UMUTeam at EXIST 2025: Multimodal Transformer Architectures and Soft-Label Learning for Sexism Detection

Notebook for the EXIST Lab at CLEF 2025

Ronghao Pan[1], Tomás Bernal-Beltrán[1], José Antonio García-Díaz[1,*] and Rafael Valencia-García[1]

[1]*Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100, Spain*

Abstract

This paper presents the approaches developed by UMUTeam in the EXIST 2025 shared task at CLEF 2025. This task focuses on identifying and classifying sexist content on social media across three modalities: text (e.g., tweets), images (e.g., memes), and videos (e.g., TikToks). Our systems address all three subtasks—binary sexism detection, source intention classification, and sexism categorization—under both hard and soft evaluation strategies. We used multilingual Transformer-based models, including XLM-RoBERTa (base and large versions) for text, ViT for image features, and VideoMAE for video input. We applied specialized preprocessing and label handling for each modality. Soft-label learning was implemented using mean squared error (MSE) loss for subtasks 1 and 2, which involve binary and multiclass classification, respectively, and binary cross-entropy (BCE) loss for subtask 3, which is a multilabel classification problem. In all cases, annotator votes were transformed into probability distributions to capture label uncertainty, in line with the LeWiDi framework. For hard-label variants, discrete predictions were obtained by selecting the class or classes with the highest probability from the model's output during the evaluation stage. Out of 244 participating teams that submitted a total of 589 hard-label and 284 soft-label runs, our systems achieved competitive results, including top-10 rankings in several subtasks.

Keywords

Vision Language Model, Natural Language Processing, Video Language Model, Sexist Identification,

1. Introduction

Sexism refers to prejudice, stereotyping, or discrimination based on a person's sex, typically directed against women. According to the Oxford English Dictionary, sexism is defined as "prejudice, stereotyping, or discrimination, typically against women, on the basis of sex." Although significant progress has been made toward gender equality in recent decades, such as the partial closing of the wage gap, the growing presence of women in leadership roles, and the implementation of educational programs to prevent gender-based violence, inequality and discrimination persist, especially in digital spaces. However, these issues persist, especially in digital environments [1].

Social media, in particular, has become a double-edged space. On one hand, it empowers movements like #MeToo, #Time'sUp, and #8M, for gender justice and provides a platform for women to report experiences of abuse and discrimination. On the other, it enables the rapid spread and normalization of sexist discourse, ranging from overtly misogynistic attacks to more subtle, seemingly benign expressions of gender bias. These forms of sexism are especially challenging to detect and address due to their implicit nature and cultural variability.

In virtual environments, sexism can manifest in many ways. While overtly misogynistic discourse is more easily identifiable, a more subtle dimension tends to go unnoticed [2]. This phenomenon can

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

*Pan et al.

ronghao.pan@um.es (R. Pan); tomas.bernalb@um.es (T. Bernal-Beltrán); joseantonio.garcia8@um.es (J. A. García-Díaz); valencia@um.es (R. Valencia-García)

0009-0008-7317-7145 (R. Pan); 0009-0006-6971-1435 (T. Bernal-Beltrán); 0000-0002-3651-2660 (J. A. García-Díaz); 0000-0003-2457-1791 (R. Valencia-García)

manifest in explicit or subtle ways. Glick and Fiske (1996) [3] proposed the notion of ambivalent sexism, distinguishing between hostile sexism, which is openly negative, and benevolent sexism, which appears positive but reproduces gender inequalities by idealizing or infantilizing women [4].

In response to the growing need for effective automatic detection of these behaviors, the sEXism Identification in Social neTworks (EXIST) shared task series was launched in 2021. EXIST seeks to promote the development of NLP-based systems that can identify a broad spectrum of sexist expressions across multiple languages and social platforms. Through a series of shared tasks, researchers have been challenged to address not only explicit hate, but also the subtleties of implicit, normalized, and "benevolent" sexism.

The upcoming fifth edition of EXIST [5, 6], which is part of CLEF 2025, is a significant advancement because it introduces a multimodal challenge centered on TikTok videos. This edition reflects the increasingly complex and multimedia nature of online communication by integrating text, audio, and visual components. Sexist messages may be embedded not only in written language, but also in tone, imagery, and gesture. This builds on previous editions, such as EXIST 2024, which used memes to examine the visual transmission of sexism. This edition therefore expands the scope of analysis by introducing nine subtasks organized across three media formats: text (tweets), images (memes), and video (TikToks), in two languages (English and Spanish). For each format, participants are asked to address three main challenges:

1. **Sexism identification**: a binary classification task to determine whether content is sexist or not.
2. **Source intention detection**: a classification of the author's communicative intent (direct, reported, or judgmental).
3. **Sexism categorization**: a multilabel classification of the sexist message according to predefined categories such as ideological inequality, stereotyping, objectification, sexual violence, and misogyny.

This paper presents the approaches developed by UMUTeam to address the three main tasks defined in the EXIST 2025 shared task: sexism identification, source intention detection, and sexism categorization, across text, image, and video modalities. For text-based tasks, we employed fine-tuned versions of `XLM-RoBERTa-base` [7] and `XLM-RoBERTa-large` [7], adapting them for both classification and regression scenarios using soft-label training strategies.

For image-based subtasks (2.1, 2.2, and 2.3), we proposed a multimodal architecture that integrates visual embeddings from the `ViT-base-patch16` [8] model and textual embeddings from `XLM-RoBERTa` or `XLM-RoBERTa-large` using late fusion techniques, as in [9]. Each multimodal classifier was fine-tuned independently per subtask, with model variations using both base and large versions of the text encoder.

For the multi-label output of subtask 2.3, regression-based heads were used and loss functions such as mean squared error (MSE) were adopted to support soft-target annotations. Video-based tasks (3.1, 3.2, and 3.3) required the handling of temporal information and multimodal alignment. We adopted `VideoMAE-base` [10] models that were pre-trained for video classification. We extended these models to fit the three subtasks using probabilistic label handling, custom data preprocessing with frame extraction via `Decord` library, and tailored loss functions (binary cross-entropy or MSE).

2. Related works

The automatic detection of sexist language has been a growing concern within NLP, initially approached as a subtask of offensive language or hate speech detection. Early works such as [11] laid the foundation by manually annotating social media posts and employing traditional classifiers like SVMs using extra-linguistic features in conjunction with character n-grams for hate speech detection. Although effective in identifying explicit abuse, these methods lacked the sensitivity to detect more subtle and normalized forms of sexism, such as paternalistic or stereotypical expressions.

The development of transformer-based models, particularly BERT and XLM-RoBERTa, significantly improved the handling of linguistic context and multilingual text, leading to better detection of nuanced forms of gender bias. Shared tasks such as AMI [12] and EXIST [13, 14] have promoted fine-grained classification of sexist content, including distinctions based on author intention and thematic categories aligned with Glick and Fiske's (1996) framework of ambivalent sexism [3]. Sexism can be analyzed at multiple levels depending on the author's intent or the specific nature of the expression. In all editions of the EXIST shared task, sexist content is categorized using a predefined set of labels, including *Ideology and Inequality*, *Stereotyping and Dominance*, *Objectification*, *Sexual Violence*, and *Misogyny and Non-Sexual Violence*. This approach aligns with the framework proposed in SemEval 2023 Task 10 – Explainable Detection of Online Sexism (EDOS) [15], which introduced a three-level taxonomy designed to facilitate explainable and hierarchical classification: (1) binary sexism detection, (2) category-level classification, and (3) fine-grained subcategory identification. Similarly, EXIST adopts a structured perspective in which sexism is addressed across three dimensions: binary classification for detecting whether a tweet is sexist, multi-class classification for identifying the author's intention, and multi-label classification for categorizing the type of sexism expressed.

While the detection of sexism has traditionally been addressed from a textual perspective, recent research has increasingly explored visual and multimodal approaches. These approaches acknowledge that sexist content may appear not only in written language but also in images or in combinations of audio and visuals as video. Multimodal methods have shown improved performance compared to single-modality systems in identifying hate speech and misogyny, as demonstrated in recent studies [16, 17]. From a visual perspective, most previous work focused on detecting offensive, inappropriate, or pornographic content. However, the symbolic and implicit nature of sexist imagery, particularly in the form of memes, has received less attention. To address this limitation, the current edition of the EXIST shared task introduces a new subtask specifically aimed at detecting sexism in memes, broadening the scope beyond purely textual analysis.

One of the ongoing challenges in sexism detection lies in addressing the biases that may compromise the fairness and reliability of model performance. Prior studies, such as [18, 19], have explored this issue primarily through the lens of textual data. While these works have contributed valuable insights, many existing approaches still fall short in capturing the full complexity of sexism. This phenomenon is inherently multifaceted and shaped by cultural, social, and individual contexts. Consequently, detection systems that overlook these dimensions risk reinforcing narrow definitions of sexism and may fail to recognize subtle or context-specific manifestations.

The multimodal direction continues in 2025 with the EXIST shared task expanding into video content, reflecting a broader research trend. The growing prevalence of short-form videos on platforms such as TikTok, YouTube Shorts, and Instagram Reels has shifted part of the research focus toward video-based sexism detection. In these platforms, sexist messages are often conveyed through a combination of speech, text overlays, facial expressions, gestures, and audio. This has introduced new challenges that require models capable of interpreting temporal dynamics and cross-modal interactions. Studies such as [20] introduced the MuSeD dataset, a Spanish-language corpus of social media videos annotated for various forms of sexism, highlighting the need for culturally aware and multimodal systems. For this reason, EXIST 2025 proposes a unified benchmark that systematically addresses sexism detection across multiple modalities and languages. The task is organized into nine subtasks, distributed across three content types: text (tweets), images (memes), and videos (TikToks), and conducted in both English and Spanish. In addition, EXIST 2025 adopts the Learning With Disagreement (LeWiDi) annotation paradigm, which represents label distributions in order to reflect inter-annotator variability.

3. Dataset

Since 2021, the EXIST shared task series has focused on detecting and categorizing sexist content in social media, initially through textual analysis. This led to the creation of several annotated tweet corpora used across different editions to evaluate systems in binary sexist detection, intention classification,

and thematic categorization. With the growing relevance of visual and audiovisual content in online communication, the EXIST 2024 campaign expanded to include memes, and in this edition EXIST 2025 introduces a significant evolution by incorporating TikTok videos into the dataset. The TikTok dataset was collected using the Apify TikTok Hashtag Scraper tool, targeting hashtags potentially associated with sexist content. A manual selection process ensured semantic relevance and balance, resulting in the identification of 185 Spanish hashtags and 61 English hashtags. To avoid user bias and enhance generalization, a chronological and author-based partitioning strategy was applied by organizers. Therefore, the final dataset comprises more than 3,000 videos, with 2,524 used for training (1,524 in Spanish and 1,000 in English) and 674 for testing (304 in Spanish and 370 in English).

For the annotation process, the organizers considered two primary sociodemographic variables: gender and age range. Six crowd-sourced annotators, recruited via the Prolific platform, labeled each meme according to detailed guidelines developed by two gender studies experts. In addition to core demographic attributes, the annotators' education level, ethnicity, and country of residence were also collected. This demographic diversity was intended to minimize potential labeling bias arising from cultural and social differences. Aligned with the Learning With Disagreement (LeWiDi) paradigm, the annotation protocol does not assume the existence of a single, definitive interpretation for each instance. Instead of producing a single "gold" label per item, the dataset includes the full set of annotations provided by all annotators. This approach captures the natural variability of human judgment, particularly in subjective tasks such as sexism detection, and encourages the development of systems that can learn from diverse perspectives. All individual annotations are made available to participants to support research into soft-label modeling, robustness to disagreement, and fairness-aware system design.

Table 1 summarizes the number of instances available for each of the three tasks defined in EXIST 2025. For Task 1 (tweets), the organizers provided separate partitions for training, validation, and test sets. In contrast, for Task 2 (memes) and Task 3 (videos), only the training and test splits were officially released. Therefore, in order to perform model selection and hyperparameter tuning, we created custom validation sets by randomly sampling 20% of the training data for each of these two tasks. This ensured consistency across experiments while preserving the integrity of the test set. As shown in the table, Task 1 includes 10,034 instances in total, Task 2 includes 5,097, and Task 3 includes 3,198. These distributions reflect the multimodal and multilingual design of the dataset, covering textual, visual, and audiovisual content.

Table 1
Number of samples per split (train, validation, test) for each EXIST 2025 task.

	Task 1	**Task 2**	**Task 3**
Train	6,920	3,235	2,019
Val	1,038	809	505
Test	2,076	1,053	674
total	10,034	5,097	3,198

4. Methodology

This section describes the models, training strategies, and evaluation procedures applied to each of the three main tasks defined in EXIST 2025: Task 1 (tweets), Task 2 (memes), and Task 3 (videos).

To address the three tasks in EXIST 2025, we designed modular pipelines that share a common structure but are specialized for each modality. As shown in Figures 1, 2, and 3, all systems begin with preprocessing and label transformation stages that prepare data for either hard-label or soft-label training. For Task 1 (tweets), we used XLM-RoBERTa-base and large variants with a custom MSE-based regression trainer to model soft label distributions and hard label distributions. In Task 2 (memes), we adopted a late-fusion strategy that combines text and image embeddings from XLM-RoBERTa and ViT,

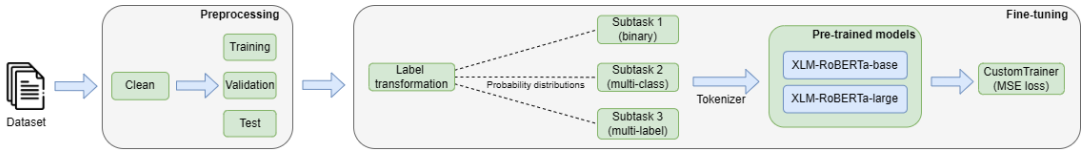

Figure 1: Pipeline overview for Task 1 (tweets), including preprocessing, subtask-specific label encoding, and soft-label fine-tuning with XLM-RoBERTa.

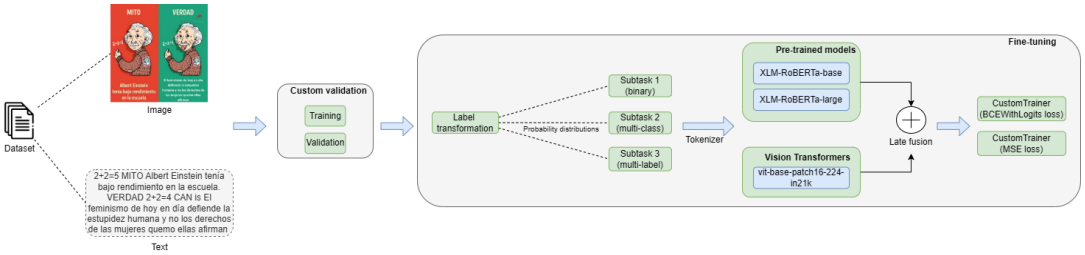

Figure 2: Pipeline overview for Task 2 (memes), showing multimodal fusion of text and image features via XLM-RoBERTa and ViT, with separate trainers for soft and hard label learning.

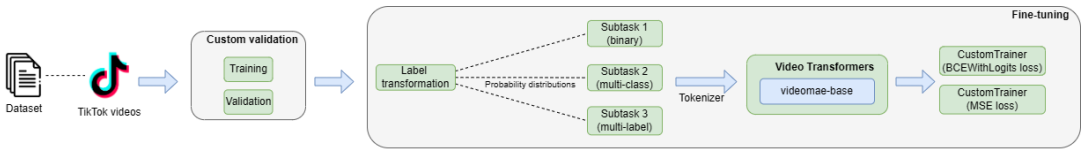

Figure 3: Pipeline overview for Task 3 (videos), based on VideoMAE for visual encoding and custom classification heads for each subtask.

respectively. For Task 3 (videos), we used VideoMAE to encode uniformly sampled frames, followed by custom classification heads.

4.1. Task 1

For Task 1, which focuses on tweet-based sexism detection, we developed a transformer-based classification system using two variants of XLM-RoBERTa (the base and large versions). Each model was fine-tuned separately for the three subtasks: (1.1) binary sexism identification, (1.2) source intention classification, and (1.3) sexism categorization. We adopted a probabilistic label encoding strategy that reflects the distribution of annotations across multiple annotators, enabling soft-label training under the LeWiDi paradigm.

Prior to tokenization, the raw tweet texts were preprocessed to remove noise and ensure consistency. This included stripping user mentions, hashtags, markdown artifacts, emojis, and extraneous whitespace. These steps aimed to standardize inputs across languages and reduce variance introduced by formatting. After preprocessing, the texts were tokenized using the corresponding XLM-RoBERTa tokenizer, with truncation to a maximum sequence length of 512 tokens.

For subtasks 1.2 and 1.3, which require multiclass and multilabel outputs, respectively, the labels were transformed into normalized probability vectors based on annotator agreement. A custom loss function based on mean squared error (MSE) was implemented to train the models to output probability distributions over the possible categories. In contrast, for the hard classification setting, we used categorical cross-entropy loss and standard F1 metrics for validation. All models were trained using the following hyperparameters: learning rate of 2e-6, weight decay of 0.01, batch size of 8, maximum

sequence length of 512 tokens, and 10 training epochs. In addition, the AdamW optimizer was used, and dropout regularization was applied as per the default configuration of the pre-trained models.

Evaluation was carried out using two complementary settings. In the hard-hard setup, models produced discrete class predictions that were compared against hard ground truth labels derived via threshold-based majority voting. For subtask 1.1, a label was accepted if at least four out of six annotators agreed. In subtask 1.2, the threshold was three votes, and in the multi-label subtask 1.3, each class was included if it received at least two votes. Instances without any class meeting the required threshold were excluded. Notably, in the hard classification setting for subtasks 1.2 and 1.3, the models were not trained to predict the "non-sexist" class explicitly. To address this limitation and avoid false positives, we implemented a two-stage pipeline. First, the best-performing model from subtask 1.1 (binary classification) was used as a filter to identify whether a tweet should be considered sexist or not. If the output of the 1.1 classifier indicated a non-sexist tweet (label "NO"), the system bypassed the 1.2 and 1.3 classifiers and directly assigned a label "NO" prediction. Only tweets predicted as sexist by the 1.1 model were passed to the corresponding models for 1.2 and 1.3, which then produced the final predictions for intention and category. This approach helped reduce false positives and improved alignment with the task definition in the absence of explicit "NO" training for the downstream models.

In the soft-soft evaluation, models output probability distributions over the classes, which were compared directly with the normalized annotator distributions. This evaluation used ICM-soft, an extension of ICM that supports soft target labels.

4.2. Task 2

Task 2 focuses on detecting sexism in memes by combining textual and visual information. It includes three subtasks: binary classification (2.1), source intention classification (2.2), and multi-label categorization (2.3). To address these tasks, we implemented multimodal models that integrate pretrained transformer encoders for text and image. Specifically, we used XLM-RoBERTa (base and large variants) for text encoding and ViT-base (ViT-base-patch16-224-in21k) for image feature extraction. The [CLS] representations from both modalities were concatenated and passed through a shared classification head tailored to each task.

For subtask 2.1, the model was designed for binary classification, producing a two-dimensional output vector corresponding to the classes YES and NO. In the hard-label setting, the target was a one-hot vector reflecting the majority-voted class, using a threshold of at least four annotators selecting YES. In the soft-label setting, we used normalized distributions derived from the proportion of YES and NO votes as targets. During evaluation, sigmoid probabilities were computed, and the predicted label was determined using argmax.

For subtask 2.2, which classifies the intention behind a meme, the output dimension depended on the label encoding. In the hard-label setting, the model produced two logits corresponding to the DIRECT and JUDGEMENTAL categories, with the NO class excluded. The final label was derived from a majority threshold of at least three votes. In the soft-label setting, the output layer returned three values representing the probability distribution over DIRECT, JUDGEMENTAL, and NO. These soft labels captured the full range of annotator responses and allowed the model to learn from uncertainty and ambiguity in interpretation.

For subtask 2.3, the model performed multi-label classification using a six-dimensional output vector, each representing one of the predefined sexism categories. Each output neuron was trained independently to predict the probability of its corresponding label. In the hard-label configuration, binary vectors were constructed by assigning a label if it was selected by at least two annotators. In the soft-label configuration, each label's target value reflected its relative frequency among annotators, resulting in a normalized probability distribution across the six categories.

As in Task 1, the models for subtasks 2.2 and 2.3 were not trained to explicitly identify non-sexist memes. To prevent misclassifications of such instances, we implemented a cascaded decision strategy in the hard evaluation mode. Specifically, we used the output of the best-performing model from subtask 2.1 as a gatekeeper. If the classifier determined that a meme was non-sexist (label "NO"), the predictions

from the models for subtasks 2.2 and 2.3 were skipped, and label "NO" were assigned. Only when a meme was classified as sexist by the binary model was it forwarded to the downstream classifiers for intention and categorization. This pipeline reduced false positives and aligned more closely with the annotation logic of the dataset.

All models were trained using 10 epochs with early stopping. The training configuration included a batch size of 8, a learning rate of 2e-5, and a weight decay of 0.01. Loss functions were selected based on the label type and task characteristics. For soft-label training, we used MSELoss in Tasks 1 and 2, as their subtasks involved binary or multiclass outputs. For Task 3, which involves multilabel classification, we employed BCEWithLogitsLoss in the soft setting, applying it independently to each output class.

4.3. Task 3

Task 3 addresses the detection of sexism in short-form video content from TikTok, which presents unique challenges due to its multimodal nature and temporal structure. The task includes three subtasks: binary sexism identification (3.1), classification of the author's communicative intention (3.2), and multi-label categorization of the type of sexism expressed (3.3). For all subtasks, we used the pre-trained VideoMAE-base model for video classification, with input videos uniformly sampled into 16 frames and processed using the associated VideoMAE processor. We implemented both hard-label and soft-label systems for each subtask to capture annotation uncertainty and enable dual evaluation.

For subtask 3.1, the model was trained to classify videos as sexist (YES) or non-sexist (NO). The classifier produced a two-dimensional output, and training was performed using binary cross-entropy with logits in the hard-label variant. In the soft-label version, we used probabilistic targets based on the proportion of YES and NO votes, training the model with MSE loss.

In subtask 3.2, the objective was to determine the author's intention. In the hard-label setup, the model produced two outputs (DIRECT and JUDGEMENTAL), trained using binary cross-entropy with majority-vote thresholds. In the soft-label version, a third class (NO) was added to reflect cases where the annotators indicated no clear sexist intention. We modeled the label distribution across these three categories and trained the model using MSE loss, regressing against soft targets.

For subtask 3.3, we framed the problem as multi-label classification across five sexism categories. In the soft-label setup, an additional NO class was added to capture annotator uncertainty. The model returned a six-dimensional output vector, and was trained using BCEWithLogitsLoss applied independently to each class, in order to approximate the soft label distributions. In the hard-label version, we considered labels valid if they were selected by at least two annotators and trained the model using binary cross-entropy.

As in the other tasks, in the hard-label setting for subtasks 3.2 and 3.3, the models were not trained to explicitly predict the "NO" class. To address this limitation, we applied a sequential filtering strategy: the output of the best-performing classifier for subtask 3.1 was used as a gating mechanism. If a video was predicted as non-sexist in 3.1, the classifiers for 3.2 and 3.3 were bypassed, and label "NO" were assigned. If a video was predicted as sexist, the corresponding models for 3.2 and 3.3 were applied to generate the final output. This approach reduced false positives and aligned with the design of the dataset and the competition evaluation criteria. It should be noted, in the evaluation and inference stages, if a video could not be processed, the model automatically assigned it the NO label or a label distribution centered on the NO class in each task. This fallback mechanism ensured robustness and completeness in prediction outputs.

All models were trained using a batch size of 2, a learning rate of 2e-5, weight decay of 0.01, and 10 training epochs. Videos were preprocessed by extracting uniformly sampled frames and resizing them to 224×224 pixels.

5. Results

Regarding evaluation, EXIST 2025 defines two complementary evaluation settings: hard-hard and soft-soft. In the hard-hard setting, systems produce a hard label, which is compared against a ground-truth

label derived by majority voting with probabilistic thresholds specific to each subtask. For example, in subtasks 1.1 and 2.1, a class is accepted if annotated by more than 3 annotators; in 1.2 and 2.2, more than 2; and in multi-label subtasks 1.3 and 2.3, any class with more than 1 vote. For video subtasks (3.1 to 3.3), the complexity of annotation requires a more lenient threshold, accepting any class labeled by more than one annotator. Instances without majority consensus are excluded from this evaluation. In the soft-soft setting, systems produce probability distributions over the possible labels, which are compared against the aggregated probability distributions from human annotators. This variant captures uncertainty and ambiguity in both system predictions and human judgments, making it particularly suitable for subjective tasks such as sexism detection.

The official metric for both evaluation modes is the *Information Contrast Measure* (ICM) [21], a generalization of pointwise mutual information that quantifies the similarity between predicted and reference distributions.

5.1. Results of Task 1

Task 1 was addressed using two variants of the UMUTeam system. UMUTeam 1 corresponds to the version based on XLM-RoBERTa-base, while UMUTeam 2 uses XLM-RoBERTa-large. The following tables summarize the official results for both systems under the soft-soft and hard-hard evaluation schemes across subtasks 1.1 (binary classification), 1.2 (intention classification), and 1.3 (sexism categorization), with results reported globally and by language (Spanish and English).

Table 2 and Table 3 show the official results for Task 1.1. In both evaluation schema, UMUTeam 2 (large version) clearly outperformed UMUTeam 1, which based on XLM-RoBERTa-base. Under the soft-soft evaluation, UMUTeam 2 achieved a positive ICM-soft score (0.0138), archiving 41st overall, while UMUTeam 1 scored negatively (-0.1729). Similarly, in the hard-hard setting, UMUTeam 2 obtained an ICM of 0.5799 and ranked 11th overall, compared to UMUTeam 1's 0.5064 and 49th place. These results demonstrate the superior performance of the large model, particularly in capturing the subjective nature of the task across both languages.

Table 2
Official results for Task 1.1 with the Soft vs Soft scheme, including the Spanish and English results. Ranking is by runs.

	Soft vs Soft ALL		Soft vs Soft ES		Soft vs Soft EN	
Team	Rank	ICM-Soft	Rank	ICM-Soft	Rank	ICM-Soft
UMUTeam 1	44	-0.1729	45	-0.2307	47	-0.1367
UMUTeam 2	**41**	**0.0138**	**42**	**-0.0183**	**43**	**0.0196**
baseline majority class	64	-2.3585	63	-2.5421	61	-2.1991
baseline minority class	66	-3.1726	64	-2.5742	66	-3.8158

Table 3
Official results for Task 1.1 with the Hard vs Hard scheme, including the Spanish and English results. Ranking is by runs.

	Hard vs Hard ALL		Hard vs Hard ES		Hard vs Hard EN	
Team	Rank	ICM-Hard	Rank	ICM-Hard	Rank	ICM-Hard
UMUTeam 1	49	0.5064	44	0.5052	71	0.4963
UMUTeam 2	**11**	**0.5799**	**19**	**0.5699**	**10**	**0.5786**
baseline majority class	154	-0.4413	143	-0.4897	154	-0.3965
baseline minority class	157	-0.5742	145	-0.5106	157	-0.6646

Subtask 1.2 results are shown in Table 4 and Table 5. Again, UMUTeam 2 consistently outperformed UMUTeam 1 in all settings. In the soft-soft evaluation, UMUTeam 2 achieved a better alignment with

annotator distributions (ICM-soft: -3.6965 vs. -3.8401), and in the hard-hard setting, it reached an ICM of 0.3064 (12th), ahead of UMUTeam 1's 0.2647 (18th). This consistent margin of improvement highlights the benefits of deeper textual representation in modeling nuanced author intention.

Table 4
Official results for Task 1.2 with the Soft vs Soft scheme, including the Spanish and English results. Ranking is by runs.

	Soft vs Soft ALL		Soft vs Soft ES		Soft vs Soft EN	
Team	Rank	ICM-Soft	Rank	ICM-Soft	Rank	ICM-Soft
UMUTeam 1	23	-3.8401	23	-3.8492	22	-3.8059
UMUTeam 2	**22**	**-3.6965**	**22**	**-3.6857**	**21**	**-3.6890**
baseline majority class	37	-5.4460	37	-5.6674	34	-5.2028
baseline minority class	56	-32.9552	54	-28.7093	55	-39.4948

Table 5
Official results for Task 1.2 with the Hard vs Hard scheme, including the Spanish and English results. Ranking is by runs.

	Hard vs Hard ALL		Hard vs Hard ES		Hard vs Hard EN	
Team	Rank	ICM-Hard	Rank	ICM-Hard	Rank	ICM-Hard
UMUTeam 1	18	0.2647	23	0.2973	15	0.2073
UMUTeam 2	**12**	**0.3064**	**19**	**0.3123**	**6**	**0.2736**
baseline majority class	106	-0.9504	102	-1.0391	105	-0.8529
baseline minority class	139	-3.1545	135	-2.9390	139	-3.4728

In subtask 1.3, as reported in Table 6 and Table 7, the large-model system also led across the board. In the soft-soft evaluation, UMUTeam 2 ranked 4th overall with an ICM-soft of -1.6711, whereas UMUTeam 1 ranked 8th with a score of -3.0327. Similarly, in the hard-hard setting, UMUTeam 2 achieved an ICM of 0.4506 and ranked 7th, while UMUTeam 1 scored 0.3276, ranking 15th. These results confirm the robustness of the large model for multi-label classification tasks involving fine-grained sexism categories.

Overall, UMUTeam 2, based on XLM-RoBERTa-large, consistently outperformed UMUTeam 1 in all subtasks and evaluation conditions. The combination of deeper language modeling and soft-label training allowed it to better capture subtle cues and disagreement patterns present in the annotations.

Table 6
Official results for Task 1.3 with the Soft vs Soft scheme, including the Spanish and English results. Ranking is by runs.

	Soft vs Soft ALL		Soft vs Soft ES		Soft vs Soft EN	
Team	Rank	ICM-Soft	Rank	ICM-Soft	Rank	ICM-Soft
UMUTeam 1	8	-3.0327	12	-3.3571	8	-2.5755
UMUTeam 2	**4**	**-1.6711**	**4**	**-1.5920**	**4**	**-1.7900**
baseline majority class	26	-8.7089	26	-9.0314	26	-8.2105
baseline minority class	53	-46.1080	49	-45.4260	51	-46.9473

5.2. Results of Task 2

Task 2 evaluates sexism detection in memes across three subtasks: binary classification (2.1), intention classification (2.2), and category classification (2.3). We submitted two system variants: UMUTeam 1,

which uses XLM-RoBERTa-base as the text encoder, and UMUTeam 2, which employs XLM-RoBERTa-large. Both systems integrate textual and visual features using a multimodal architecture that combines outputs from the text encoder with ViT-based image representations.

Table 8 and Table 9 report the official results for subtask 2.1 under the soft-soft and hard-hard evaluation schemes. In the soft-soft setting, UMUTeam 2 ranked 6th overall with an ICM-soft of -0.9623, slightly ahead of UMUTeam 1, which ranked 7th with -1.2113. In the hard-hard evaluation, UMUTeam 2 ranked 13th, compared to UMUTeam 1 at 14th. Both systems outperformed the majority and minority baselines by a substantial margin, confirming the benefit of multimodal modeling even in the binary case.

In subtask 2.2 (author intention classification), UMUTeam 1 obtained the best performance among all participants, ranking 1st in the soft-soft evaluation with an ICM-soft of -1.6327 overall, -1.7469 in Spanish, and -1.5643 in English. UMUTeam 2 ranked 4th overall in the same setting, with a global ICM-soft of -2.4994. Under the hard-hard evaluation (Table 11), UMUTeam 2 ranked 8th globally with an ICM of -0.7265, outperforming UMUTeam 1, which ranked 12th with -0.7730. These results suggest that while the base model better captures soft-label ambiguity in intention, the large model maintains a slight advantage in hard classification settings.

For subtask 2.3 (category classification), UMUTeam 1 again achieved the top rank in the soft-soft evaluation, placing 1st across all language subsets with a global ICM-soft of -4.7791. UMUTeam 2 followed closely, ranking 2nd with -4.8825. In the hard-hard setting, UMUTeam 1 ranked 10th overall,

Table 7
Official results for Task 1.3 with the Hard vs Hard scheme, including the Spanish and English results. Ranking is by runs.

	Hard vs Hard ALL		Hard vs Hard ES		Hard vs Hard EN	
Team	Rank	ICM-Hard	Rank	ICM-Hard	Rank	ICM-Hard
UMUTeam 1	15	0.3276	17	0.3130	8	0.3258
UMUTeam 2	**7**	**0.4506**	**9**	**0.4650**	**4**	**0.4168**
baseline majority class	108	-1.5984	108	-1.7269	105	-1.4563
baseline minority class	128	-3.1295	125	-3.3196	127	-2.9279

Table 8
Official results for Task 2.1 with the Soft vs Soft scheme, including the Spanish and English results. Ranking is by runs.

	Soft vs Soft ALL		Soft vs Soft ES		Soft vs Soft EN	
Team	Rank	ICM-Soft	Rank	ICM-Soft	Rank	ICM-Soft
UMUTeam 1	7	-1.2113	7	-1.3398	7	-1.0907
UMUTeam 2	**6**	**-0.9623**	**6**	**-1.0271**	**6**	**-0.9012**
baseline majority class	8	-2.3568	8	-2.4997	9	-2.2236
baseline minority class	10	-3.5089	9	-3.9408	10	-3.1235

Table 9
Official results for Task 2.1 with the Hard vs Hard scheme, including the Spanish and English results. Ranking is by runs.

	Hard vs Hard ALL		Hard vs Hard ES		Hard vs Hard EN	
Team	Rank	ICM-Hard	Rank	ICM-Hard	Rank	ICM-Hard
UMUTeam 1	14	-0.3043	13	-0.3064	18	-0.3021
UMUTeam 2	**13**	**-0.2957**	**14**	**-0.3169**	**17**	**-0.2744**
baseline majority class	17	-0.4038	15	-0.4001	19	-0.4076
baseline minority class	20	-0.6468	16	-0.6557	20	-0.6381

while UMUTeam 2 ranked 12th. These results demonstrate the base model's capacity to model fine-grained, multi-label decisions effectively, especially under soft-label conditions.

In summary, UMUTeam 2 (XLM-RoBERTa-large) performed best in binary classification and in most hard evaluations, while UMUTeam 1 (XLM-RoBERTa-base) showed stronger performance in subjective and soft-label tasks. Across all three subtasks and both evaluation schemes, both models consistently outperformed the baselines. Notably, UMUTeam 1 ranked 1st in Task 2.2 and Task 2.3 (soft-soft), positioning the system among the top performers in the competition.

Table 10
Official results for Task 2.2 with the Soft vs Soft scheme, including the Spanish and English results. Ranking is by runs.

	Soft vs Soft ALL		Soft vs Soft ES		Soft vs Soft EN	
Team	Rank	ICM-Soft	Rank	ICM-Soft	Rank	ICM-Soft
UMUTeam 1	**1**	**-1.6327**	**1**	**-1.7469**	**1**	**-1.5643**
UMUTeam 2	4	-2.4994	4	-2.7757	3	-2.2635
baseline majority class	6	-5.0745	5	-5.5913	6	-4.6049
baseline minority class	7	-18.9382	6	-20.3091	7	-18.1227

Table 11
Official results for Task 2.2 with the Hard vs Hard scheme, including the Spanish and English results. Ranking is by runs.

	Hard vs Hard ALL		Hard vs Hard ES		Hard vs Hard EN	
Team	Rank	ICM-Hard	Rank	ICM-Hard	Rank	ICM-Hard
UMUTeam 1	12	-0.7730	10	-0.7553	12	-0.7912
UMUTeam 2	**8**	**-0.7265**	**7**	**-0.6791**	**11**	**-0.7735**
baseline majority class	14	-1.0445	11	-1.0504	14	-1.0385
baseline minority class	17	-2.0637	13	-2.0866	17	-2.0410

Table 12
Official results for Task 2.3 with the Soft vs Soft scheme, including the Spanish and English results. Ranking is by runs.

	Soft vs Soft ALL		Soft vs Soft ES		Soft vs Soft EN	
Team	Rank	ICM-Soft	Rank	ICM-Soft	Rank	ICM-Soft
UMUTeam 1	**1**	**-4.7791**	**1**	**-4.9258**	**1**	**-4.6710**
UMUTeam 2	2	-4.8825	2	-4.9552	2	-4.8909
baseline majority class	6	-9.8173	5	-10.4121	6	-9.2886
baseline minority class	8	-50.0353	6	-47.3973	8	-53.0762

Table 13
Official results for Task 2.3 with the Hard vs Hard scheme, including the Spanish and English results. Ranking is by runs.

	Hard vs Hard ALL		Hard vs Hard ES		Hard vs Hard EN	
Team	Rank	ICM-Hard	Rank	ICM-Hard	Rank	ICM-Hard
UMUTeam 1	**10**	**-1.5624**	**8**	**-1.6943**	**11**	**-1.4714**
UMUTeam 2	12	-1.8869	10	-1.9764	13	-1.8424
baseline majority class	14	-2.0711	11	-2.1173	14	-2.0015
baseline minority class	16	-3.3135	13	-3.2599	16	-3.3506

5.3. Results of Task 3

Task 3 focused on detecting sexism in short-form video content from TikTok across three subtasks: binary identification (3.1), intention classification (3.2), and multi-label categorization (3.3). Only one run was submitted for this task, denoted as UMUTeam 1, based on the VideoMAE-base architecture. The evaluation was conducted under both the soft-soft and hard-hard paradigms using the ICM metric.

Table 14 and Table 15 present the results for subtask 3.1. In the soft-soft evaluation, UMUTeam 1 achieved an ICM-soft of -1.9857 (ranked 30th), outperforming the minority baseline (-2.0051) and close to the majority baseline (-1.2877). In the hard-hard evaluation, UMUTeam 1 obtained an ICM of -0.6926 (rank 43), performing better than the minority baseline (-0.6036) but slightly behind the majority class baseline (-0.4244). These results suggest that the system could effectively learn from distributional annotations but may still face challenges under strict classification constraints, particularly given the complexity and ambiguity of video data.

Table 14
Official results for Task 3.1 with the Soft vs Soft scheme, including the Spanish and English results. Ranking is by runs.

	Soft vs Soft ALL		Soft vs Soft ES		Soft vs Soft EN	
Team	Rank	ICM-Soft	Rank	ICM-Soft	Rank	ICM-Soft
UMUTeam 1	**30**	**-1.9857**	**30**	**-2.2412**	**31**	**-1.9222**
baseline majority class	26	-1.2877	20	-0.8905	30	-1.7009
baseline minority class	31	-2.0051	32	-2.5626	29	-1.6647

Table 15
Official results for Task 3.1 with the Hard vs Hard scheme, including the Spanish and English results. Ranking is by runs.

	Hard vs Hard ALL		Hard vs Hard ES		Hard vs Hard EN	
Team	Rank	ICM-Hard	Rank	ICM-Hard	Rank	ICM-Hard
UMUTeam 1	**43**	**-0.6926**	**39**	**-0.8216**	**43**	**-0.6192**
baseline majority class	40	-0.4244	35	-0.3708	40	-0.4836
baseline minority class	42	-0.6036	38	-0.7411	42	-0.5174

For subtask 3.2, the results are shown in Table 16 and Table 17. In the soft-soft setting, UMUTeam 1 ranked 26th overall with an ICM-soft of -3.0703, which was superior to the minority baseline (-15.4368) and comparable to the majority class baseline (-3.1337). In the hard-hard evaluation, the system achieved an ICM of -1.1856 and ranked 36th overall. These results indicate that the model was able to identify intent with reasonable alignment to annotator disagreement, although its performance slightly decreased in the hard classification scenario.

Table 16
Official results for Task 3.2 with the Soft vs Soft scheme, including the Spanish and English results. Ranking is by runs.

	Soft vs Soft ALL		Soft vs Soft ES		Soft vs Soft EN	
Team	Rank	ICM-Soft	Rank	ICM-Soft	Rank	ICM-Soft
UMUTeam 1	**26**	**-3.0703**	**19**	**-3.9102**	**28**	**-2.9509**
baseline majority class	27	-3.1337	16	-3.3756	29	-3.3354
baseline minority class	31	-15.4368	30	-35.3991	31	-9.7700

Subtask 3.3, the multi-label categorization task, posed a more significant challenge. In the soft-soft evaluation (Table 18), UMUTeam 1 ranked 26th with an ICM-soft of -9.0825, outperforming the

minority class baseline (-11.6668) and the majority baseline (-6.8222). In the hard-hard setting (Table 13), UMUTeam 1 ranked 40th with an ICM of -2.7332, again better than the minority baseline (-6.7467), though below the majority class (-0.9530). The relatively low ICM scores in this subtask reflect the complexity of modeling nuanced, co-occurring sexist categories in audiovisual content.

It should be noted, a critical factor affecting the overall performance in Task 3 was the model's inability to process a substantial portion of the test videos. Due to frame extraction or decoding failures during inference, these videos were automatically assigned the label "NO" in all subtasks. As a result, many true positive cases were misclassified as non-sexist, significantly increasing the number of false negatives. This issue had a direct and negative impact on both soft and hard evaluation metrics, particularly in the hard-hard setting, where missing detections are penalized more severely. However, in summary, UMUTeam 1 consistently outperformed the minority baseline and was competitive with the majority baseline in some settings.

6. Conclusions and further work

In this work, we presented a comprehensive system for sexism detection in social media, developed for the EXIST 2025 shared task. Our approach covered all three content modalities: text, images, and videos, and addressed each of the three subtasks across hard and soft evaluation settings. We employed multilingual transformer-based models, such as XLM-RoBERTa (base and large versions) for text, ViT for image features, and VideoMAE for video input. Specialized preprocessing and label handling were applied for each modality. Despite the fact that we leveraged transformer-based encoders adapted to

Table 17
Official results for Task 3.2 with the Hard vs Hard scheme, including the Spanish and English results. Ranking is by runs.

	Hard vs Hard ALL		Hard vs Hard ES		Hard vs Hard EN	
Team	Rank	ICM-Hard	Rank	ICM-Hard	Rank	ICM-Hard
UMUTeam 1	**36**	**-1.1856**	**36**	**-1.3289**	**35**	**-1.0901**
baseline majority class	34	-0.7537	33	-0.6304	34	-0.8657
baseline minority class	38	-2.4749	37	-2.9537	38	-2.1890

Table 18
Official results for Task 3.3 with the Soft vs Soft scheme, including the Spanish and English results. Ranking is by runs.

	Soft vs Soft ALL		Soft vs Soft ES		Soft vs Soft EN	
Team	Rank	ICM-Soft	Rank	ICM-Soft	Rank	ICM-Soft
UMUTeam 1	**26**	**-9.0825**	**25**	**-10.4828**	**27**	**-9.8929**
baseline majority class	9	-6.8222	11	-7.5409	11	-6.8246
baseline minority class	28	-11.6668	12	-7.7430	29	-11.4332

Table 19
Official results for Task 3.3 with the Hard vs Hard scheme, including the Spanish and English results. Ranking is by runs.

	Hard vs Hard ALL		Hard vs Hard ES		Hard vs Hard EN	
Team	Rank	ICM-Hard	Rank	ICM-Hard	Rank	ICM-Hard
UMUTeam 1	**40**	**-2.7332**	**39**	**-2.8851**	**40**	**-2.6572**
baseline majority class	31	-0.9530	24	-0.8103	34	-1.0678
baseline minority class	41	-6.7467	40	-5.6377	41	-8.4611

each modality, we implemented a modular architecture that incorporates both discrete and probabilistic label modeling. The use of soft-label learning with MSE and BCE loss allowed our systems to account for annotator disagreement, while the hard-label setup used thresholded ground truth for comparison. Additionally, to reduce false positives in downstream classification (subtasks 1.2, 1.3, 2.2, 2.3, and 3.2, 3.3), we integrated a binary prediction filter based on the outputs of subtasks 1.1, 2.1, and 3.1.

Our systems demonstrated strong performance across tasks. In the text modality (Task 1), UMUTeam 2 (XLM-RoBERTa-large) ranked as high as 12th and 7th in subtasks 1.2 and 1.3 (hard-hard), confirming the value of large-scale models and soft-label learning. In the image modality (Task 2), UMUTeam 1 achieved 1st place in both subtasks 2.2 and 2.3 under the soft-soft setting, and both models consistently outperformed the baselines across evaluation types. For video (Task 3), while the system faced limitations due to frame extraction failures, it still outperformed the minority baseline and remained competitive with the majority baseline in several subtasks. Overall, our results validate the effectiveness of multimodal, disagreement-aware approaches in modeling the complexity and subjectivity inherent in sexism detection tasks.

Although our systems achieved competitive results, several areas remain open for improvement. First, in the video modality, decoding errors and frame extraction failures affected the completeness and reliability of the inference stage. Future work will focus on improving preprocessing robustness and integrating additional modalities such as audio and automatic speech recognition to capture spoken content. Additionally, we plan to use LLM techniques, such as prompt engineering, as described in [22] and [23] to improve sexism identification and source intention detection. Finally, in the image and video pipelines, we plan to explore stronger multimodal fusion strategies such as cross-attention to better capture interactions between modalities.

Acknowledgments

This work is part of the research project LaTe4PoliticES (PID2022-138099OB-I00) funded by MCIN/AEI/10.13039/501100011033 and the European Regional Development Fund (ERDF)-a way to make Europe. Mr. Tomás Bernal-Beltrán is supported by University of Murcia through the predoctoral programme.

Declaration on Generative AI

During the preparation of this work, the author(s) used DeepL in order to: Grammar and spelling check.

References

[1] M. V. Campuzano, Force and inertia: A systematic review of women's leadership in male-dominated organizational cultures in the united states, Human Resource Development Review 18 (2019) 437–469.

[2] M. Dosil, J. Jaureguizar, E. Bernaras, J. B. Sbicigo, Teen dating violence, sexism, and resilience: A multivariate analysis, International journal of environmental research and public health 17 (2020) 2652.

[3] P. Glick, S. T. Fiske, Hostile and benevolent sexism: Measuring ambivalent sexist attitudes toward women, Psychology of women quarterly 21 (1997) 119–135.

[4] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, Detecting sexism in social media: an empirical analysis of linguistic patterns and strategies, Applied Intelligence 54 (2024) 10995–11019.

[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), Proceedings of the 16th International Conference of the CLEF Association (CLEF 2025), Springer Lecture Notes in Computer Science LNCS, Madrid, Spain, 2025.
[6] 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, Proceedings of the 16th International Conference of the CLEF Association (CLEF 2025), Springer Lecture Notes in Computer Science LNCS, Madrid, Spain, 2025.
[7] 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.
[8] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, P. Vajda, Visual transformers: Token-based image representation and processing for computer vision, arXiv preprint arXiv:2006.03677 (2020).
[9] R. Pan, J. A. García-Díaz, M. Ángel Rodríguez-García, R. Valencia-García, Spanish meacorpus 2023: A multimodal speech–text corpus for emotion analysis in spanish from natural environments, Computer Standards & Interfaces 90 (2024) 103856. URL: https://www.sciencedirect.com/science/article/pii/S0920548924000254. doi:https://doi.org/10.1016/j.csi.2024.103856.
[10] Z. Tong, Y. Song, J. Wang, L. Wang, Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training, Advances in neural information processing systems 35 (2022) 10078–10093.
[11] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on Twitter, in: J. Andreas, E. Choi, A. Lazaridou (Eds.), Proceedings of the NAACL Student Research Workshop, Association for Computational Linguistics, San Diego, California, 2016, pp. 88–93. URL: https://aclanthology.org/N16-2013/. doi:10.18653/v1/N16-2013.
[12] E. Fersini, P. Rosso, M. Anzovino, et al., Overview of the task on automatic misogyny identification at ibereval 2018., Ibereval@ sepln 2150 (2018) 214–228.
[13] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, T. Donoso, Overview of exist 2021: sexism identification in social networks, Procesamiento del Lenguaje Natural 67 (2021) 195–207.
[14] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, A. Mendieta-Aragón, G. Marco-Remón, M. Makeienko, M. Plaza, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2022: sexism identification in social networks, Procesamiento del Lenguaje Natural 69 (2022) 229–240.
[15] 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. URL: https://aclanthology.org/2023.semeval-1.305. doi:10.18653/v1/2023.semeval-1.305.
[16] A. Chhabra, D. K. Vishwakarma, Multimodal hate speech detection via multi-scale visual kernels and knowledge distillation architecture, Engineering Applications of Artificial Intelligence 126 (2023) 106991. URL: https://www.sciencedirect.com/science/article/pii/S0952197623011752. doi:https://doi.org/10.1016/j.engappai.2023.106991.
[17] N. Jindal, P. K. Kumaresan, R. Ponnusamy, S. Thavareesan, S. Rajiakodi, B. R. Chakravarthi, Mistra: Misogyny detection through text–image fusion and representation analysis, Natural Language Processing Journal 7 (2024) 100073. URL: https://www.sciencedirect.com/science/article/pii/S2949719124000219. doi:https://doi.org/10.1016/j.nlp.2024.100073.
[18] M. Wiegand, J. Ruppenhofer, T. Kleinbauer, Detection of abusive language: the problem of biased datasets, in: North American Chapter of the Association for Computational Linguistics, 2019. URL: https://api.semanticscholar.org/CorpusID:174799974.
[19] X. Zhou, M. Sap, S. Swayamdipta, Y. Choi, N. Smith, Challenges in automated debiasing for toxic language detection, in: P. Merlo, J. Tiedemann, R. Tsarfaty (Eds.), Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics:

Main Volume, Association for Computational Linguistics, Online, 2021, pp. 3143–3155. URL: https://aclanthology.org/2021.eacl-main.274. doi:10.18653/v1/2021.eacl-main.274.

[20] L. De Grazia, P. Pastells, M. V. Chas, D. Elliott, D. S. Villegas, M. Farrús, M. Taulé, Mused: A multimodal spanish dataset for sexism detection in social media videos, arXiv preprint arXiv:2504.11169 (2025).

[21] 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. URL: https://aclanthology.org/2022.acl-long.399/. doi:10.18653/v1/2022.acl-long.399.

[22] R. Pan, J. A. García-Díaz, R. Valencia-García, Optimizing few-shot learning through a consistent retrieval extraction system for hate speech detection, Procesamiento del Lenguaje Natural 74 (2025) 241–252.

[23] R. Pan, J. A. García-Díaz, R. Valencia-García, Spanish mtlhatecorpus 2023: Multi-task learning for hate speech detection to identify speech type, target, target group and intensity, Computer Standards & Interfaces 94 (2025) 103990. URL: https://www.sciencedirect.com/science/article/pii/S0920548925000194. doi:https://doi.org/10.1016/j.csi.2025.103990.