder to detect. Another potential factor is variation in annotation consistency across languages, though further analysis would be needed to confirm this.

The leaderboard results demonstrate the effectiveness of confidence-based ensembling, particularly when integrating predictions from both LLMs and fine-tuned models. Among the submissions, *warwick_1*, which employs the full confidence-based ensemble, consistently outperforms the other approaches. These findings also highlight the advantages of fine-tuning monolingual models separately for each language. Specifically, the ensemble approach using BERTweet and RoBERTuito in the MMT setup (*warwick_2*) outperforms the LLaMA3.2-1B model fine-tuned on both languages (*warwick_3*) on the Spanish test set (ranked 5th) and across all instances (ranked 7th).

5. Conclusion

This work introduced a multi-stage, confidence-based ensemble framework for sexism detection in multilingual social media data. The proposed pipeline dynamically integrates the outputs of fine-tuned monolingual and multilingual transformer models with zero-shot predictions from instruction-tuned LLMs, using model confidence to guide the decision process. This mechanism enabled the system to balance precision and generalization by leveraging each model type where they are most reliable. Through extensive evaluation on the EXIST 2025 shared task, our system demonstrated competitive performance, ranking 2nd on the Spanish test set and 4th overall. The experimental results confirm that the confidence-based routing not only enhances robustness to ambiguous or borderline cases, but also improves performance over individual models and naïve ensembles. A promising direction for future research is exploring additional prompting techniques, for example, based on multi-expert prompting [41] to model annotators as experts and improve the reliability of the predictions.

Declaration on Generative AI

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

References

[1] D. Q. Nguyen, T. Vu, A.-T. Nguyen, Bertweet: A pre-trained language model for english tweets, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 9–14.

[2] J. M. Pérez, D. A. Furman, L. A. Alemany, F. M. Luque, Robertuito: a pre-trained language model for social media text in spanish, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 7235–7243.

[3] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).

[4] Anthropic, Claude 3.7 sonnet: a hybrid reasoning model with adjustable "thinking" mode, https://www.anthropic.com/news/claude-3-7-sonnet, 2025. (accessed June 2025).

[5] 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), 2025.

[6] T. L. Sutejo, D. P. Lestari, Indonesia hate speech detection using deep learning, in: 2018 International Conference on Asian Language Processing (IALP), IEEE, 2018, pp. 39–43.

[7] T. Gröndahl, L. Pajola, M. Juuti, M. Conti, N. Asokan, All you need is" love" evading hate speech detection, in: Proceedings of the 11th ACM workshop on artificial intelligence and security, 2018, pp. 2–12.

[8] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, O. Frieder, Hate speech detection: Challenges and solutions, PloS one 14 (2019) e0221152.

[9] G. Pergola, L. Gui, Y. He, Tdam: A Topic-Dependent Attention Model for Sentiment Analysis, Information Processing & Management 56 (2019) 102084.

[10] N. Albadi, M. Kurdi, S. Mishra, Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere, in: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2018, pp. 69–76.

[11] 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.
[12] A. S. Saksesi, M. Nasrun, C. Setianingsih, Analysis text of hate speech detection using recurrent neural network, in: 2018 international conference on control, electronics, renewable energy and communications (ICCEREC), IEEE, 2018, pp. 242–248.
[13] 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.
[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
[15] M. Mozafari, R. Farahbakhsh, N. Crespi, A bert-based transfer learning approach for hate speech detection in online social media, in: International conference on complex networks and their applications, Springer, 2019, pp. 928–940.
[16] R. T. Mutanga, N. Naicker, O. O. Olugbara, Hate speech detection in twitter using transformer methods, International Journal of Advanced Computer Science and Applications 11 (2020).
[17] 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.
[18] F. M. Plaza-del Arco, M. D. Molina-González, L. A. Urena-López, M. T. Martín-Valdivia, Comparing pre-trained language models for spanish hate speech detection, Expert Systems with Applications 166 (2021) 114120.
[19] 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. URL: https://aclanthology.org/2024.tsar-1.10/.
[20] Anjum, R. Katarya, Hate speech, toxicity detection in online social media: a recent survey of state of the art and opportunities, International Journal of Information Security 23 (2024) 577–608.
[21] S. Khan, G. Pergola, A. Jhumka, Multilingual sexism identification via fusion of large language models, in: Conference and Labs of the Evaluation Forum (CLEF 2024), 2024.
[22] 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.
[23] 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, pp. 361–382.
[24] M. Reuver, I. Sen, M. Melis, G. Lapesa, Tell me what you know about sexism: Expert-llm interaction strategies and co-created definitions for zero-shot sexism detection, in: Findings of the Association for Computational Linguistics: NAACL 2025, 2025, pp. 8438–8467.
[25] A. M. Davani, V. Prabhakaran, M. Diaz, Dealing with disagreements: Looking beyond the majority vote in subjective annotations, in: ACL, 2022.
[26] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019).
[27] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, arXiv preprint arXiv:2006.03654 (2020).

[28] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
[29] J. Tavarez-Rodríguez, F. Sánchez-Vega, A. Rosales-Pérez, A. P. López-Monroy, Better together: Llm and neural classification transformers to detect sexism, Working Notes of CLEF (2024).
[30] S. Khan, G. Pergola, A. Jhumka, Multilingual sexism identification via fusion of large language models, Working Notes of CLEF (2024).
[31] 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 tweets and memes, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2024, pp. 93–117.
[32] 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).
[33] S. Ruder, An overview of multi-task learning in deep neural networks, arXiv preprint arXiv:1706.05098 (2017).
[34] A. F. M. de Paula, G. Rizzi, E. Fersini, D. Spina, Ai-upv at exist 2023–sexism characterization using large language models under the learning with disagreements regime, arXiv preprint arXiv:2307.03385 (2023).
[35] 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, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.
[36] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, Hugging face: The ai community building the future., https://huggingface.co, 2020. Accessed: 2025-06-02.
[37] H. Chase, Langchain: Building applications with llms through composability, https://www.langchain.com/, 2022. Accessed: 2025-06-02.
[38] O. Team, Ollama: Run and deploy large language models locally, https://ollama.com/, 2023. Accessed: 2025-06-02.
[39] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 2623–2631. URL: https://doi.org/10.1145/3292500.3330701. doi:10.1145/3292500.3330701.
[40] E. Amigo, A. Delgado, Evaluating extreme hierarchical multi-label classification, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 5809–5819.
[41] D. X. Long, D. N. Yen, A. T. Luu, K. Kawaguchi, M.-Y. Kan, N. F. Chen, Multi-expert prompting improves reliability, safety, and usefulness of large language models, arXiv preprint arXiv:2411.00492 (2024).