UMUTeam at EXIST2026: Multimodal Stacking Architecture for Sexism Detection in Short Videos

Notebook for the EXIST Lab at CLEF 2026

Alba de-Rojas-Ortuño[1,*], Ronghao Pan[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 for the EXIST 2026 shared task at CLEF 2026, focusing on sexism detection in TikTok short videos. Recognizing the complexity of this content, we propose a multimodal framework combining text, audio, and video modalities through Late Fusion strategies. For the textual modality, we employ TF-IDF with character n-grams as a feature extractor. For the audio modality, traditional Mel-Frequency Cepstral Coefficients (MFCCs) are used as acoustic features. For the video modality, a SigLIP architecture serves as a visual feature extractor. Each modality is independently combined with a traditional machine learning classifier. We addressed two subtasks exclusively under the hard-hard evaluation variant: Task 3.1 (Binary Sexism Identification), combining text and audio modalities, and Task 3.2 (Source Intention Classification), combining text and video modalities, both through Stacking-based Late Fusion. This system was developed as part of a Final Degree Project and faced certain constraints, such as being trained on a subset of the official lab data and excluding the newly introduced physiological features. Due to these limitations, the system obtained modest results compared to other participants, ranking 70th out of 73 for Task 3.1 and 64th out of 68 for Task 3.2.

Keywords

Sexism Detection, Multimodal Learning, Natural Language Processing, Late Fusion, Hard-Label Classification

1. Introduction

Gender-based discrimination remains a widespread problem in contemporary society. At its core, sexism involves treating people unfairly because of their sex, often stemming from deeply ingrained assumptions about what each gender is capable of or deserves. These attitudes play out in very different ways — sometimes as outright hostility or abuse, and other times as everyday habits and expectations that go largely unnoticed but still hold people back. Women tend to bear the brunt of this, and the rise of social media has made things worse, giving sexist content new platforms to reach wider audiences [1].

The classification of sexism is important to maintain a safer and more equitable social environment, as it helps us understand how prejudice manifests linguistically and behaviorally [2].

Short-form videos have established themselves as the dominant format on social media platforms, especially on TikTok, which is built entirely around this format. Their virality amplifies the spread of harmful content while simultaneously hindering its automatic detection. This is because, unlike traditional formats, the message does not reside in a single modality but is distributed simultaneously across text, audio, and video.

Given the importance of detecting harmful content in a format that is consumed daily by millions of users, some initiatives are emerging that encourage users to identify this type of content, focusing on specific domains. This is the case with the EXIST (sEXism Identification in Social neTworks) competition, which, in its 2025 edition, proposed for the first time the task of detecting sexism in TikTok videos as part of the CLEF laboratories [1].

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

a.derojasortuno@um.es (A. de-Rojas-Ortuño); ronghao.pan@um.es (R. Pan); valencia@um.es (R. Valencia-García)

0009-0002-8086-8725 (A. de-Rojas-Ortuño); 0009-0008-7317-7145 (R. Pan); 0000-0003-2457-1791 (R. Valencia-García)

In this context, this paper describes the system developed by UMUTeam for this edition of EXIST as part of a Final Degree Project, with the main objective of exploring classification techniques for the detection of harmful content in short videos, using sexism detection as a benchmark case within the EXIST 2026 shared task framework. We participated in two subtasks involving TikTok video content: Task 3.1 (Binary Sexism Identification) and Task 3.2 (Source Intention Classification), exclusively under the hard-hard evaluation variant. For the textual modality, we employed TF-IDF with character n-grams (range 3–5) [3] as a feature extractor combined with a Support Vector Machine (SVM) classifier. For the audio modality, Mel-Frequency Cepstral Coefficients (MFCCs) [4] were extracted and fed into a Multilayer Perceptron (MLP) classifier. For the video modality, a SigLIP [5] architecture was used as a visual feature extractor, integrated with a Long Short-Term Memory (LSTM) network and an SVM for final classification. The unimodal representations were then combined through Late Fusion strategies: for Task 3.1, text and audio features were fused via a Stacking meta-classifier with an MLP as meta-learner; for Task 3.2, text and video features were combined using the same framework but with Logistic Regression (LR) as meta-learner.

The remainder of this paper is organized as follows. Section 2 reviews the related work. Section 3 describes the dataset and the tasks evaluated. Section 4 presents the methodology. Section 5 reports and analyzes the results. Finally, Section 6 outlines conclusions and future work.

2. Related Work

Research into the automatic classification of social media content has historically progressed along unimodal lines, analysing text messages, visual aesthetics and the presence of faces in videos, or musical features and speech patterns, independently. Although these unimodal approaches have yielded valuable results, they provide an incomplete view of current content generation, which involves multimodal content [6].

The shift towards multimodal approaches has been gradual and presents significant limitations. Of the few existing works, many combine only two modalities, typically text and image, and do so using simple fusion strategies that do not allow for an analysis of how each modality contributes individually or how they interact with one another [6]. Added to this is the difficulty of constructing high-quality annotated *datasets* for audiovisual content, the creation of which requires considerable manual effort and whose data quickly becomes outdated given the constant evolution of these platforms [6, 7].

In the specific domain of harmful content detection in video, research is still scarce. Das et al. [7] presented HateMM, one of the first multimodal resources for classifying hate speech videos, integrating text, audio and images together. Lang et al. [8] proposed MoRE, a system for detecting harmful content in short videos that demonstrates that treating all modalities equally in fusion is suboptimal, requiring the contribution of each modality must be dynamically adapted according to the content of each video. MM-HSD [9] explores fusion mechanisms based on *Cross-Modal Attention* for hate speech videos, noting that the modalities are not always individually informative and that simple fusion methods do not adequately capture the dependencies between them. In the specific context of TikTok, Arcos and Rosso [10] proposed a trimodal system for detecting sexism on this platform, achieving improvements over unimodal systems.

Complementary studies have also explored aspects that are relevant to the present study. In the area of speech-based affective computing, *speech-emotion* [11] introduces a multilingual and multimodal toolkit for emotion recognition from speech, integrating speech processing, transcription, text encoding, and fusion mechanisms within a unified framework. Although focused on emotion recognition rather than sexism detection, this work is relevant because it demonstrates the usefulness of combining acoustic and textual signals in speech-centred classification tasks. In the field of harmful content analysis, Spanish MTLHateCorpus 2023 [12] presents a fine-grained Spanish corpus for hate speech detection based on multi-task learning, addressing speech type, target, target group, and intensity. This resource highlights the importance of modelling harmful online content beyond binary classification, which is closely related to the fine-grained nature of sexism and intention detection addressed in the present work.

Sexism presents additional challenges compared to other forms of hate speech. Its detection requires an understanding not only of the explicit content of the message, but also of implicit attitudes and stereotypes that can manifest in very subtle ways. Consequently, the proposed classification tackles a greater challenge that will enable the recognition of nuances in short videos, with the potential to be generalised to other concepts that may be easier to identify.

3. Dataset

The dataset used in this work corresponds to the one provided by the organizers of the EXIST 2026 shared task [2] [13], which focuses on the detection of sexism in TikTok short videos and introduces physiological data as a new modality. It comprises a total of 3198 videos, split into a training set of 2524 samples and a test set of 674 samples. The videos are multilingual, covering both Spanish and English content. In this work, we address Task 3.1 and Task 3.2 exclusively.

For the textual modality, the dataset includes the text field, which contains the automatic transcription of the video (ASR), the description without hashtags, and the text on the screen (OCR). For the video and audio modalities, the organizers provide the videos directly, from which the corresponding audio can be extracted easily. For the sake of simplicity, we have chosen to consolidate the labels into a single category per sample using the absolute majority criterion, thereby adopting a hard-labelling approach. During an initial examination of the training dataset, we identified 2524 samples corresponding to videos, without null values, incomplete records, or incorrectly loaded data types.

About the proportions of the different classes across the two tasks evaluated, we observed that for Task 3.1, the frequency of the YES and NO labels is fairly even, as shown in Figure 1.

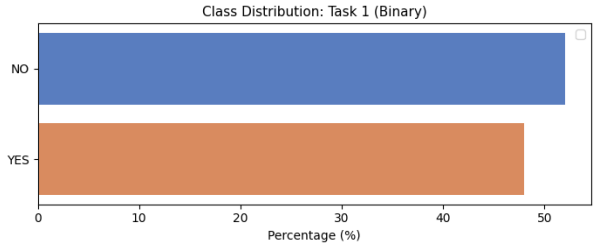


Figure 1: Distribution of the labels for Task 3.1

Figure 2 shows the label distribution for Task 3.2. The NO label remains consistent with Task 3.1, whilst sexist videos are split between DIRECT and JUDGEMENTAL. The former encompasses content with explicit messages such as insults or misogynistic attacks, whilst the latter covers more subtle moral judgements regarding women's behaviour. DIRECT is slightly more frequent than JUDGEMENTAL, indicating that explicit sexism predominates slightly in the corpus. This may also be an advantage for classification, as direct examples are easier to identify due to their explicit nature.

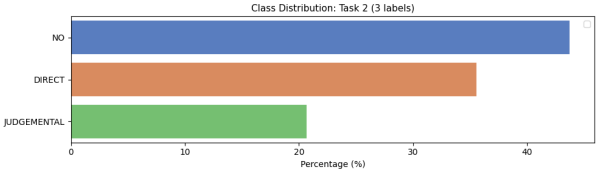
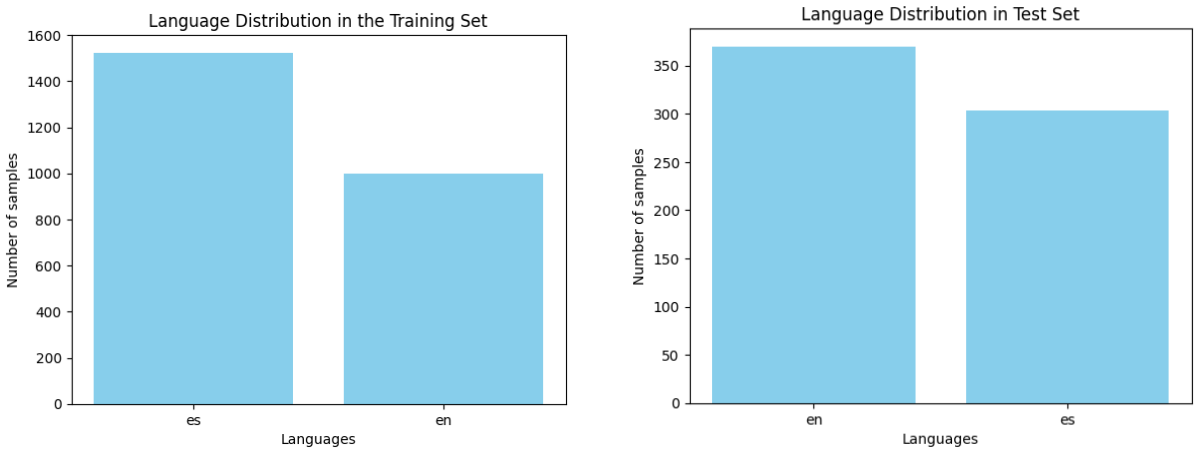


Figure 2: Distribution of the labels for Task 3.2

Regarding the language distribution, both the training and test sets maintain a similar proportion of Spanish and English videos, with Spanish being the majority language in both cases, as illustrated in Figures 3a and Figure 3b.



(a) Language Distribution in Training Set

(b) Language Distribution in Test Set

Figure 3: Language Distribution on the Training and Test Set

The duration of the clips was analysed to determine whether special processing would be required. The data show that the average duration is 27 seconds, ranging from 5 seconds to approximately 1 minute, as illustrated in Figure 4. A notable spike is observed in the 55 to 60-second range, likely due to many TikTok audio clips reaching that maximum duration. Overall, no special treatment or complex long-term memory management was needed for the content.

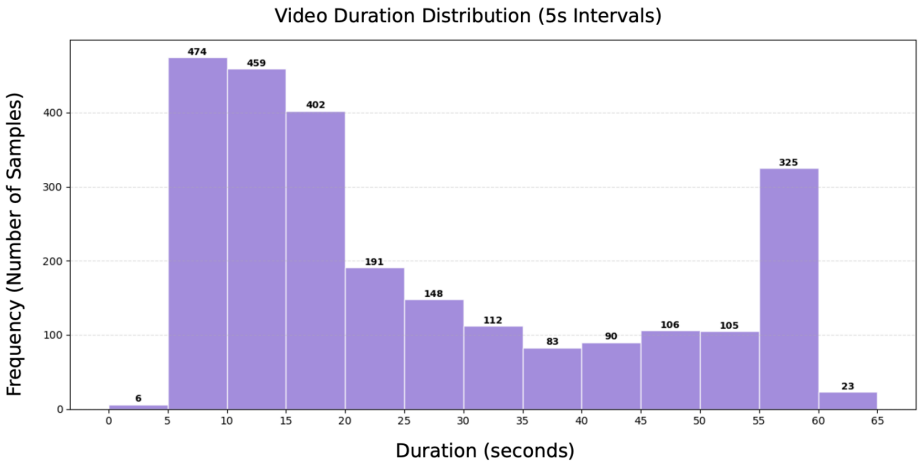


Figure 4: Video Duration Distribution (5s Intervals)

Next, we applied minimal textual preprocessing on the text field. This preprocessing consisted of lowercase conversion and the removal of redundant spaces, with the objective of reducing superficial variations. We prioritised the conservation of the original content over more aggressive cleaning techniques, so as not to lose the meaning of the text. For example, consideration was given to spell-checking, but applying text correction could strip the comments of their informal essence. This principle also applies to punctuation marks and stop-words, as these reflect intentions and feelings that are fundamental to the analysis of sexism. Regarding the average length of the text strings, it is around 596 characters in total. Assuming that an average length of 596 characters suggests that most texts fall below the usual limit of 512 tokens in the Transformer models used, truncation was retained as a safety mechanism. No additional processing was applied to the audio accompanying these videos.

4. Methodology

This section describes the models, training strategies, and evaluation procedures applied to Task 3.1 and Task 3.2 defined in EXIST 2026. The general pipeline proposed for both tasks is illustrated in Figure 5. It is based on a Late Fusion strategy in which each modality is independently processed through a feature extraction stage followed by a dedicated classifier, yielding a set of prediction probabilities for the evaluated labels. These unimodal predictions are then combined to produce the final output.

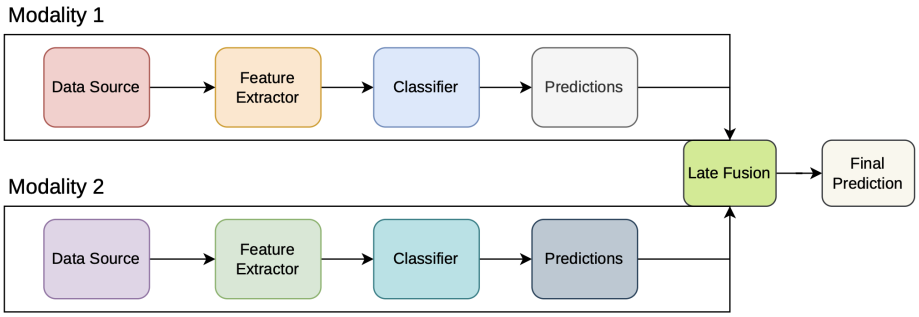


Figure 5: General pipeline of the proposed multimodal Late Fusion system.

As shown in the figure, both tasks employ bimodal combinations. Specifically, Task 3.1 combines text and audio modalities, while Task 3.2 combines text and video modalities.

4.1. Text Modality

For the text modality, which is shared across both tasks, we used TF-IDF with character n-grams (3–5 grams) as a feature extractor, combined with a Support Vector Machine (SVM) classifier. The same TF-IDF vectorizer, fitted on the training set, was applied to both tasks. The SVM was configured with `class_weight='balanced'` to account for class imbalance and `probability=True` to enable probability estimation, with the remaining hyperparameters set to the default values of `scikit-learn`.

The use of TF-IDF with character n-grams is justified by its ability to capture morphological and sublexical patterns, which are particularly useful for detecting the informal language and misspellings typical of social media content. Preliminary experiments conducted as part of this work (summarized in Table 1) showed that TF-IDF with character n-grams outperformed Transformer-based representations for these two tasks, while also reducing computational cost. It is worth noting that these preliminary trials were evaluated on a constrained subset of 1,543 training samples due to technical issues with video downloads encountered within the framework of the degree thesis, prior to official registration and data acquisition from the competition.

In this limited data regime, deep architectures are more susceptible to suboptimal behavior. As evidenced in Table 1, for the binary scenario (Task 3.1), our proposed character-level TF-IDF + SVM performs virtually on par with the frozen XLM-RoBERTa feature extractor (0.6693 vs. 0.6701 in macro F_1-score), yielding a negligible difference under 0.1%. Meanwhile, for the three-class scenario (Task 3.2), the TF-IDF approach significantly outperforms all Transformer alternatives, including the fine-tuned XLM-RoBERTa model (0.5682 vs. 0.5402). This performance gap is likely due to the short and noisy nature of the text field in the dataset and the fact that the tasks do not require deep semantic understanding. Consequently, the statistical approach provides a highly competitive performance while drastically minimizing computational overhead, making it the most robust choice for our final pipeline.

Combined with an SVM classifier, this approach is well-suited for high-dimensional feature spaces, where SVMs are known for their robustness against overfitting [14, 15].

Table 1
Preliminary text modality experiments (Validation Set) for Task 3.1 and Task 3.2.

		Task 3.1		Task 3.2	
Extractor	**Classifier**	**Acc**	**F1-M**	**Acc**	**F1-M**
TF-IDF (Char 3–5)	**SVM**	**0.6776**	**0.6693**	0.5628	**0.5682**
TF-IDF (N-grams)	SVM	0.6503	0.6498	**0.5656**	0.5544
BERT	MLP	0.6000	0.5954	0.4767	0.4565
XLM-RoBERTa	MLP	0.6712	0.6701	0.5370	0.5320
DistilBERT	MLP	0.5918	0.5886	0.4411	0.4186
Fine-Tuning XLM-RoBERTa	—	0.6685	0.6636	0.5589	0.5402

4.2. Audio Modality

For the audio modality, the audio files were first synchronized with their corresponding text samples by matching each video identifier to its associated MP3 file. Feature extraction was performed using Mel-Frequency Cepstral Coefficients (MFCCs) [4], extracting 13 coefficients per frame using the original sampling rate of each file (`sr=None`). The first 13 coefficients are retained, where C0 represents the mean energy of the signal and C1–C12 capture the spectral shape of the audio. The resulting MFCC matrix was averaged over time to obtain a fixed-length feature vector of 13 dimensions per sample. MFCCs were chosen as the acoustic feature extractor as they are among the most widely used features for speech-related tasks, capturing the sensitivity of human auditory perception to frequencies through a compact spectral representation [16].

As with the textual modality, these acoustic models were tuned and selected based on preliminary experiments (summarized in Table 2) trained on the constrained subset of 1,543 samples available from the initial data-collection phase. In this low-data regime, the MFCC + MLP pipeline proved to be highly robust, achieving an accuracy of 0.5979 and a macro F_1-score of 0.6027 for Task 3.1 and an accuracy of 0.4314 and a macro F_1-score of 0.4658 for Task 3.2. This setup systematically outpaced more complex end-to-end architectures and massive pre-trained audio transformers like Wav2Vec2-BERT 2.0, which severely suffered from overfitting and failed to converge effectively given the limited number of training samples. Consequently, the combination of MFCC features and a non-linear MLP offered the most consistent and data-efficient baseline to generate reliable prediction probabilities for the subsequent late fusion stages.

Before classification, features were standardized using `StandardScaler`, fitting the scaler to the training set and applying it to both training and test sets to avoid data leakage. The classifier used was a Multilayer Perceptron (MLP) with two hidden layers of 512 and 256 units respectively, trained for a maximum of 150 iterations. An MLP classifier was chosen given its ability to model complex, non-linear relationships between MFCC coefficients through hidden layers with non-linear activation functions, which is crucial since acoustic features are highly correlated and rarely linearly separable.

Table 2
Preliminary audio modality experiments (Validation Set) for Task 3.1 and Task 3.2.

		Task 3.1		Task 3.2	
Extractor	**Classifier**	**Acc**	**F1-M**	**Acc**	**F1-M**
Pat. Paralinguistic ($sr = 16000$)	SVM	0.5918	0.5900	0.4247	0.4193
MFCC ($sr = None$)	**MLP**	**0.5979**	**0.6027**	**0.4314**	**0.4658**
MFCC	CNN	0.5223	0.5233	0.2436	0.3973
MFCC	CNN+LSTM	0.5217	0.5233	0.3890	0.2753
Wav2Vec2-BERT 2.0 (Frozen)	MLP	0.5425	0.3517	0.4493	0.2067
Fine-Tuning Wav2Vec2-BERT 2.0	—	0.5425	0.3517	0.4493	0.2067

4.3. Video Modality

For the video modality, the video files were synchronized with their corresponding text samples by matching each video identifier to its associated MP4 file, following the same approach as for the audio modality. Corrupted or unreadable videos were assigned a zero vector of 768 dimensions as a fallback mechanism. For feature extraction, frames were sampled at 1 frame per second (1 FPS) to reduce computational cost while retaining sufficient temporal information. Each frame was processed using SigLIP [5] (`google/siglip-base-patch16-224`), a vision-language model that extends CLIP by replacing the contrastive loss with a sigmoid-based loss, producing a 768-dimensional embedding per frame. SigLIP was selected due to its contrastive text-image training, which enables the model to capture rich semantic context from visual content, combined with its improved image-text alignment through sigmoid loss, which produces more discriminative visual representations [5].

The resulting sequence of frame embeddings was padded to a maximum length of 60 frames using post-padding. This sequence was then fed into an LSTM [17] network with 256 units, which summarizes the temporal information into a single fixed-length vector. The LSTM was chosen to model the temporal dependencies between frames, as it is specifically designed to capture long-range sequential patterns [17]. The LSTM encoder output was extracted as the video representation and standardized using `StandardScaler` before being passed to an SVM classifier with RBF kernel, `C=1`, and `class_weight='balanced'`.

As summarized in Table 3, the SigLIP-based architectures consistently achieved the highest performance across both tasks. The SigLIP + LSTM sequence reached a macro F_1-score of 0.6862 when paired with an MLP for Task 3.1, and 0.5745 when paired with an RBF-SVM for Task 3.2. Conversely, complex end-to-end architectures and video-level frameworks like X-CLIP suffered from severe overfitting, showing early saturation during training due to the constrained data regime. Consequently, leveraging SigLIP as a robust frozen visual extractor coupled with sequential modeling provided the most reliable foundation for the multimodal fusion stages.

Table 3
Preliminary video modality experiments (Validation Set) for Task 3.1 and Task 3.2.

		Task 3.1		Task 3.2	
Extractor	**Aggregation / Classifier**	**Acc**	**F1-M**	**Acc**	**F1-M**
ResNet-50	Mean Pooling + MLP	0.6055	0.6022	0.4411	0.4033
Efficient Net-B4	Mean Pooling + MLP	0.5890	0.5821	0.4603	0.4271
ViT	Mean Pooling + MLP	0.6055	0.6036	0.4932	0.4778
SigLIP	**LSTM + MLP**	**0.6877**	**0.6862**	0.5589	0.5506
SigLIP	**LSTM + SVM**	0.5753	0.5745	**0.5753**	**0.5745**
SigLIP	Mean Pooling + MLP	0.6274	0.6235	0.5479	0.5379
Fine-Tuning X-CLIP	—	0.6247	0.6196	0.4904	0.4595

4.4. Late Fusion

For the final classification of the short videos, we implemented bimodal combinations for the two tasks. Our development tests revealed that the trimodal combinations yielded worse performance than the bimodal ones. This demonstrates that when all three modalities are combined simultaneously, they interfere with each other, consequently introducing noise that harms the classification. Regarding the specific Late Fusion strategy chosen, we observed that, for both tasks, the Stacking meta-classifier outperformed or performed virtually on par with simpler paradigms, such as Hard or Soft Voting. For this reason, we opted for this approach.

For Task 3.1, the fusion framework integrates the textual and acoustic modalities. This bimodal combination was chosen because text and audio provided the highest synergy during development, whereas adding the visual modality consistently dropped the macro F_1-score. The meta-features consist of the prediction probabilities yielded by the character-level TF-IDF + SVM pipeline and the MFCC +

MLP model. A Multilayer Perceptron (MLP) was selected as the meta-learner for this task, configured with the default `scikit-learn` hyperparameters and trained to map the bimodal probability space into the final binary classes (YES and NO). The pipeline is shown in the Figure 6 below.

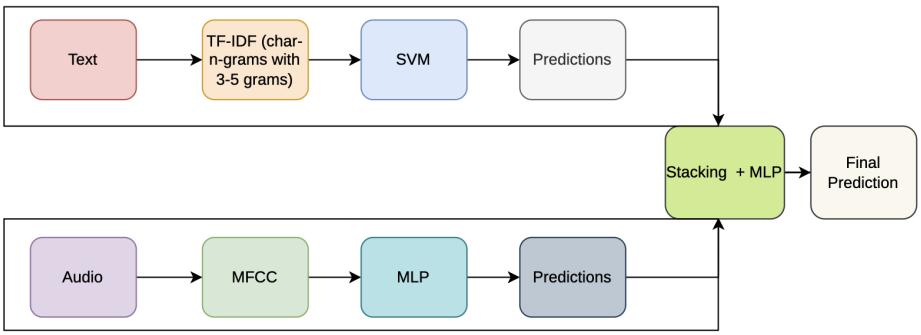


Figure 6: Pipeline of the proposed multimodal Stacking architecture for Task 3.1.

For Task 3.2 (source intention classification), the system combines the textual and visual modalities, as video features proved to be highly discriminative for intention nuances during our development phase, while the acoustic modality introduced noise. In this scenario, the meta-features are derived from the character-level TF-IDF + SVM and the SigLIP + LSTM + SVM models. A Logistic Regression (LR) model with default `scikit-learn` hyperparameters was employed as the meta-learner. Logistic Regression was selected for this three-class problem due to its better performance based on its simplicity and its ability to act as a robust linear combination layer, minimizing the risk of overfitting in the meta-learning stage. The pipeline is shown in the Figure 7.

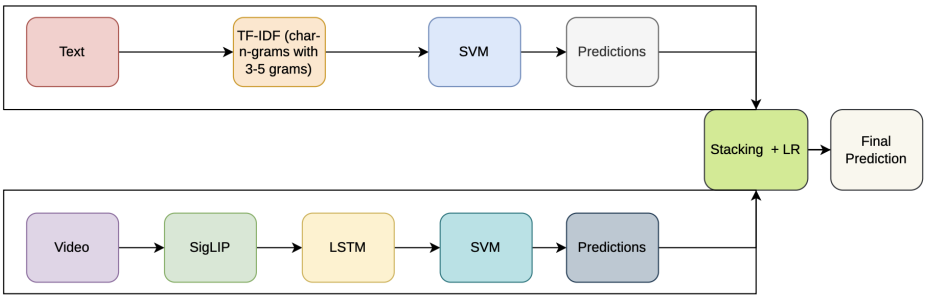


Figure 7: Pipeline of the proposed multimodal Stacking architecture for Task 3.2.

To ensure robust generalization and prevent data leakage during the late fusion training, the meta-learners were trained using a 5-fold stratified cross-validation scheme on the meta-features. The resulting ensembled pipeline aligns with our core objective of maintaining a lightweight yet mathematically sound multimodal architecture capable of capturing multi-source signals without the prohibitive computational overhead of end-to-end deep multimodal networks.

4.5. Hyperparameter Configuration

To enhance the reproducibility of this work, Table 4 summarizes the key hyperparameters used for each base classifier and meta-learner described in Sections 4.1–4.4. Unless otherwise specified, all remaining hyperparameters were kept at their `scikit-learn` default values.

Table 4
Key hyperparameters for the base classifiers and meta-learners used in the proposed pipeline.

Component	Model	Key hyperparameters
Text (3.1 & 3.2)	SVM	`kernel='rbf'`, `C=1.0`, `gamma='scale'`, `class_weight='balanced'`, `probability=True`, `random_state=42`
Audio (3.1)	MLP	`hidden_layer_sizes=(512,256)`, `activation='relu'`, `solver='adam'`, `max_iter=150`, `alpha=0.0001`, `random_state=42`
Video (3.2)	LSTM encoder	`lstm_units=256`, `dense_units=256`, `dropout=0.4`, Adam `learning_rate=0.0001`, `epochs=15`, `batch_size=16`, early stopping (`patience=3`, monitored on validation loss), sequences padded/ truncated to 60 frames
Video (3.2)	SVM	`kernel='rbf'`, `C=1`, `gamma='scale'`, `class_weight='balanced'`, `probability=True`, `random_state=42`
Meta-learner T+A (3.1)	MLP	`random_state=42`
Meta-learner T+V (3.2)	Logistic Regression	`class_weight='balanced'`, `random_state=42`
Stacking scheme	—	5-fold stratified CV on meta-features

5. Results

This section details the official evaluation of the models developed for the EXIST 2026 shared task by UMUTeam for Task 3.1 and Task 3.2. Consistent with the hard-hard paradigm, the systems output discrete labels instead of probability distributions. These predictions are then validated against a consolidated target baseline derived from majority voting using specific probabilistic thresholds for each subtask. The systems are evaluated upon the official metric, which is the Information Contrast Measure (ICM), which generalizes pointwise mutual information to evaluate how closely the predicted distributions match the reference data. In the tables, the Macro F_1 metric is also included.

Before analyzing each task individually, it is worth noting that ICM-Hard scores are negative across all configurations. Within the information-theoretic framework of this metric [18], this indicates that the cumulative penalty of false positives and false negatives marginally outweighs the correct predictive updates. Despite this, the proposed system consistently outperforms both official baselines across all tasks and languages, indicating that the model captures meaningful patterns from the multimodal data rather than defaulting to trivial predictions.

5.1. Modality Contribution Analysis

Since gold labels for the official test set are not released to participants in the EXIST shared task, the comparative analysis in this section is instead computed on the training set (2,524 samples) using an outer 5-fold stratified nested cross-validation scheme: in each outer fold, both the unimodal base classifiers and the stacking meta-classifier are retrained from scratch on the remaining folds, and predictions are collected only for the held-out fold. This yields held-out predictions for every training sample without any risk of data leakage, allowing us to directly compare the individual contribution of each modality against the final fused system under identical conditions.

Table 5 reports this comparison for Task 3.1. The textual modality alone already accounts for most of the fusion's performance, while the acoustic modality performs considerably worse in isolation. The Stacking-based fusion yields only a marginal improvement over the text-only model (+0.0069 in macro F1), indicating that, for this task, audio contributes limited complementary information beyond what is

Table 5
Comparison of individual modalities against the bimodal fusion for Task 3.1, evaluated via nested cross-validation on the training set.

Modality	Accuracy	F_1-macro
Text (TF-IDF + SVM)	0.6589	0.6507
Audio (MFCC + MLP)	0.5325	0.5302
Fusion T+A (Stacking + MLP)	**0.6601**	**0.6576**

already captured by the textual signal.

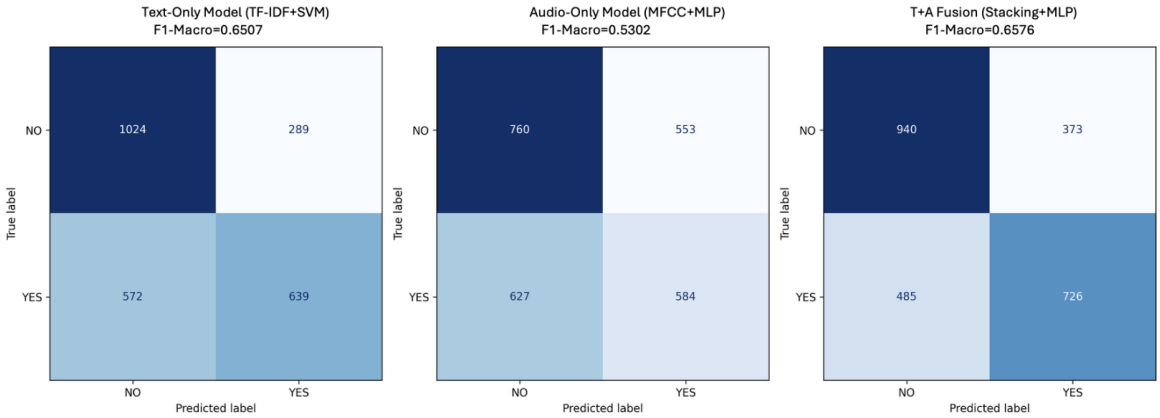


Figure 8: Confusion matrices for Task 3.1: text-only, audio-only, and bimodal fusion, evaluated via nested cross-validation on the training set.

Figure 8 shows that the audio-only model performs close to random for the YES class, misclassifying it as NO in nearly half of the cases, whereas the fusion improves recall on the YES class at a small cost to NO precision.

Table 6 and Figure 9 present the equivalent analysis for Task 3.2. In this case, the video modality alone clearly outperforms the text-only model, and the fusion yields a further, more substantial gain over the best individual modality (+0.0115 in macro F1 over video alone), supporting the choice of combining text and video for this task. The confusion matrices reveal that the text-only model struggles to distinguish DIRECT from NO, whereas the video modality is comparatively stronger at identifying JUDGEMENTAL content. The fusion improves further still on the video-only model, correctly classifying additional NO and JUDGEMENTAL instances.

Table 6
Comparison of individual modalities against the bimodal fusion for Task 3.2, evaluated via nested cross-validation on the training set.

Modality	Accuracy	F_1-macro
Text (TF-IDF + SVM)	0.5368	0.4948
Video (SigLIP + LSTM + SVM)	0.6145	0.6075
Fusion T+V (Stacking + LR)	**0.6252**	**0.6190**

Note that the scores reported in this subsection are higher than the official results in Tables 7 and 8. This is expected, since here we evaluate on our own training data, whereas the official tables report performance on the competition's test set, which is different and presumably more challenging. This subsection is not meant to replace those official results, but rather to compare modalities fairly against

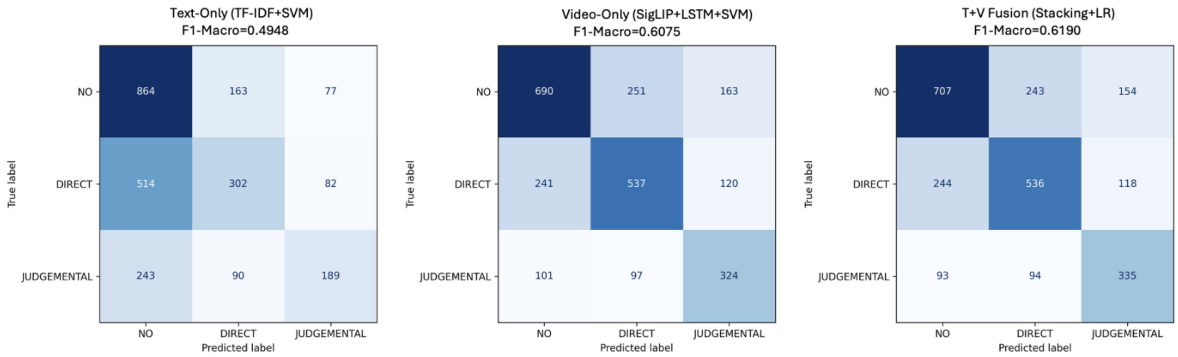


Figure 9: Confusion matrices for Task 3.2: text-only, video-only, and bimodal fusion, evaluated via nested cross-validation on the training set.

each other.

5.2. Results of Task 3.1

Table 7 presents the comparative evaluation of our Stacking-based Late Fusion pipeline for Task 3.1 (Binary Sexism Identification) against the official baselines and the top-performing configurations, including our final standing on the global leaderboard and covering both overall results and language-specific benchmarks for English and Spanish.

Table 7
Official evaluation results for Task 3.1 under the Hard-Hard paradigm, including global (ALL), Spanish (ES), and English (EN) benchmarks with ICM-Hard and F1-score metrics.

	Hard vs Hard ALL			Hard vs Hard ES			Hard vs Hard EN		
Team / System	Rank	ICM-H	F_1	Rank	ICM-H	F_1	Rank	ICM-H	F_1
EXIST2025-test_gold	0	0.9907	1.0000	0	0.9631	1.0000	0	0.9997	1.0000
Top-1 Submissions	1	0.3307	0.7627	1	0.2662	0.7177	1	0.3967	0.7956
UMUTeam_1	**70**	**-0.1237**	**0.5644**	**70**	**-0.2680**	**0.5251**	**70**	**-0.0079**	**0.5974**
baseline majority class	74	-0.4244	0.0000	72	-0.3708	0.0000	74	-0.4836	0.0000
baseline minority class	75	-0.6036	0.6117	75	-0.7411	0.5558	75	-0.5174	0.6545

In Task 3.1, a considerable margin exists between the F_1-score of 0.7627 achieved by the top-1 submission on the global dataset and the 0.5644 attained by our architecture. While our model demonstrates a notable improvement within the English subset, reaching an F_1-score of 0.5974, a parallel upward trend is observed in the top-1 English submission, which scales to 0.7956. Consequently, this performance gap remains virtually unchanged between both evaluation setups. Nevertheless, this localized improvement in our absolute results highlights a higher algorithmic affinity for English linguistic structures, due to the higher statistical discriminative power of TF-IDF features over the English vocabulary distribution, coupled with a more consistent acoustic variance captured by MFCCs across English-speaking instances.

5.3. Results of Task 3.2

Following the evaluation paradigm described for Task 3.1, Table 8 displays the official performance metrics obtained by our system for Task 3.2 (Source Intention Classification).

Regarding Task 3.2, a substantial performance gap is observed when comparing our framework against the top-performing system. Globally, the top-1 submission achieved a F_1-score of 0.7065, whereas our model attained 0.5024, representing a distinct margin of 0.2041. Furthermore, a notable discrepancy

Table 8
Official evaluation results for Task 3.2 under the Hard-Hard paradigm, including global (ALL), Spanish (ES), and English (EN) benchmarks with ICM-Hard and Macro F_1 metrics.

	Hard vs Hard ALL			Hard vs Hard ES			Hard vs Hard EN		
Team / System	Rank	ICM-H	Macro F_1	Rank	ICM-H	Macro F_1	Rank	ICM-H	Macro F_1
EXIST2025-test_gold	0	1.3244	1.0000	0	1.2232	1.0000	0	1.3913	1.0000
Top-1 Submissions	1	0.3344	0.7065	1	0.2565	0.7089	1	0.3757	0.7008
UMUTeam_1	**64**	**-0.4232**	**0.5024**	**65**	**-0.6183**	**0.4677**	**56**	**-0.2885**	**0.5254**
baseline majority class	69	-0.7537	0.2375	67	-0.6304	0.2522	69	-0.8657	0.2246
baseline minority class	70	-2.4749	0.0586	70	-2.9537	0.0431	70	-2.1890	0.0709

emerges when analyzing the results across different languages. While our model demonstrates a clear improvement within the English subset—reaching a higher Macro F_1-score of 0.5254 and climbing to the 56th position of the official leaderboard a parallel upward trend is observed in the top-1 English submission, which reaches 0.7008. Consequently, this performance gap remains persistent between both evaluation setups. Nevertheless, this localized enhancement in our absolute results can be primarily attributed to the architectural priors of the visual extractor. The pre-training phase of the SigLIP model relies heavily on massive English-centric corpora, which inherently yields more robust and discriminative multimodal representations in this language than in Spanish.

6. Conclusions and Future Work

In this work, we developed and evaluated a multimodal framework that demonstrates a solid capability to capture underlying patterns within complex datasets, consistently outperforming random baselines across both evaluated tracks.

For both tasks, a cohesive Stacking-based late fusion strategy was adopted, utilizing different base classifiers adapted to the specificities of each problem. For Task 3.1, a bimodal fusion architecture was implemented, combining textual and acoustic modalities through their respective classifiers. For Task 3.2, the pipeline fused textual features with visual embeddings. These architectures were the result of extensive preliminary experimentation aimed at determining the optimal combination of features and classification techniques.

Regarding the official leaderboards, our systems secured the 70th position out of 73 official submissions in Task 3.1, and the 64th position out of 68 in Task 3.2 (excluding baseline performance metrics). While these absolute positions remain modest, they provide a transparent benchmark that highlights specific areas for improvement, which can be primarily attributed to two main architectural and methodological limitations.

First, our pipeline did not incorporate the physiological modalities introduced in this edition of the shared task. Comparing the top results with the overview of the last edition of EXIST [1], it is clear that including physiological data makes a huge difference for the performance. In Task 3.1, the inclusion of these signals by the top teams nearly doubled the official ICM-Hard scores and yielded an increase of approximately 0.06 in Macro F_1-score compared to last year's top metrics. In Task 3.2, this performance gap is even more pronounced, exhibiting an increase of over a tenth in the top-1 submission compared to the previous edition. Consequently, omitting this data source significantly constrained our model's competitive ceiling.

Second, our research had some limitations regarding the amount of data we used. Due to problems downloading the full video corpus, we trained our models with about 1,000 fewer samples than the official training dataset. This missing data means that our training was not as complete, and the conclusions we drew during our local tests might not be the same as if we had used all the original videos.

Despite these limitations, this research delivers a functional, structured pipeline capable of performing reasoned multimodal classification well above baseline performance thresholds.

As an improvement of this work, it is advisable to directly utilize the official downloaded videos provided by the competition after registration, ensuring the research is based on the exact same data. Since we started working on our pipeline before the official registration, we could not access this resource at that time and had to download the videos manually via URL, which was not ideal and caused the missing samples. Additionally, given the importance of physiological features, future work should integrate physiological feature extraction into our stacking framework to close the competitive gap identified in this edition. Finally, future efforts could explore the incorporation of native Spanish language models to mitigate the linguistic bias observed in our visual encoder and achieve a more balanced performance across both language subsets.

Acknowledgments

This work is part of the research project LaTe4PoliticES (PID2022-138099OB-I00) funded by MICIU/AEI/10.13039/501100011033 and the European Regional Development Fund (ERDF/EU - FEDER/UE)-a way of making Europe.

Declaration on Generative AI

During the preparation of this work, the author(s) used Gemini 3.5 Flash in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, 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, Overview of exist 2025: Learning with disagreement for sexism identification and characterization in tweets, memes, and tiktok videos, in: J. Carrillo-de Albornoz, J. Gonzalo, L. Plaza, A. García Seco de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer Nature Switzerland, Cham, 2025. doi:10.1007/978-3-031-88720-8_65.

[2] L. Plaza, J. Carrillo-de Albornoz, I. Arcos, M. Aloy-Mayo, E. Gomis-Vicent, P. Rosso, E. García-Arias, D. Spina, Overview of EXIST 2026: Physiological data for multimodal sexism characterization in social media, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), 2026.

[3] K. Sparck Jones, A statistical interpretation of term specificity and its application in retrieval, Journal of Documentation 28 (1972) 11–21.

[4] S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Transactions on Acoustics, Speech, and Signal Processing 28 (1980) 357–366. URL: http://ieeexplore.ieee.org/document/1163420/. doi:10.1109/TASSP.1980.1163420.

[5] X. Zhai, B. Mustafa, A. Kolesnikov, L. Beyer, Sigmoid Loss for Language Image Pre-Training, 2023. URL: https://arxiv.org/abs/2303.15343. doi:10.48550/ARXIV.2303.15343, version Number: 4.

[6] M. Zha, H.-C. H. Chang, Interpreting Multimodal Communication at Scale in Short-Form Video: Visual, Audio, and Textual Mental Health Discourse on TikTok, 2026. URL: http://arxiv.org/abs/2601.15278. doi:10.48550/arXiv.2601.15278, arXiv:2601.15278 [cs.MM].

[7] M. Das, R. Raj, P. Saha, B. Mathew, M. Gupta, A. Mukherjee, HateMM: A Multi-Modal Dataset for Hate Video Classification, 2023. URL: https://arxiv.org/abs/2305.03915. doi:10.48550/ARXIV.2305.03915, version Number: 1.

[8] J. Lang, R. Hong, J. Xu, Y. Li, X. Xu, F. Zhou, et al., Biting off more than you can detect: Retrieval-augmented multimodal experts for short video hate detection, in: Proceedings of the ACM on Web Conference 2025, 2025, pp. 2763–2774. doi:10.1145/3696410.3714560.

[9] B. Céspedes-Sarrias, C. Collado-Capell, P. Rodenas-Ruiz, O. Hrynenko, A. Cavallaro, MM-HSD: Multi-Modal Hate Speech Detection in Videos, 2025. URL: https://arxiv.org/abs/2508.20546. doi:10.48550/ARXIV.2508.20546, version Number: 1.

[10] I. Arcos, P. Rosso, Sexism Identification on TikTok: A Multimodal AI Approach with Text, Audio, and Video, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, G. M. Di Nunzio, L. Soulier, P. Galuščáková, A. García Seco De Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, volume 14958, Springer Nature Switzerland, Cham, 2024, pp. 61–73. URL: https://link.springer.com/10.1007/978-3-031-71736-9_2. doi:10.1007/978-3-031-71736-9_2, series Title: Lecture Notes in Computer Science.

[11] R. Pan, T. Bernal-Beltrán, J. A. García-Díaz, R. Valencia-García, speech-emotion: A multilingual and multimodal toolkit for emotion recognition from speech, SoftwareX 34 (2026) 102677. URL: https://www.sciencedirect.com/science/article/pii/S235271102600169X. doi:https://doi.org/10.1016/j.softx.2026.102677.

[12] 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.

[13] L. Plaza, J. Carrillo-de Albornoz, I. Arcos, M. Aloy-Mayo, E. Gomis-Vicent, P. Rosso, E. García-Arias, D. Spina, Overview of EXIST 2026: Physiological data for multimodal sexism characterization in social media (extended overview), in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.

[14] T. Joachims, Text categorization with Support Vector Machines: Learning with many relevant features, in: J. G. Carbonell, J. Siekmann, G. Goos, J. Hartmanis, J. Van Leeuwen, C. Nédellec, C. Rouveirol (Eds.), Machine Learning: ECML-98, volume 1398, Springer Berlin Heidelberg, Berlin, Heidelberg, 1998, pp. 137–142. URL: http://link.springer.com/10.1007/BFb0026683. doi:10.1007/BFb0026683, series Title: Lecture Notes in Computer Science.

[15] C. Cortes, V. Vapnik, Support-Vector Networks, Machine Learning 20 (1995) 273–297. URL: https://link.springer.com/10.1023/A:1022627411411. doi:10.1023/A:1022627411411.

[16] 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.

[17] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation 9 (1997) 1735–1780.

[18] E. Amigó, A. Delgado, Evaluating extreme hierarchical multi-label classification, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 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.