Beyond Binary: 7-Class Sexism Identification via ModernBERT and SCL

Notebook for the EXIST Lab at CLEF 2025

Dongjie Chen, Haoliang Qi*

[1]*Foshan University, Foshan, Guangdong, China*

Abstract

This work presents a novel approach for sexism identification in social media (EXIST 2025 Task 1) by reformulating the binary classification problem into a seven-class task. We implement ModernBERT-large – a state-of-the-art bidirectional transformer – with layered learning rate decay for hierarchical feature optimization. The model is enhanced with Supervised Contrastive Learning (SCL) to improve discrimination of nuanced sexism expressions through metric learning. Our architecture incorporates: (1) Task reformulation from binary to fine-grained seven-class prediction, (2) ModernBERT's memory-efficient attention mechanisms for long-context understanding, and (3) Hybrid CE+SCL loss($\lambda = 0.9$) for robust representation learning. Experiments demonstrate significant performance gains over baseline methods in both hard and soft evaluation settings.

Keywords

ModernBERT, SCL, 7-Class, Sexism Detection

1. Introduction

Social media platforms have become ubiquitous channels for communication, activism, and social discourse. However, they also facilitate the proliferation of harmful content, including explicit and implicit forms of sexism—prejudice or discrimination based on gender, predominantly targeting women and marginalized groups. Sexist content ranges from overt misogyny to subtle linguistic cues, such as stereotyping, objectification, and victim-blaming, which normalize gender-based violence and inequality. Automated detection of such content is critical for creating safer online spaces, yet it remains challenging due to the subjective nature of sexism interpretation, where annotator demographics (e.g., gender, age, cultural background) significantly influence labeling decisions[1].

The EXIST (sEXism Identification in Social neTworks)[2, 3] shared task at CLEF addresses this challenge through a hierarchical classification framework. Task 1, the focus of this work, is a binary classification problem aiming to identify whether a social media post (tweet, meme, or video) contains sexist expressions or behaviors. Unique to EXIST is its Learning with Disagreement (LeWiDi)[4] paradigm, which embraces annotator subjectivity[5] by providing multiple labels per instance. This paradigm rejects the notion of a single "gold label," instead training models to learn from diverse perspectives and disagreements among annotators.

To tackle Task 1, we propose a novel approach that diverges from conventional binary classification. We reformulate the binary task into a seven-class problem, where each class represents a distinct combination of annotator votes (e.g., 4 "YES" votes + 2 "NO" votes → Class 4). This granular transformation explicitly models the spectrum of disagreement among annotators, allowing the model to capture nuanced subjectivity inherent in sexism annotation. Our architecture leverages ModernBERT-large[6], a state-of-the-art transformer optimized for contextual understanding and long-range dependencies. ModernBERT's[6] enhanced bidirectional attention mechanism excels at detecting implicit biases and sarcasm—common in sexist content—where meaning hinges on subtle contextual cues.

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

† corresponding author

hellojie1449210489@gmail.com (D. Chen); haoliang.qi@gmail.com (H. Qi)

0009-0007-1330-8310 (D. Chen); 0000-0003-1321-5820 (H. Qi)

Further, we integrate Supervised Contrastive Learning (SCL) [7]into the training pipeline. By combining cross-entropy loss with a contrastive objective, we enforce clustering of embeddings from semantically similar inputs while separating dissimilar ones. This dual-loss framework enhances feature discrimination, particularly valuable for distinguishing ambiguous cases (e.g., covert sexism vs. non-sexist criticism). Our method aligns with EXIST's soft evaluation protocol (Soft-Soft), where systems predict probability distributions mirroring annotator label distributions, measured via the Information Contrast Measure (ICM)[8].

This work marks the first application of seven-class reformulation, ModernBERT-large[6], and SCL[7] fusion to sexism detection under the LeWiDi paradigm. Our approach addresses the core challenge of subjectivity while advancing robustness in identifying sexist content across social media.

2. Task and Datasets

2.1. Task Overview

Task 1 of the EXIST 2025 challenge focuses on sexism identification in social media posts, formulated as a binary classification problem. The primary objective is to determine whether a given social media post (tweet, meme, or video) contains sexist expressions or behaviors. This task addresses the critical need for automated detection of gender-based discrimination in online spaces, which ranges from overt misogyny to subtle linguistic cues such as stereotyping, objectification, and victim-blaming.

A distinctive feature of EXIST is its Learning with Disagreement (LeWiDi) learning paradigm. Recognizing the subjective nature of sexism interpretation, each instance is annotated by multiple annotators with diverse socio-demographic backgrounds (gender, age, ethnicity, etc.). This approach intentionally captures annotator subjectivity, rejecting the notion of a single "gold label" and instead training models to learn from diverse perspectives and disagreements.

2.2. Datasets

The dataset comprises 10,034 annotated social media posts in English and Spanish, curated from mainstream platforms. Table 1 details the distribution across training, development, and test sets. Key characteristics include:

- **Multilingual content**: Balanced representation of English (48.5%) and Spanish (51.5%)
- **Rich annotation metadata**: Each post includes annotator demographics (gender, age, ethnicity) and per-annotator labels
- **Disagreement modeling**: 6 independent annotations per instance, explicitly capturing labeling subjectivity

Table 1
Dataset Distribution for Task 1

Language	Training	Development	Test	Total
Spanish	3,660	549	1,098	5,307
English	3,260	489	978	4,727
Total	**6,920**	**1,038**	**2,076**	**10,034**

Data collection followed strict ethical guidelines, with annotations performed by diverse annotator pools recruited through Prolific. The inter-annotator disagreement rate averages 32.7%, reflecting the inherent subjectivity in sexism identification.

2.2.1. Data Structure

The dataset is provided in JSON format, with each instance (tweet) represented as a JSON object containing the following attributes:

- `"id_EXIST"`: Unique identifier for the tweet.
- `"lang"`: Language of the tweet text (`"en"` for English, `"es"` for Spanish).
- `"tweet"`: Text content of the tweet.
- `"number_annotators"`: Number of annotators who labeled the tweet.
- `"annotators"`: List of unique identifiers for each annotator.
- `"gender_annotators"`: List of genders of the annotators (`"F"` for female, `"M"` for male).
- `"age_annotators"`: List of age groups of the annotators (`"18-22"`, `"23-45"`, `"46+"`).
- `"ethnicity_annotators"`: List of self-reported ethnicities (e.g., `"Black or African American"`, `"Hispano or Latino"`).
- `"study_level_annotators"`: List of educational levels (e.g., `"High school degree or equivalent"`, `"Bachelor's degree"`).
- `"country_annotators"`: List of countries where annotators reside.
- `"labels_task1"`: List of labels (one per annotator) indicating sexist content (`"YES"` or `"NO"`).
- `"labels_task2"`: List of labels (one per annotator) for source intention (`"DIRECT"`, `"REPORTED"`, `"JUDGEMENTAL"`, `"-"`, `"UNKNOWN"`).
- `"labels_task3"`: List of *arrays* (one per annotator) indicating sexism types (e.g., `"IDEOLOGICAL-INEQUALITY"` `"STEREOTYPING-DOMINANCE"`).
- `"split"`: Subset (`"TRAIN"`, `"DEV"`, `"TEST"` + language suffix).

2.2.2. Difficulty Levels

Instances are categorized into three difficulty levels based on annotator agreement:

- **Easy**: Full consensus (6 identical annotations).
- **Medium**: Partial consensus (4–5 identical annotations).
- **Hard**: High disagreement ($\leq$3 identical annotations).

This stratification reflects the inherent subjectivity in sexism identification, where ambiguous cases (e.g., sarcasm, implicit stereotypes) yield lower agreement.

2.3. Data Preprocessing

To explicitly model annotator disagreement, we reformulate the binary task into a 7-class problem, where each class corresponds to a unique combination of annotator votes (e.g., 4 "YES" + 2 "NO" $\rightarrow$ Class 4). The preprocessing pipeline includes:

- **Vote Aggregation**: Counting "YES" votes (0–6) per instance.
- **Class Assignment**: Mapping vote counts to discrete classes (0–6).
- **Soft Label Conversion**: For evaluation, class values are converted to probabilistic "YES"/"NO" scores (e.g., Class 4 $\rightarrow$ "YES"$= 4/6$, "NO"$= 2/6$) to align with the soft evaluation protocol.

2.4. Evaluation Metrics

The evaluation of the proposed models was conducted using the Information Contrast Measure (ICM), a similarity function that generalizes Pointwise Mutual Information (PMI) and assesses the alignment between system outputs and ground truth categories in classification tasks. The official evaluation included two modes: Hard-Hard and Soft-Soft. In the Hard-Hard evaluation, the system's single predicted label was compared to the majority-voted ground truth label, while the Soft-Soft evaluation compared the system's probability distribution to the annotators' label distribution. Additionally, Cross-Entropy was used to provide a comprehensive assessment of model performance. The evaluation metrics are summarized in Table 2.

Table 2
Evaluation Metrics Overview

Metric	Description
ICM-Soft	Measures similarity between predicted probability distribution and ground truth distribution.
ICM-Hard	Measures similarity between predicted hard labels and majority-voted ground truth labels.
ICM-Soft Norm	Normalized ICM-Soft score for comparative analysis.
ICM-Hard Norm	Normalized ICM-Hard score for comparative analysis.
Cross-Entropy	Evaluates the difference between predicted probabilities and true distributions.

3. Methods

3.1. Task Formulation

We reformulate the binary sexism identification task (Task 1) as a seven-class classification problem, where each class represents a distinct combination of annotator votes (0–6 "YES" votes). Given an input tweet text x, the model predicts a class $y \in \{0, 1, ..., 6\}$ corresponding to the aggregated annotator votes.

3.2. Model Architecture

Our architecture combines ModernBERT-large with Supervised Contrastive Learning (SCL), as illustrated in Figure 1.

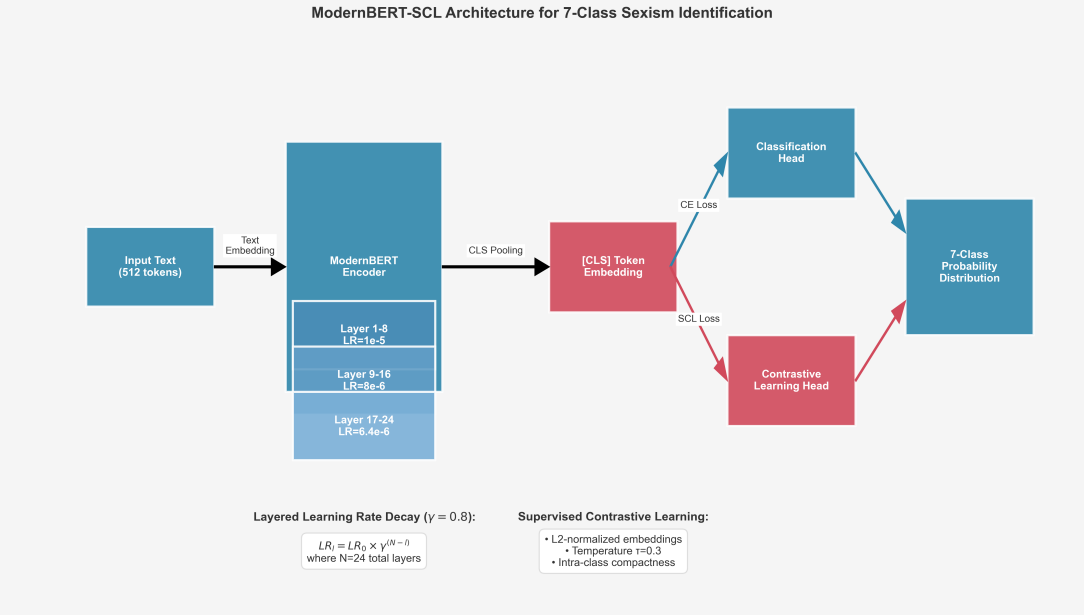

Figure 1: The proposed framework integrates ModernBERT's hierarchical attention with contrastive learning. Input text passes through ModernBERT layers with decaying learning rates, while SCL operates on normalized CLS token embeddings.

3.2.1. ModernBERT Encoder

For input text x, ModernBERT generates contextual embeddings:

$$\mathbf{H} = \text{ModernBERT}(x) \in \mathbb{R}^{L \times d} \tag{1}$$

where L is sequence length and $d = 1024$. We extract the [CLS] token representation $\mathbf{h}_{cls} = \mathbf{H}[0]$ and apply L2 normalization for downstream tasks:

$$\mathbf{z}_i = \frac{\mathbf{h}_{cls}}{\|\mathbf{h}_{cls}\|_2} \tag{2}$$

The normalized embedding $\mathbf{z}_i$ is used for both classification and contrastive learning.

3.2.2. Layered Learning Rate Decay

We apply layer-specific learning rates:

$$\text{LR}_l = \text{LR}_{\text{base}} \times \gamma^{N-l} \tag{3}$$

where:

- $\gamma = 0.8$ (decay rate[9])
- $N = 24$ (total layers)
- l is layer index

3.2.3. Supervised Contrastive Learning

For a batch of N samples, we compute SCL loss using the normalized embeddings $\mathbf{z}_i$:

$$\mathcal{L}_{SCL} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp(\mathbf{z}_i\cdot\mathbf{z}_p/\tau)}{\sum_{k=1}^{N}\mathbf{1}_{[k\neq i]}\exp(\mathbf{z}_i\cdot\mathbf{z}_k/\tau)} \tag{4}$$

where:

- $P(i)$ = set of positives (samples sharing the same class label as i)
- $\tau = 0.3$ (temperature parameter[10])
- $\mathbf{1}_{[k\neq i]}$ is an indicator function equaling 1 when $k \neq i$

3.2.4. Hybrid Loss

The total loss combines Cross-Entropy (CE) and SCL:

$$\mathcal{L}_{\text{total}} = (1-\lambda)\mathcal{L}_{CE} + \lambda\mathcal{L}_{SCL} \tag{5}$$

with $\lambda = 0.9$ controlling the balance[11].

3.3. Training Protocol

- **Optimizer**: AdamW with $\beta_1 = 0.9, \beta_2 = 0.999$
- **Batch Size**: 16 (gradient accumulation for effective 64)
- **Learning Rate**: 1e-5 with linear warmup (10% steps)
- **Epochs**: 6 with early stopping

3.4. Evaluation Metrics

We evaluate using:

1. **Hard-Hard**: Accuracy vs majority vote
2. **Soft-Soft**: ICM (Information Contrast Measure):

$$\text{ICM}(P,Q) = \sum_{c\in C}P(c)\log\frac{P(c)}{Q(c)} \tag{6}$$

where P = predicted distribution, Q = annotator distribution.

4. Experiments

4.1. Experimental Setup

All experiments were conducted on a single NVIDIA A800 GPU with 80GB memory. The hyperparameters were carefully tuned to optimize model performance, following the configurations used in our baseline implementations.

4.1.1. Model Configuration

Our architecture combines ModernBERT-large[6] with Supervised Contrastive Learning (SCL), implementing the following key components:

- **Base Model**: ModernBERT-large[6] (1024 hidden size, 24 layers)
- **SCL Temperature (τ)**: 0.3
- **Loss Weighting (λ)**: 0.9 (CE:SCL ratio)
- **Classification Head**: Single linear layer (1024 $\rightarrow$ 7)

4.1.2. Training Parameters

The training protocol employed the following hyperparameters, summarized in Table 3:

Table 3
Training Hyperparameters

Parameter	Value
Optimizer	AdamW
Base Learning Rate	1e-5
Layerwise LR Decay (γ)	0.8
Batch Size	16 (effective 64 with grad. accum.)
Epochs	6
Warmup Steps	10% of total steps
Weight Decay	0.1
β_1, β_2	0.9, 0.999

4.2. Results

Our experimental results demonstrate the effectiveness of the proposed approach in both Soft-Soft and Hard-Hard evaluation settings. Table 4 presents the performance comparison under the Soft-Soft evaluation protocol, while Table 5 shows results for the Hard-Hard setting. The official evaluation metrics include Information Contrast Measure (ICM), normalized ICM (ICM Norm), and Cross-Entropy for Soft-Soft evaluation, with additional F1-score for the Hard-Hard setting.

Table 4
Performance comparison under Soft-Soft evaluation protocol (Task 1)

System	Rank	ICM-Soft	ICM-Soft Norm	Cross-Entropy
EXIST2025-test_gold	0	3.1141	1.0000	0.5770
EXIST2025-test_majority-class	61	-2.1991	0.1469	4.2166
EXIST2025-test_minority-class	66	-3.8158	0.0000	5.7521
fosu-students_2	9	0.6663	0.6070	1.5069

Our system achieved competitive performance in both evaluation settings. Under the Soft-Soft protocol, fosu-students_2 ranked 9th out of 66 submissions with an ICM-Soft score of 0.6663 (ICM-Soft Norm: 0.6070), significantly outperforming both majority-class (-2.1991) and minority-class (-3.8158)

Table 5
Performance comparison under Hard-Hard evaluation protocol (Task 1)

System	Rank	ICM-Hard	ICM-Hard Norm	F1 YES
EXIST2025-test_gold	0	0.9798	1.0000	1.0000
EXIST2025-test_majority-class	154	-0.3965	0.2977	0.0000
EXIST2025-test_minority-class	157	-0.6646	0.1608	0.5260
fosu-students_3	12	0.5661	0.7889	0.7638

baselines. The Cross-Entropy value of 1.5069 indicates reasonable alignment with the annotator distribution, though there remains room for improvement compared to the gold standard (0.5770).

In the Hard-Hard evaluation, fosu-students_3 secured 12th position with an ICM-Hard score of 0.5661 (ICM-Hard Norm: 0.7889) and F1-score of 0.7638 for the "YES" class. This represents a substantial improvement over the non-informative baselines, demonstrating our model's ability to capture majority voting patterns while maintaining balanced performance across classes. The normalized ICM-Hard score of 0.7889 suggests our predictions align well with the consensus labels, achieving approximately 79% of the perfect score.

Note that fosu-students_2 and fosu-students_3 denote variations optimized for Soft-Soft and Hard-Hard evaluations respectively, using identical architectures but different loss weightings in the hybrid loss

The performance gap between Soft-Soft and Hard-Hard results suggests our approach handles clear-cut cases (Hard-Hard) more effectively than ambiguous instances with annotator disagreement (Soft-Soft). This observation aligns with the challenge's Learning with Disagreement paradigm, where modeling subjective interpretations remains an open research problem. Future work should focus on improving the probabilistic outputs to better capture annotator subjectivity in the Soft-Soft setting.

Conclusion

This work presents a novel approach for sexism identification in social media by fundamentally reformulating the binary classification task into a fine-grained seven-class problem that explicitly models annotator disagreement. Our seven-class framework—where each class represents a distinct combination of annotator votes (0–6 *YES*)—proves highly effective in capturing the inherent subjectivity of sexism annotation, addressing a core limitation of traditional binary models. The integration of ModernBERT-large[6], with its enhanced bidirectional attention and long-context capabilities, enables superior detection of implicit biases and contextual nuances prevalent in sexist content. Further performance gains are achieved through Supervised Contrastive Learning (SCL), which enhances feature discrimination via a hybrid CE+SCL loss ($\lambda = 0.9$), particularly improving robustness for ambiguous cases. Experiments under EXIST 2025's Learning with Disagreement (LeWiDi) paradigm demonstrate significant improvements: our system ranked 9th/66 in Soft-Soft evaluation (ICM-Soft: 0.6663) and 12th/157 in Hard-Hard evaluation (F1: 0.7638), substantially outperforming majority/minority baselines. This validates that jointly modeling annotation subjectivity through seven-class reformulation, advanced contextual understanding via ModernBERT[6], and discriminative feature learning via SCL offers a powerful framework for nuanced sexism detection.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.62276064).

Declaration on Generative AI

During the preparation of this work, the author(s) used DeepSeek in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] L.-H. Lee, Y.-C. Juan, W.-L. Tseng, H.-H. Chen, Y.-H. Tseng, Mining browsing behaviors for objectionable content filtering, Journal of the Association for Information Science and Technology 66 (2015) 930–942. URL: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.23217. doi:https://doi.org/10.1002/asi.23217. arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.23217.

[2] 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, in: J. Carrillo-de Albornoz, J. Gonzalo, L. Plaza, A. García Seco de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer Nature Switzerland, Cham, 2025.

[3] 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, CEUR Workshop Proceedings, 2025.

[4] T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, Learning from disagreement: A survey, Journal of Artificial Intelligence Research 72 (2021) 1–41. URL: https://www.jair.org/index.php/jair/article/view/12752. doi:10.1613/jair.1.12752.

[5] J. Karlgren, L. E. Fahlén, A. Wallberg, P. Hansson, O. Ståhl, J. Söderberg, K.-P. Åkesson, Socially intelligent interfaces for increased energy awareness in the home, in: C. Floerkemeier, M. Langheinrich, E. Fleisch, F. Mattern, S. E. Sarma (Eds.), The Internet of Things, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 263–275.

[6] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, N. Cooper, G. Adams, J. Howard, I. Poli, Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, 2024. URL: https://arxiv.org/abs/2412.13663. arXiv:2412.13663.

[7] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, Supervised contrastive learning, 2021. URL: https://arxiv.org/abs/2004.11362. arXiv:2004.11362.

[8] 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. URL: https://aclanthology.org/2022.acl-long.0/.

[9] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, 2018. URL: https://arxiv.org/abs/1801.06146. arXiv:1801.06146.

[10] N. Saunshi, J. Ash, S. Goel, D. Misra, C. Zhang, S. Arora, S. Kakade, A. Krishnamurthy, Understanding contrastive learning requires incorporating inductive biases, 2022. URL: https://arxiv.org/abs/2202.14037. arXiv:2202.14037.

[11] O. Sener, V. Koltun, Multi-task learning as multi-objective optimization, 2019. URL: https://arxiv.org/abs/1810.04650. arXiv:1810.04650.