ECA-SIMM-UVa at EXIST 2025: A Segmentation Oriented Approach to Sexism Detection in Tik Tok Videos Based on a "One Is Enough" Paradigm

Notebook for the EXIST Lab at CLEF 2025

David Fernández García[1,*], Enrique Amigó Cabrera[2] and Valentín Cardeñoso Payo[1]

[1]*ECA-SIMM Research group, University of Valladolid, Spain*
[2]*Universidad Nacional a Distancia (UNED), Spain*

Abstract

This paper details the ECA-SIMM-UVa team's participation in Task 7: Sexism Identification in TikToks as part of the EXIST 2025 challenge. The focus is on automatic detection of potentially harmful sexist behaviours on social platforms. We adopted a segmentation oriented approach, splitting TikTok videos into textual, audio, and video channels, on the hypothesis that sexism can manifest in spoken words, embedded text, speaker tone, or visual content (text, pictures or other images). We trained individual deep learning classifiers for each channel and explored various prediction fusion mechanisms like One Is Enough (OIE), Majority Voting, and Probabilistic OIQ for hard evaluation, as well as Logistic Regression and Weighted Sum for soft evaluation, to combine predictions. As a significant finding, models using the textual channel show superior performance, specially when using the original text provided with each sample in the dataset. They consistently outperformed audio and video channels, indicating textual information as the most informative source for sexism detection in this context. Although fusion mechanisms achieved good estimation performance, it was frequently associated, almost exclusively, to the presence of decisions made on the original-text specific model being fused with the others, effectively disregarding contributions from the audio and video channels due to high thresholds. Our systems ranked 1st, 3rd, and 7th out of 41 submissions in the hard evaluation category, and 15th, 17th, and 18th out of 35 submissions in the soft evaluation category, considering instances of any language in both cases. Our results emphasizes the challenges that multimodal sexism detection still faces and the need to further improve pre-trained audio and video models.

Keywords

Segmentation, Fusion Mechanism, TikToks, Multimodal, Sexism

1. Introduction

Nowadays we can find many different social platforms like Twitter, Instagram or Tik Tok, where people can share huge variety of multimedia and hypermedia publications brought to users in a multimodal fashion. This is a great opportunity to encourage positive social interaction and discussions but it also opens way for potentially dangerous and harmful behaviours, like sexism or misogyny, by becoming huge loudspeakers for many kinds of discriminatory content. Due to that, one of the most important problems nowadays is how to deploy realistic regulations and mechanisms to detect, control and mitigate these types of behaviour. The vast amount of content upload to these platforms makes it impossible to address this controls under a manual fashion. Thus, the development of automatic tools that can help to address control and mitigation of harmful information and behaviours become an open challenge for the following years.

The EXIST challenge (sEXism Identification in Social neTworks) is a group of tasks that try to promote research related to designing, implementing and evaluating automatic sexism detection systems on social networks content. This year's challenge include three different global tasks:

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
david.fernandez@uva.es (D. F. García); enrique@lsi.uned.esnl (E. A. Cabrera); valentin.cardenoso@uva.es (V. C. Payo)
0009-0009-6826-0195 (D. F. García); 0000-0003-1482-824X (E. A. Cabrera); 0000-0003-1460-158X (V. C. Payo)


- **Global Task 1**: This is a **binary classification task**, where systems have to decide whether one post is sexist or not.
- **Global Task 2**: This is a **multi-class classification task**, where systems have to identify the author's intention of the posts classified as sexist in *Task 1*. There are three different source intention classes: *direct*, *reported* or *judgemental*.
- **Global Task 3**: This is **multi-label classification task**, where systems have to categorize sexist posts based on a set of defined types: *ideological-inequality*, *stereotyping-dominance*, *objectification*, *sexual-violence*, *misogyny-non-sexual-violence*.

Each of these global tasks can be faced from three different perspectives, depending on the kind of input media being considered for the source: textual posts, meme posts and video posts. For a more complete description of the challenge and overview documents, see [1, 2, 3].

This paper describes the participation of ECA-SIMM-UVa research team in this challenge. We focus on Task 7: Sexism Identification in TikToks, trying to deal with one of the most important social media platform of our days. We built systems for both soft and hard evaluation modalities. Our approach is based on the initial segmentation of videos into three different source channels: text, audio and image sequences (images). The textual channel includes both the audio transcription of the words in the video and any kind of text material that can be recognized as embedded in the video itself. Audio channel includes the sound track extracted from the video. Images channel includes the sequence of image frames in the video without audio. For each channel, a deep learning based classifier is trained, through a fine-tuning of a specialized pre-trained model on each type of data. Then, we explore different classification fusion mechanisms, in order to merge the three individual predictions for text, audio and images into a final decision. Different fusion mechanisms were applied in this work, depending of the evaluation type. For hard evaluation we tried: *OIQ*, *One Is Enough (OIE)* and *Majority Voting*. For soft evaluation we try: *Logistic regression* and *Weighted sum of channels*.

Therefore, the study aims not only to obtain good competition results on the classification task, but also addresses a comparative study of classification models for different information channels, analysing the relative importance of each channel for the sexism identification task. More specifically, we seek to answer the following research questions:

- How do pre-trained text, audio and image models compare in terms of performance to classify contents as sexist or not?
- What is the relative impact of the three information channels (text, audio and image sequence) on classification performance, both in an isolated and combined fashion?

The paper is organised as follows. Section 2 investigates related studies for sexism identification on video. Section 3 describes the ECA-SIMM-UVa approach for Task 7 of EXIST 2025. Section 4 presents results and rankings. Finally, in Section 5, we include discussion, conclusions, and suggestions for future work.

2. Related Work

The importance of automated detection of sexism on digital platforms has recently increased, as a consequence of the endless amount of multimedia content delivered every hour through social networks and other distribution channels. Researchers have made serious efforts to develop these ML and AI based automated systems, promoting and participating in initiatives like EXIST [4, 5, 6, 7] or SemEval 2023 challenges [8], and through the publication of relevant studies [9, 10]. Research has mainly focused on textual data, which is easily obtained from social networks like X (Twitter) or Gab, and also from other audio and video sources by means of automatic speech recognizers of ever increasing quality.

In the realm of sexism identification based on text, as addressed in EXIST 2024 [7] Tasks 1, 2, and 3, and the textual component of Tasks 4, 5, and 6, the state of the art is primarily characterized by the dominant use of encoding-based transformer models fine-tuned on the EXIST dataset, frequently enhanced with

additional components. Meticulous data preprocessing was a key factor in improving performance, involving removal of irrelevant elements and the application of data augmentation techniques like AEDA [11] and automatic English-Spanish translation. Ensembles of encoding-based transformer models such as BERT [12], RoBERTa [13], and DeBERTa [14] (including multilingual versions or those pre-trained on domains like tweets or hate speech) proved highly effective, particularly in the soft evaluation setting, which was linked to their training with soft labels. Ensemble strategies varied, from assigning higher weight to the best-performing model for significant performance differences, to using a proportion of votes when differences were smaller. A significant distinguishing factor, especially successful in the hard evaluation setting, was the incorporation of Large Language Models (LLMs) like Llama [15], Mistral [16], and GPT, primarily used for zero-shot or few-shot learning due to computational costs, relying heavily on prompt engineering. While encoding-based transformers generally excelled in soft evaluation, LLMs showed superior performance in hard evaluation. For multimodal tasks (4, 5, 6), top performances were unexpectedly achieved by models focusing solely on text, with systems using encoding-based transformers for text analysis often outranking truly multimodal approaches.

Video automated detection of sexism has attracted much less interest for researchers, probably due to the difficulties associated with its collection, processing, information extraction and model training for them. Thus, few works can be found that specifically deal with sexism. A novel corpus of 11 hours of video extracted from Tik Tok and BitChute, was presented in [17], a videos' dataset which is annotated at three different levels: text, audio and image sequences. In [18], a segmentation and multimodal approach is explored to face the problem. They use a wide variety of models such as RoBERTa [13] (textual model), Wav2Vec [19] (audio model) and ViT [20] (video model).

As the field of interest broadens, a higher number of works using video sources is found, as in, for example, hate speech detection [21, 22, 23, 24]. A common practice is to obtain or extract text transcriptions from videos, and train classifiers just with that textual data. Other approaches prefer the multimodal way, combining text, audio and video features [25, 26]. Regarding these approximations, we can find the use of Multimodal deep learning systems [27, 28, 29] or models ensemble approaches [30, 31], which mainly use majority vote to make their decisions.

The shortage of works centered around video detection of sexism is a clear symptom that the research must pay more attention to this field, specially because of its increasing importance. Development of new corpora, improvement of multimodal models and an increase of consciousness on the importance of this kind of research, become crucial factors for boosting this field.

3. ECA-SIMM-UVa Approach

3.1. Data

Table 1 shows the composition and distribution of the EXIST 2025 Tik Tok Dataset [1] both in English and Spanish. This dataset was specifically developed for the challenge, extending sexism detection tasks to TikTok videos. TikTok's recommendation algorithm could reinforce sexism and normalize misogynistic attitudes, significantly impacting adolescents' self-esteem and gender perceptions. Apify's TikTok Hashtag Scraper was used for data collection, and, as a crucial feature of the dataset, annotation was performed by trained annotators from Servipoli, organized into mixed-gender pairs to avoid biases. A Learning With Disagreement paradigm was adopted, incorporating diverse human perspectives and disagreements to foster human-centric AI, thereby reducing bias and promoting inclusive decision-making. The dataset supports three main tasks: Sexism Identification, Source Intention Detection, and Sexism Categorization.

3.2. Segmentation

We adopted a **segmentation-based approach** to video detection of sexism (see Figure 1). This decision was based on the hypothesis that a Tik Tok video can be perceived as sexist because of four distinct reasons:

Table 1
Distribution of the dataset across different partitions, segmented by language. The table details the number of instances in the training and test sets for both Spanish and English.

Language	Training Set	Test Set	Total
Spanish	1,524	304	1,828
English	1,000	370	1,370
Total	**2,524**	**674**	**3,198**

1. The **semantic content** of the spoken words is sexist. This involves using audio processing and speech transcription as a source.
2. The **embedded text** within the video conveys sexist content. This includes any on-screen textual elements, such as real posters or comments added on the video.
3. The **tone or speech intention** of the speaker carries a sexist attitude. This involves audio signal processing and analysis.
4. The **visual content** of the video is sexist. This involves visual scene analysis.

Based on those four paths to sexism detection in videos, we conduct a segmentation of Tik Tok videos to split them into 3 input channels: **text**, **audio** and **video**.

In our experiments, we considered three different sources for the text channel. First one was the text input that was given with the dataset. This text includes a combination of textual transcription and embedded text. Then, we obtain two additional textual sources: we use *Whisper-X* [32] to get a detailed time-aligned text transcription of videos. This allowed us to identify what was said and when was it said; in a second place, a textual channel was obtained using *DeepSeek-VL* [33] to extract text messages embedded as images in the video frames. After experimenting with various prompts, we opted for a *zero-shot prompting* strategy with this tool (see Listing 1). As the output of this processing, we get three text tiers in the text channel: **original**, which mixes transcription and embedded text, **transcription** and **embedded text**.

Listing 1: Prompt for zero-shot prompting video text extraction with*DeepSeek-VL* [33]

```
{
    "role": "User",
    "content": "<image_placeholder> Extract ONLY the main visible
       text in this image and give it back. Ignore TikTok interface
       elements such as: hashtags, user IDs, counters, buttons,
       mentions, tags, or any text overlaid by the platform. Focus
       exclusively on text that is part of the original video
       content. The text may be in English or Spanish. Preserve the
       original format (uppercase/lowercase) and organize the text
       in natural reading order. Do not add interpretations, context
       , or explanations. Return ONLY the extracted text. In order
       to give back the text extracted use this form: The text
       extracted in the image is: <here put the text you have
       extracted >.",
    "images": [image_path]
},
{
    "role": "Assistant",
    "content": ""
}
```

To extract audio and video channels, we used *ffmpeg* library, which is a framework commonly used for audio, video and multimedia file and stream processing.

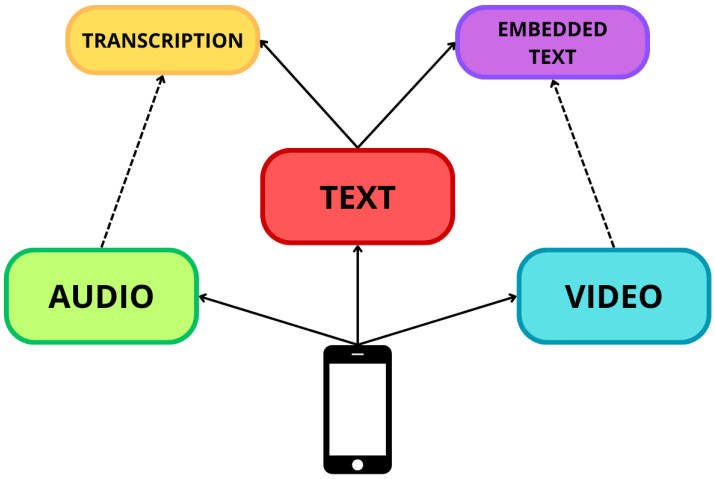

Figure 1: Diagram of the segmentation-based approach adopted in this work. (Continuous/Dotted) lines indicate (direct/indirect) relationship.

3.3. General Approach

We trained three classifiers, one for each input channel (audio, text, video). To ensure comparability across models, we applied a common architecture and training strategy for each of them (see Figure 2). All classifiers consist of a channel-specific encoder followed by a *Multi-Layer Perceptron (MLP)*. Given that our task is binary classification (sexist vs. non-sexist), we used the *Binary Cross-Entropy (BCE)* loss function during training.

We conducted hyper-parameter tuning using the *optuna* library, with an 80/20 train-validation split. The original approach of this work consisted of the hypothesis shown in Figure 1, which assumes that a video will be classified as sexist just if one of the models classifies the video as such, ignoring the decisions of the rest of models. This implies that the precision of each classifier is critical, since a single false positive leads to a global misclassification. In this scenario, we have to use a more specific metric, which gives greater weight to the precision, in order to decide which hyper-parameter set it is the optimal one. We chose *F-Beta*, with $\beta = 0.5$, as primary optimization metric, in order to give twice as much importance to *Precision* as compared to *Recall*. For each fixed set of hyper-parameters, we also performed threshold calibration through exhaustive threshold search, in order to maximize our primary metric. For completion, we also monitored alternative metrics: *ICM* [34], *ICM_norm* and *F1 Score*.

Once the hyper-parameter search was complete, we need to estimate the real error of our system. To get a global estimation over the entire dataset, we apply *5-Fold-Cross-Validation*, which allows us to obtain an evaluation for each specific sample. As in the tuning phase, threshold adjustment was performed after each fold.

Finally, after selecting the optimal hyper-parameters, we retrained each model on the full dataset in order to coin the final classifiers we used for downstream analysis.

3.4. Models

Procedure described in Figure 2 was followed for every channel-specific system we trained. However, due to the unique characteristics of each channel, certain differences emerged in the training processes for the three models. We trained three distinct textual models, each of them with one of the three textual representations we mentioned in Section 3.2. All three were trained using an identical set of hyper-parameters to ensure comparability. As a pre-trained model for the textual channel, we used **XLM-RoBERTa-Large** [35]; **Wav2Vec-Large-XLRS-53** [36] was used for audio channel and **ViViT-b-16x2-Kinetics400** [37] for video channel. Specific hyper-parameter values for each model can be seen in Table 2.

In our experiments, we consider three configurations based on the source of textual input. **Original**

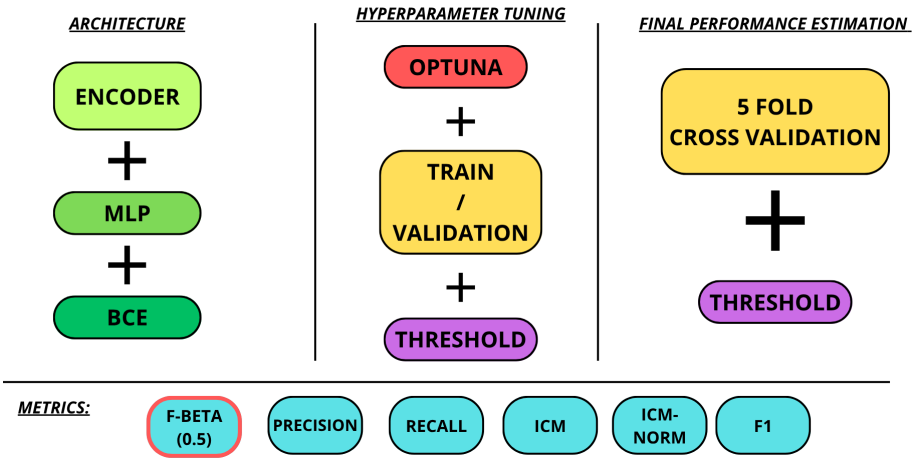

Figure 2: A common methodological pipeline adopted to train all channel-specific models.

configuration uses the model trained with textual input provided directly by the dataset. **Own configuration** includes two separate textual models, one trained on automatically extracted transcriptions and the other on automatically extracted embedded text. **All configuration** includes all three textual models.

All experiments were carried out using a *NVIDIA A-40* GPU with 48GB RAM. Due to high memory requirements, mainly when we process video and audio, we had to apply *gradient accumulation* technique in order to simulate larger batch sizes during training. It should be noted that the available hardware turns out to be a limitation when training audio and video models, since these require a large amount of resources.

Table 2
Hyper-parameters used for fine-tuning different modality models. (1) **XLM-RoBERTa-Large** [35], (2) **Wav2Vec-Large-XLRS-53** [36], (3) **ViViT-b-16x2-Kinetics400** [37].

Hyperparameter	Text(1)	Audio(2)	Video(3)
Learning Rate	8.65e-06	2.65e-04	4.68e-05
Batch Size	8	32	8
Epochs	11	2	2
Warmup Ratio	0.09826	0.00000	0.23611
Weight Decay	0.01707	0.00000	0.01269
Model Dropout Rate	0.1	0.0	0.1
Classifier Dropout Rate	0.0	0.0	0.1

3.5. Fusion Mechanisms

As pointed out in Section 3.3, our first attempt was to follow a *One Is Enough (OIE)* approach, so all the training process was focus on that. However, we also explored alternative fusion mechanisms to combine model outputs. Different fusion alternatives were chosen, depending on the type of evaluation, because, while in hard evaluation we have to get a final label, in soft evaluation we should obtain a correct likelihood distribution of labels. We implemented three fusion mechanisms for hard evaluation:

1. **One is Enough (OIE)**: A video is classified as sexist if any individual model detects it as such. This mechanism follows a similar idea to what can happen when screening critical patients at hospital. If the objective is to determine whether the patient is sick or not, it is enough for one of the specialists to affirm it, without the need of further opinions of other doctors.
2. **Majority Voting**: A simple majority rule is applied, where at least half plus one, out of total number models, must classify the video as sexist for the final label to be positive.

3. **Probabilistic OIQ**: This method considers all possible combinations of binary outputs from the models. For each pattern (e.g., [1, 0, 1]), we estimate the empirical probability that the video is sexist, based on the classification distribution of each sequence. Notice that before applying this method we have to adjust an individual threshold for each model, which should maximize that model's performance. At inference time, the model's predicted output pattern is matched against these empirical probabilities. A final label is assigned based on whether this probability exceeds a tuned decision threshold.

Regarding soft fusion mechanisms, we explored two fusion strategies:

1. **Logistic Regression Fusion**: A meta-model is trained to map the predicted probabilities from each channel to a final soft label. The model is optimized to approximate the ground truth probability distribution.
2. **Weighted Sum of Predictions**: Fixed weights are assigned to each channel's output. These weights are optimized using the *SLSQP algorithm* [38], minimizing the cross-entropy loss between the weighted prediction and the soft ground truth.

4. Results

Table 3 shows the best achieved estimation performance for each individual channel using 5-fold cross-validation. For all individual channels, performance is better than for baselines. The results for *audio* and *video* channels show only small differences between them. The *textual* channel, however, clearly outperforms the others, as the three best-performing models are all based on text. *Transcription* and *embedded text* channels show also a very similar performance, while the *original text* channel stands out as the best-performing channel overall.

These findings reveal that textual channel is the best selection for prediction, either due to its own information content or because baseline and pre-trained models for this channel are better. Nevertheless, all channels contribute meaningful information, as each surpasses the baseline performance. Baselines refer to the naive *all-negative classification* (majority class) and *all-positive classification* (minority class).

Table 3
Performance of each channel individually for hard evaluation. **MAJORITY** and **MINORITY** are the baselines. **EMB. TEXT** refers to the text that is embedded in the video. **TRANS** refers to transcription extract from video. **ORI. TEXT** refers to the text that was given with the corpus.

METRIC	MAJORITY	MINORITY	AUDIO	VIDEO	EMB. TEXT	TRANS	ORI. TEXT
ICM	-0.4695	-0.5342	-0.2355	-0.2308	-0.0584	-0.0007	**0.2058**
ICM-Norm	0.2649	0.2325	0.3821	0.3845	0.4708	0.4997	**0.6030**
F1	0.3409	0.3225	0.5814	0.5897	0.6467	0.6578	**0.7350**
THRESHOLD	-	-	0.455	0.444	0.485	0.737	0.889

Tables 4 and 5 show performance values of runs sent for hard evaluation and soft evaluation, respectively. *OIQ*, considering all possible channels, achieves the best estimation performance for hard evaluation. Even so, the differences with the other two fusion mechanisms (voting and OIE) are very small, which denotes a great similarity between algorithms. Something similar happens with soft evaluation, where *Logistic Regression fusion* with all possible channel gets the best estimation performance, but again, the results of the other methods are really close to it.

Focusing on hard evaluation –as it was the main emphasis of this work–, we observed that results for the different fusion mechanisms were very close to *original-text-specific* model results. In short, we found that all the decisions made by the fusion mechanism were mainly based on *original-text-specific* information, frequently ignoring the predictions from the audio and video channels. A clear example of that behaviour can be seen in Figure 3, where the *OIE* fusion mechanism relies almost exclusively on the *original-text-specific* model for decision-making. This happens because during threshold adjustment,

Table 4
Performance estimation results, through 5-fold-cross-validation, for hard evaluation. **ALL** means that all textual models are used. **ORIGINAL** means that only textual model trained on original dataset text is used.

FUSION MECHANISM	CONFIGURATION	ICM-Hard	ICM-Hard-Norm	F1
OIQ	ALL	**0.2316**	**0.6100**	**0.7435**
VOTING	ORIGINAL	0.2102	0.6052	0.7363
OIE	ORIGINAL	0.2059	0.6031	0.7351

Table 5
Performance estimation results, through 5-fold-cross-validation, for soft evaluation. **ALL** means that all textual models are used. **ORIGINAL** means that only textual model trained on original dataset text is used.

FUSION MECHANISM	CONFIGURATION	ICM-Soft	ICM-Soft-Norm	CE
Logistic Regression	ALL	**-0.3088**	**0.4459**	**0.8005**
Weight Sum	ALL	-0.3551	0.4378	0.8071
Logistic Regression	ORIGINAL	-0.4404	0.4229	0.8226

audio and video thresholds are set so high that few, if any, examples exceed them, effectively excluding these channels from the final decision.

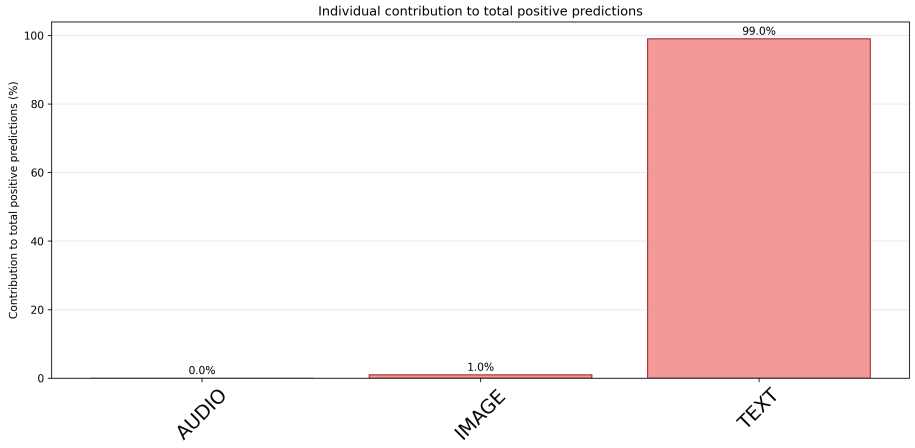

Figure 3: Percentage of cases classified as positive by each model when OIE is applied to *original-configuration*.

4.1. Rankings

Tables 6 and 7 respectively show our final submissions with hard and soft test set for Task 7: Sexism Identification in TikToks. In the hard category, considering all instances, our models were ranked 1^{st}, 3^{rd} and 7^{th}, respectively, out of 41 submissions. In the soft category, considering all instances, our systems were ranked 15^{th}, 17^{th} and 18^{th}, out of 35 submissions.

5. Discussion and Conclusions

This work presented the ECA-SIMM-UVa team's participation in Task 7 of the EXIST 2025 challenge, which focuses on sexism identification in TikTok videos. The overarching goal of the EXIST challenge is to develop automatic detection systems for harmful behaviours like sexism on social platforms, a crucial task given the vast amount of content uploaded. Our approach involved a segmentation-based strategy, splitting TikTok videos into textual, audio, and video channels, based on the hypothesis that sexism can manifest through spoken words, embedded text, speaker's tone, and/or visual content. A significant finding from our experiments is the clear performance superiority of the textual channel as

Table 6
Final results on the test set for hard evaluation of Task 7. **Run 1**: OIQ applied to *all-configuration*. **Run 2**: VOTING applied to *original-configuration*. **Run 3**: OIE applied to *original-configuration*.

Lang	Run	Hard Rank	ICM-Hard	ICM-Hard-Norm	F1
All	1	7	0.1445	0.5730	0.6643
	2	3	0.1827	0.5922	0.6833
	3	1	**0.1984**	**0.6001**	**0.6935**
ES	1	14	0.0176	0.5091	0.5991
	2	6	0.0702	0.5364	0.6228
	3	4	**0.0838**	**0.5435**	**0.6320**
EN	1	6	0.2302	0.6151	0.7095
	2	4	0.2553	0.6277	0.7246
	3	3	**0.2721**	**0.6361**	**0.7353**

.

Table 7
Final results on the test set for soft evaluation of Task 7. **Run 1**: LOGISTIC REGRESSION applied to *all-configuration*. **Run 2**: WEIGHT SUM applied to *all-configuration*. **Run 3**: LOGISTIC REGRESSION applied to *original-configuration*.

Lang	Run	Soft Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
All	1	15	**-0.4517**	**0.4207**	**0.8378**
	2	17	-0.4774	0.4162	0.8387
	3	18	-0.5234	0.4081	0.8475
ES	1	18	-0.8479	0.3494	0.8888
	2	17	-0.8129	0.3556	0.8723
	3	16	**-0.8057**	**0.3569**	**0.8657**
EN	1	16	**-0.2018**	**0.4650**	**0.7959**
	2	17	-0.2799	0.4514	0.8111
	3	19	-0.3711	0.4355	0.8325

compared to audio and video channels. The original text channel, which combines textual transcription and embedded text provided with the dataset, outperformed all other individual channels, including automatically extracted transcriptions and embedded text. This highlights that while all channels contribute meaningful information and individually surpass baseline performance, the textual content appears to be the most informative modality for sexism detection in this context, possibly due to better available models or the inherent nature of the textual information. Same conclusion was obtained in [18], which shows that there has not been great progress in terms of the development of multimodal approaches in the last year. This directly answers one of our research questions, confirming a notable performance gap between pre-trained text, audio, and video models, with text being significantly stronger.

Regarding the fusion mechanisms, for hard evaluation, the Probabilistic OIQ method, considering all possible channels, yielded the best estimation performance, though only marginally better than Majority Voting and One Is Enough (OIE). Similarly, for soft evaluation, Logistic Regression fusion with all channels showed the best performance. However, final results did not follow same order as our estimations, which clearly indicates the inclusion of some type of bias in the estimation during the training phase. We also found that fusion mechanisms performance were really close to *original-text-specific* model results. This circumstance shows us that fusion mechanisms frequently relied almost exclusively on the original-text-specific model's decisions, effectively ignoring predictions from the audio and video channels. This behaviour implicitly addresses our second research question, indicating that the textual channel was deemed disproportionately important during the decision-making process of the fusion models, rather than all three channels being considered equally important.

Despite these observations regarding channel contributions within the fusion mechanisms, our team achieved remarkable results in the official EXIST 2025 challenge. Our competition ranking results demonstrate the effectiveness of our segmentation-based approach and the fine-tuning of specialized

deep learning models.

The findings underscore the challenges in multimodal sexism detection, particularly the comparatively underdeveloped research in video automated detection of sexism due to data collection difficulties and high resource requirements. While our textual models leveraged robust pre-trained architectures like XLM-RoBERTa-Large, the performance and integration issues with audio (Wav2Vec-Large-XLRS-53) and video (ViViT-b-16x2-Kinetics400) models suggest areas for future improvement. Future work should focus on improving audio and video models under a mutual reinforcement learning strategy. These models are crucial for advancing this field, especially as platforms like TikTok continue to be significant vectors for potentially harmful content.

Acknowledgments

This work was carried out in the Project PID2021-126315OB-I00 that was supported by MCIN / AEI / 10.13039/501100011033 / FEDER, EU. Also, this work is partially funded by the Spanish Ministry of Science, Innovation and Universities (project FairTransNLP PID2021-124361OB-C32) funded by MCIN/AEI/10.13039/501100011033.

Declaration on Generative AI

During the preparation of this work, the author(s) used NotebookLM in order to: Drafting content and Abstract drafting. Further, the author(s) used GPT-4 in order to: Improve writing style and 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. Plaza, J. Carrillo-de Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, Exist 2025: Learning with disagreement for sexism identification and characterization in tweets, memes, and tiktok videos, in: European Conference on Information Retrieval, Springer, 2025, pp. 442–449.

[2] L. Plaza, J. C. de Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, Overview of exist 2025: Learning with disagreement for sexism identification and characterization in tweets, memes, and tiktok videos, in: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. G. S. de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, Springer, 2025.

[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, 2025.

[4] 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.

[5] 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.

[6] L. Plaza, J. Carrillo-de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2023–learning with disagreement for sexism identification and characterization, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2023, pp. 316–342.

[7] 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.

[8] 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.

[9] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, Automatic classification of sexism in social networks: An empirical study on twitter data, IEEE Access 8 (2020) 219563–219576.

[10] P. Chiril, V. Moriceau, F. Benamara, A. Mari, G. Origgi, M. Coulomb-Gully, An annotated corpus for sexism detection in french tweets, in: 12th Conference on Language Resources and Evaluation (LREC 2020), ELRA: European Language Resources Association, 2020, pp. 1–7.

[11] A. Karimi, L. Rossi, A. Prati, AEDA: An easier data augmentation technique for text classification, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 2748–2754. URL: https://aclanthology.org/2021.findings-emnlp.234/. doi:10.18653/v1/2021.findings-emnlp.234.

[12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 4171–4186.

[13] 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).

[14] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, 2021. URL: https://arxiv.org/abs/2006.03654. arXiv:2006.03654.

[15] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. URL: https://arxiv.org/abs/2302.13971. arXiv:2302.13971.

[16] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, 2023. URL: https://arxiv.org/abs/2310.06825. arXiv:2310.06825.

[17] L. D. 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, 2025. URL: https://arxiv.org/abs/2504.11169. arXiv:2504.11169.

[18] I. Arcos, P. Rosso, Sexism identification on tiktok: a multimodal ai approach with text, audio, and video, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2024, pp. 61–73.

[19] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, Advances in neural information processing systems 33 (2020) 12449–12460.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[21] C. S. Wu, U. Bhandary, Detection of hate speech in videos using machine learning, in: 2020 international conference on computational science and computational intelligence (CSCI), IEEE, 2020, pp. 585–590.

[22] M. Das, R. Raj, P. Saha, B. Mathew, M. Gupta, A. Mukherjee, Hatemm: A multi-modal dataset for hate video classification, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 17, 2023, pp. 1014–1023.

[23] H. Wang, T. R. Yang, U. Naseem, R. K.-W. Lee, Multihateclip: A multilingual benchmark dataset for hateful video detection on youtube and bilibili, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 7493–7502.
[24] F. T. Boishakhi, P. C. Shill, M. G. R. Alam, Multi-modal hate speech detection using machine learning, in: 2021 IEEE International Conference on Big Data (Big Data), 2021, pp. 4496–4499. doi:10.1109/BigData52589.2021.9671955.
[25] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The hateful memes challenge: Detecting hate speech in multimodal memes, Advances in neural information processing systems 33 (2020) 2611–2624.
[26] R. Velioglu, J. Rose, Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge, arXiv preprint arXiv:2012.12975 (2020).
[27] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, Advances in neural information processing systems 32 (2019).
[28] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, K.-W. Chang, Visualbert: A simple and performant baseline for vision and language, arXiv preprint arXiv:1908.03557 (2019).
[29] J. Lu, C. Clark, R. Zellers, R. Mottaghi, A. Kembhavi, Unified-io: A unified model for vision, language, and multi-modal tasks, arXiv preprint arXiv:2206.08916 (2022).
[30] Y. Li, J. Duan, Z. Qu, Tri-robust learning: Robust multi-neural networks against extremely noisy labels, Available at SSRN 4911734 (????).
[31] A. Shrotriya, A. K. Sharma, A. K. Bairwa, R. Manoj, Hybrid ensemble learning with cnn and rnn for multimodal cotton plant disease detection, IEEE Access (2024).
[32] M. Bain, J. Huh, T. Han, A. Zisserman, Whisperx: Time-accurate speech transcription of long-form audio, arXiv preprint arXiv:2303.00747 (2023).
[33] H. Lu, W. Liu, B. Zhang, B. Wang, K. Dong, B. Liu, J. Sun, T. Ren, Z. Li, H. Yang, et al., Deepseek-vl: towards real-world vision-language understanding, arXiv preprint arXiv:2403.05525 (2024).
[34] E. Amigó, F. Giner, J. Gonzalo, F. Verdejo, On the foundations of similarity in information access, Information Retrieval Journal 23 (2020) 216–254.
[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, arXiv preprint arXiv:1911.02116 (2019).
[36] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, M. Auli, Unsupervised cross-lingual representation learning for speech recognition, arXiv preprint arXiv:2006.13979 (2020).
[37] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, C. Schmid, Vivit: A video vision transformer, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 6836–6846.
[38] D. Kraft, A Software Package for Sequential Quadratic Programming, Technical Report DFVLR-FB 88-28, Institut für Dynamik der Flugsysteme, Deutsche Forschungs- und Versuchsanstalt für Luft- und Raumfahrt (DLR), Cologne, Germany, 1988.