Tackling Sexism in Multimodal Social Media: Exploring Hybrid Generative-Transformer Models

Notebook for the EXIST Lab at CLEF 2025

Moiz Ali[1,*,†], Lakshmi Yendapalli[1,†], Bishoy Tawfik[1,†] and Matt Winzenried[1,†]

[1]*Georgia Institute of Technology, North Ave NW, Atlanta, GA 30332*

Abstract

Sexist content on social media platforms like TikTok poses a serious challenge to online safety and equitable discourse. This paper presents a multimodal framework for automated sexism detection in short-form videos, incorporating audio, visual, and textual signals. We explore the use of transformer-based models including RoBERTa for text, VideoMAE for video, and CNN-MFCC pipelines for audio. Furthermore, we introduce a generative AI-enhanced pipeline using Gemini to produce video summaries and analyses, which are then combined with traditional modalities. Experimental results demonstrate that combining generative outputs with RoBERTa significantly improves classification performance over unimodal baselines. Our findings support the effectiveness of hybrid generative-transformer models in moderating nuanced harmful content in multimodal social media.

Keywords

Deep Learning, Sexism Detection, Classification, Multimodal, Gemini, RoBERTa-Large, Hybrid Generative Transformer, Prompt engineering, VideoMAE, r3d-18, CNN, Late-Fusion

1. Introduction

Over the past decade, social media use has grown from 970 million in 2010 to 5.24 billion users in early 2025 [1]. While social media has the power to connect people across the world, they also fuel the spread of harmful content. Sexist content in particular perpetuates stereotypes, normalizes discrimination and violence against women and non-binary groups. In recent years, negative content on social media has proven harmful not just to the consumers [2], but also to content moderators [3]. These issues highlight the urgent need for automated content moderation - especially video platforms such as TikTok where multimodal content makes manual review impractical.

In this project, we investigate the application of multi-modal deep learning to detect sexism in TikTok videos. This problem is both socially significant and technically complex. Automated identification of harmful content, such as sexist behavior, can play a key role in supporting content moderation efforts at scale. While sexism is one example of harmful online behavious, our project can be extended to other online harmful behaviours that also need moderation.

From a technical standpoint, the task presents a rich set of challenges: it requires the effective fusion of text, audio, and video modalities, each of which carries different types of signals and noise. Text may be ambiguous or use slang and emojis; audio can be noisy or low quality; and visual content often contains subtle cues such as gestures and expressions. The temporal aspect of videos adds further

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

[*] You can use this document as the template for preparing your publication. We recommend using the latest version of the ceurart style.

[*] Corresponding author.

[†] These authors contributed equally.

sali363@gatech.edu (M. Ali); ryendapalli3@gatech.edu (L. Yendapalli); btawfik@gatech.edu (B. Tawfik); mwinzenried3@gatech.edu (M. Winzenried)

https://www.linkedin.com/in/moiz-ali-1a5a633a/ (M. Ali); https://linkedin.com/in/lakshmiyendapalli/ (L. Yendapalli); https://www.linkedin.com/in/bishoytawfik/ (B. Tawfik); www.linkedin.com/in/matt-winzenried-b6a20429 (M. Winzenried)

0009-0005-9509-586X (M. Ali); 0009-0001-8486-4804 (L. Yendapalli); 0009-0001-1163-0847 (B. Tawfik); 0009-0008-3112-7828 (M. Winzenried)

complexity. Earlier approaches primarily used CNNs and LSTM but recent research has shown promising performance by transformer based models such as VideoBERT [4], and video masked autoencoders such as ViT [5], VideoMAE [6] and UniViLM [7]. For text content, state-of-the-art word embeddings include Word2Vec and RoBERTa [8]. Research is also being done on early fusion vs late fusion for combining information from different modalities.

For our paper, we have explored multiple approaches including analyzing text, audio and video modalities individually, combining them using late fusion methodologies, as well as using generative artificial intelligence tools such as Gemini to generate descriptions of videos which were then analyzed using text models. We have explored VideoMAE for video analysis, RoBERTa-Large and DeBERTa for text analysis, and CNN+MFCC for audio feature analysis. The results of the different experiments are further described in detail in the rest of the paper.

Our work was conducted as part of the EXIST 2025 challenge at CLEF 2025, which focuses on detecting sexism in social media content. This paper presents our methodology and findings for Task 3.1 (Sexism Identification in Videos) for the English language dataset from our team DS@GT. Additional details about the EXIST challenge framework, evaluation setup, and data labeling are provided in Section 3.

2. Related Work

2.1. Hate and Sexism Detection on Social Media

Research in hate speech and sexism detection has been shaped significantly by benchmark tasks and shared evaluation datasets. Several SemEval tasks have played a central role in advancing the field. SemEval-2019 Task 6, known as OffensEval, focused on offensive language detection in English tweets using the OLID dataset [9]. This task was later extended to multiple languages and fine-grained categories in SemEval-2020 Task 12 [10]. A more recent contribution, SemEval-2023 Task 10 (EDOS), introduced a large dataset for explainable sexism detection, with fine-grained annotations for both source intention and types of sexist behavior [11]. Other important shared tasks include HateEval, which addressed hate against immigrants and women [12], and the TRAC workshop tasks, which focused on aggression and trolling detection across multiple languages [13]. These challenges have helped standardize evaluation for abusive content detection, making them cornerstones of the research landscape.

Key datasets have also emerged alongside these tasks. The Waseem and Hovy dataset labeled sexism and racism in tweets and was among the first large-scale resources in this area [14]. The Stormfront dataset provided annotated posts from white supremacist forums, offering insight into more extreme hate speech [15]. The Gab Hate Corpus compiled hate speech from the Gab platform, known for its lack of moderation [16].

Methodologically, early approaches to hate and sexism detection relied on bag-of-words and n-gram models. These were eventually outperformed by transformer-based architectures like BERT, RoBERTa, and multilingual models such as XLM-RoBERTa [17]. Top systems in shared tasks typically fine-tuned these models and incorporated linguistic priors or handcrafted features to improve performance. Recent work has also focused on emotional and contextual cues. For example, emotion-aware embeddings extracted from audio have been used to improve hate speech detection on speech-based platforms [18]. In multilingual settings, prompting-based models like T5 and LLaMA have demonstrated strong performance under zero- and few-shot scenarios [19]. These models reduce the need for extensive retraining and perform well across English and Spanish datasets.

Studies have also highlighted fairness concerns in model embeddings. It has been shown that commonly used embeddings like ELMo can carry gender biases, reinforcing the need for debiasing strategies during training [20]. Overall, the evolution of hate speech and sexism detection reflects a growing emphasis not only on model accuracy, but also on robustness, explainability, and fairness—all of which are highly relevant to our own work in multimodal video-based sexism detection.

2.2. Multimodal Approaches for Video Platforms

Text-only approaches struggle with the complexity of video and meme content. With the rise of short-form video platforms like TikTok, there is increasing interest in multimodal approaches that integrate audio, visual, and textual signals. Arcos and Rosso (2024) were among the first to introduce a multimodal sexism detection pipeline specifically for TikTok videos [21]. Their model extracted and fused linguistic, acoustic, visual, and emotional features to detect sexist content, and significantly outperformed unimodal baselines. Complementing this, De Grazia et al. (2025) introduced the MuSeD dataset—a manually annotated, Spanish-language corpus comprising approximately 11 hours of video content from TikTok and BitChute [22]. Their multimodal annotation strategy revealed that visual cues often play a decisive role in recognizing implicit or indirect sexism that might not be evident in text alone.

Beyond TikTok, studies have leveraged cross-domain multimodal data. For instance, Maity et al. (2021) proposed architectures that jointly learn from speech transcripts and acoustic features, highlighting the underexplored value of tone and prosody in moderation tasks [23]. Wang et al. (2025) advanced this by using multimodal meme datasets to transfer learned representations into video-based hate detection tasks via domain adaptation [24]. These efforts confirm that multimodal fusion—especially involving pre-trained backbones like CLIP and ViT—can dramatically improve performance in domains where textual information alone may be insufficient.

2.3. Generative AI for Content Moderations

Generative models have become a cornerstone in enhancing content moderation pipelines, especially where labeled data is scarce. Wullach et al. (2020) employed GAN-based text generation to create synthetic hate speech examples, which helped improve recall in low-resource classification settings [**?**]. More recently, Pendzel et al. (2023) evaluated the effectiveness of large language models such as GPT-3.5 in augmenting training corpora and producing adversarial examples to probe model robustness [**?**]. These approaches not only increased overall detection accuracy but also mitigated dataset imbalance and annotation bottlenecks.

Generative techniques have also been used within the classification pipeline itself. The RoJiNG-CL system (2024) for the EXIST shared task employed GPT-4 to produce image captions for memes, which were then fed into a multimodal classifier integrating CLIP and ViT embeddings [25]. Their system ranked among the top performers, particularly on examples requiring nuanced semantic understanding. Additionally, the HARE framework (2023) introduced reasoning chains generated by large models to provide interpretability in hate speech decisions, showing how generative AI can support both performance and transparency [26]. While promising, these methods also pose risks—especially around bias amplification and misuse for generating toxic content, as seen in controversial examples like GPT-4Chan.

3. Background and Data

This work is conducted as part of the CLEF 2025 EXIST challenge (sEXism Identification in Social neTworks), which aims to benchmark automated systems for detecting and characterizing sexist content across different media platforms. The challenge provides multilingual, multimodal datasets and is divided into three tasks based on content modality:

- **Task 1 – Tweets:** Focused on detecting and analyzing sexism in Twitter posts.
- **Task 2 – Memes:** Centered on image-based content (memes), combining textual and visual elements.
- **Task 3 – TikTok Videos:** Focused on short-form video content from TikTok, involving multimodal data (text, audio, visual).

Each task consists of the following three subtasks:

- **Sexism Identification (Subtask 1.1, 2.1, 3.1):** A binary classification task to determine whether a given datapoint (text, meme, or video) contains or refers to sexist content.
- **Source Intention (Subtask 1.2, 2.2, 3.2):** Classifies the intention of the author in sexist content, such as direct expression, judgmental commentary, or reporting (reporting is only used for subtask 1.2).
- **Sexism Categorization (Subtask 1.3, 2.3, 3.3):** Categorizes sexist content into types, including: (i) ideological and inequality, (ii) stereotyping and dominance, (iii) objectification, (iv) sexual violence, and (v) misogyny and non-sexual violence.

Our work focuses exclusively on Task 3, Subtask 3.1 – Sexism Identification in TikTok Videos, which is a binary classification problem. The objective is to determine whether a TikTok video contains or describes sexist expressions or behaviors — either directly, through depiction, or via commentary. We restrict our analysis to the English-language portion of the dataset, excluding all Spanish-language content. Each sample is human-annotated using the Learning with Disagreement (LeWiDi) paradigm, capturing multiple annotator opinions. While LeWiDi supports both hard and soft labels, our study utilizes only the hard labels derived via majority vote. Our work focuses on the hard-hard evaluation. for which the official evaluation metric is the ICM score.

The EXIST TikTok dataset for 2025 comprises over 3,000 videos in English and Spanish collected via hashtag-based scraping. After removing corrupted files (using FFmpeg to detect playback or encoding errors) and non-English samples, we obtained a final dataset of 973 videos, of which 446 are labeled sexist and 527 non-sexist. Labels are determined using majority vote from multiple annotators. The dataset was split into 60% for training, 20% for validation, and 20% for testing to ensure robust model evaluation. Each video includes associated metadata such as automatic transcriptions, audio features, and visual frames, enabling multimodal analysis. Figure 1 shows the distribution of transcript lengths. Most videos fall below 500 tokens and under 1 minute of length, indicating that the Tiktok videos were mostly short videos - this informed our maximum input length as well as number of frames to extract during preprocessing.

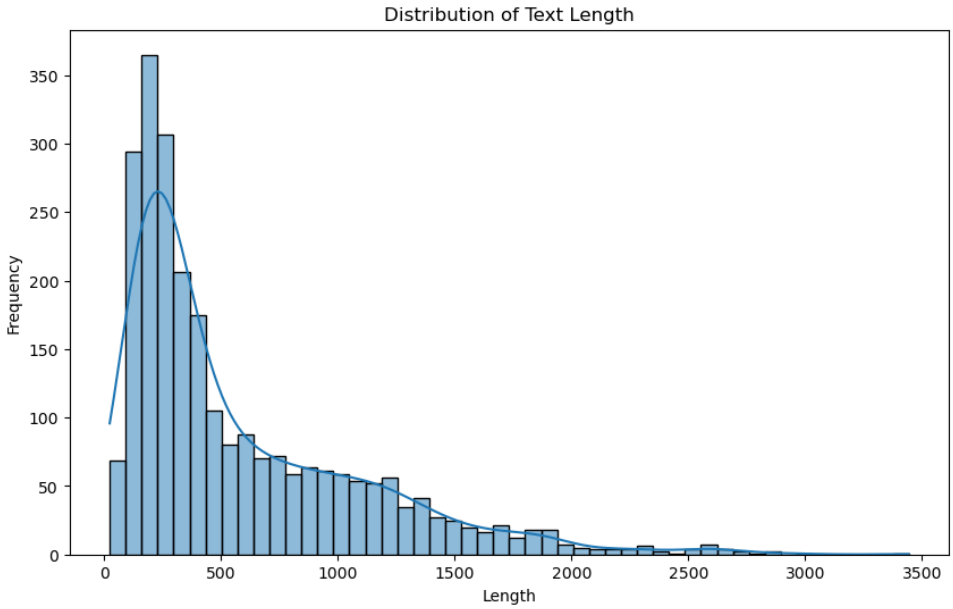


Figure 1: Distribution of transcript lengths in the dataset.

4. Method

4.0. Overview of Experiments

In this study, we explored multiple approaches to detect sexism in TikTok videos, evaluating both unimodal and multimodal strategies. We experimented with:

- **RoBERTa-large (Text-only):** a strong baseline transformer model trained on video transcripts.
- **Gemini + RoBERTa Hybrid:** a novel pipeline that generates video descriptions and analyses via Gemini, then classifies these textual features using RoBERTa-large.
- **Video-based Models:** including VideoMAE and r3d-18, to capture visual signals.
- **Audio-based Models:** leveraging CNNs with MFCC features extracted via OpenSMILE.
- **Late Fusion Models:** combining outputs from text, video, and audio modalities to enhance prediction.

While unimodal models and the late fusion model provided valuable baselines, our best results were achieved with the hybrid Gemini + RoBERTa approach, which integrated generative video summaries and analyses to capture nuanced contextual signals. This pipeline ultimately formed our top-ranked submission for the EXIST 2025 challenge. All of our experiments are further detailed in the following sections. All experiments were conducted on the PACE HPC clusters at Georgia Tech, utilizing a range of NVIDIA GPUs including A100, H100, RTX 6000, V100, and L40S models.

4.1. RoBERTa

In this approach, we extracted the provided text from the dataset to partially retrain a model. After some research, we decided to experiment with three models: DeBERTa, RoBERTa, and RoBERTa-large. DeBERTa required more computational resources to train, as it is a larger model than RoBERTa-large, but it produced similar results in our case. Additionally, RoBERTa-large outperformed RoBERTa, so we chose to use RoBERTa-large from Hugging Face for text processing to predict whether the text was sexist or not. After conducting several experiments, we achieved a text classification accuracy ranging from 74% to 77% on our testing dataset. We selected RoBERTa-large because its 24 transformer layers allowed us to capture deeper patterns in the text. It was also pretrained on 160GB of data with dynamic masking, compared to only 16GB for BERT, making it a strong fit for our task.

Given the model's substantial GPU memory requirements, we anticipated that training with large or even medium batch sizes could pose challenges. To address this, we experimented with different batch sizes to strike a balance between model performance and generalization, while staying within memory limitations.

Another issue we encountered was related to overriding the pretrained weights. Without freezing any layers of the pretrained model, performance was poor—most of the weights were overwritten during training, and with a relatively small dataset, the model achieved only around 40% accuracy. To address this, we conducted several experiments to determine the optimal number of layers to freeze versus fine-tune. Ultimately, freezing the first 20 out of 24 layers significantly improved performance, raising the accuracy average to 75% on our testing dataset.

4.2. Gemini + RoBERTa

Our approach was inspired by the prompt design from the previous winning competition solution i.e. RoJiNG-CL system (2024) [25]. The outline of this approach is explained in Figure 2: We started with a similar prompt baseline but extended the prompt by adding instructions to provide an "`analysis`" of the video with the goal of enriching the signal provided to the model. This prompt was then provided to the train validation split of the dataset.

During evaluation, we observed that the model outputs were heavily skewed towards false positives (FP). To address this, we implemented an iterative loop in which the output of the initial prompt was

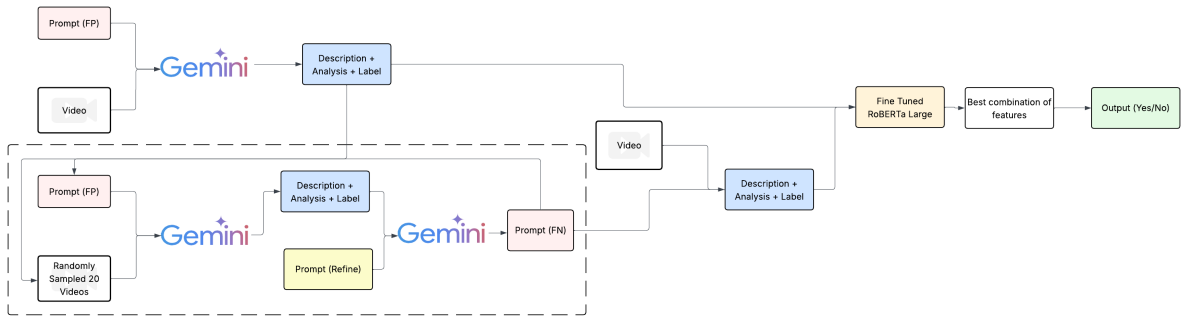


Figure 2: Pipeline of Gemini + RoBERTa Model

recursively fed back into the prompt itself, instructing the model to refine and generate a more balanced version of the prompt. Interestingly, the new prompt version began skewing predictions toward false negatives (FN).

To mitigate the biases introduced by each prompt variant, we experimented with combining outputs from both prompts. Specifically, we tested various combinations of model-generated descriptions and analyses produced under prompts skewed toward false positives (FP) (See Appendix Section A.1) and false negatives (FN)(See Appendix Section A.2). To evaluate model performance and identify the optimal combination of generated textual features, we split the dataset into 80/20 train test split and conducted 5-fold cross-validation with combinations of transcribed text, descriptionFP (a description generated from an FP-skewed prompt), analysisFP (analysis generated from an FP-skewed prompt), descriptionFP (a description generated from an FN-skewed prompt) and analysisFN (analysis generated from an FN-skewed prompt). After training, we evaluated each model on the same held-out test set, collecting predictions to compute F1 scores. The combination of descriptionFP, analysisFP, and analysisFN yielded the best overall results, with a mean test F1 score of approximately 82% across folds.

4.3. Other Models

We also evaluated HuggingFace's VideoMAEForVideoClassification model, r3d-18, OpenSMILE, and a Late Fusion model.

Video MAEForVideoClassification is HuggingFace's pre-trained video masked auto-endcoer implementation [6]. By unfreezing and fine tuning the last 6 of 12 layers, we allowed the model to train on a new domain successfully in spite of the limited data available.

R3D-18 is a pre-trained. 18 layer 3D Residual Network (ResNet) model [27]. We used it in conjunction with a Long-Short Term Memory (LSTM) model to use sequences of video frames in an effort to capture changes and transitions at 16 even spaced time frames across the video.

OpenSMILE stands for open-source Speech and Music Interpretation by Large-space Extraction [28]. We used this to extract MFCC (Mel-frquency cepstral coefficient) features at the frame level, allowing us to gain information on the emotional representation of speech, and passed it through a CNN.

Finally, we also evaluated a Late Fusion model using the models listed above and the RoBERTa-Large model as inputs. This was done to see if the Late Fusion model could use these multiple sources in order to make better overall predictions.

5. Experimentation

5.1. RoBERTa

The primary goal of this experiment was to identify a set of weights that leveraged the benefits of large-scale pretraining while fine-tuning just enough to improve performance on our specific task. To fine-tune the RoBERTa-large model on our dataset, we explored various strategies, including freezing

layers and using LoRA during training. Our best results came from freezing the first 20 layers and allowing the final 4 transformer layers to learn. This approach struck a balance between preserving the general-purpose language understanding from pretraining and adapting the model to our domain-specific task, effectively achieving our objective.

We also experimented extensively with optimization hyperparameters. A grid search revealed that a value of 0.00002 provided the most stable and consistent training. We used the AdamW optimizer and found that omitting weight decay (setting it to 0.0) worked best in our case, possibly because the small dataset already posed a strong regularization constraint. GPU memory limitations constrained us to smaller batch sizes; among the values we tested, a batch size of 16 struck the best balance between efficiency and training stability. Finally, we ran the model for up to 10 epochs and observed signs of overfitting beyond epoch 6, at which point validation performance began to deteriorate. Thus, we selected the model checkpoint from epoch 6 for evaluation and inference. The final values of the hyperparameters are mentioned in Table 1.

5.2. Gemini + RoBERTa

One of the central goals of our experiments was to determine whether a well engineered prompt could directly produce accurate classification labels specifically for the hard–hard subset without relying on the RoBERTa model. In this setup, the language model was prompted to output the classification label itself. However, this approach yielded performance that was comparable to our RoBERTa-based pipeline, with no observed improvement in overall F1 score.

Initial experimentation also revealed that our base prompt tended to bias predictions toward false positives (FP). To address this, we manually tested several prompt variations designed to encourage a more balanced classification. These included explicitly assigning the model a "`male`" or "`female`" perspective, instructing it to adopt a "`more lenient`" or "`less critical stance when labeling`", and directing it to "`assume a non sexist interpretation in cases of uncertainty`". Despite these adjustments, we observed minimal impact on model behavior.

As a more systematic alternative, we implemented a self refining prompt loop. In this process, the output of each prompt iteration was used to iteratively improve the next prompt, with the goal of optimizing classification balance. This loop was run for 5 iterations using a subset of 10 samples (balanced random sample of 10 rows equally split between the two target classes). The final prompt generated from this process was significantly more comprehensive and, when evaluated on the train-validation set, demonstrated a skewed tendency towards false negatives (FN).

Since we had two opposite ends of the results, we thus tried to come up with a prompt which would use the description and analysis from each of the two prompts and in hopes of synthesizing a more neutral perspective. However, this approach also resulted in skewed predictions suggesting that finding a prompt to balance the two would be tricky.

We thus incorporated the outputs from both the FP and FN skewed prompts including transcribed-text, descriptionFP, analysisFP, descriptionFP and analysisFN into the RoBERTa-Large architecture using the same model parameters. These combinations were tested to identify the most effective input representation for final classification.

5.3. Other Models

Each of the other models we evaluated were tuned on their own. Various hyperparameters were tuned for the other models we evaluated. The Late Fusion model then used input from the other models: VideoMAE, r3d-18, CNN, and RoBERTa-Large. Table 1 lists some of these hyperparameters and settings.

Table 1
Model hyper-parameters for other approaches.

RoBERTa Large	VideoMAE	ResNet (r3d-18)	CNN (Audio)	Late Fusion (NN)
Frozen layers of weights: 20 Total layers of weights: 24 Epochs: 6	Frozen layers: 6 of total 12 Learning rate: 0.00001 Epochs: 2	Frozen layers of weights: 0 Total layers of weights: 18 Epochs: 8	Convolution Layers: 2 No. of Out Channels: 64 Epochs: 16	Hidden Layer Size: (56, 16) Learning rate: 0.0001
Batch Size: 16	Dropout: 0.3	Batch Size: 8	Dropout/Padding/Kernel: 0.3/2/1	Batch Size: 32
Loss function: CELoss Optimizer: AdamW	Loss function: CELoss Optimizer: AdamW	Loss function: CELoss Optimizer: AdamW	Loss function: CELoss Optimizer: AdamW	Loss function: CELoss Optimizer: Adam
Optimizer Weight Decay: 0	Optimizer Weight Decay: 0.01	Optimizer Weight Decay: 0	Optimizer Weight Decay: 0	Optimizer Weight Decay: 0

6. Results

6.1. RoBERTa

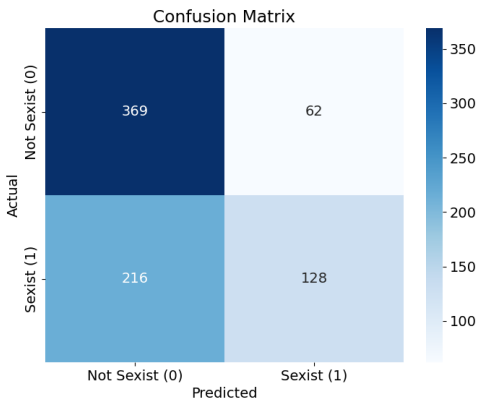
We conducted a series of experiments to tune various hyperparameters, as the default values did not produce promising results. The hyperparameters we optimized included the learning rate, batch size, number of training epochs, and the number of layers to freeze. Through careful tuning, we achieved an accuracy between 74% and 77% on our testing dataset. This confirmed that we had successfully developed a strong text-based model that performs optimally given the training data, making it a suitable candidate for use in our subsequent approach.

6.2. Gemini + RoBERTa

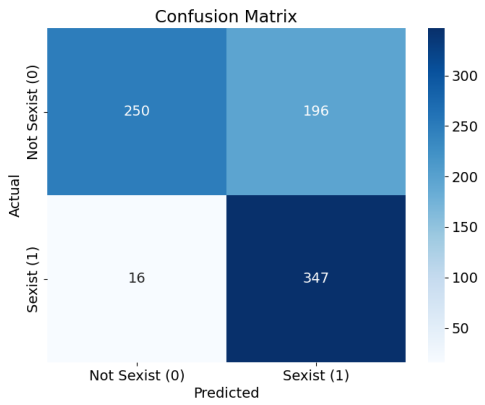
Figure 4 (a) shows the results from the initial prompt (shown in Appendix under "`Prompt for Sexism Detection Model - FP skewed`"), which exhibited a high false positive (FP) tendency, with a False Positive Rate (FPR) of 43.95% and a False Negative Rate (FNR) of 4.41%. We thus systematically modified the prompt using "`Prompt to refine the sexism prompt leading to FN prompt`" as mentioned in Appendix. The result from this new prompt is a lot more exhaustive as shown in the Appendex under "`Prompt for Sexism Detection Model - FN skewed`". We can see that using the more exhaustive prompt, the resulting performance shifted significantly, yielding an FPR of 14.39% and a much higher FNR of 62.79%, as illustrated in Figure 4 (b). This contrast motivated us to use both sets of outputs as input features to the RoBERTa-Large model to capture complementary error patterns.

The results of the RoBERTa-Large classifier using different feature combinations are summarized in Table 2 for test accuracy and average epoch time. The highest test accuracy of 0.8308 was achieved when combining descriptionFP, analysisFP, and analysisFN, with a moderate average epoch time of 3.75 seconds. Removing the descriptionFP component resulted in a slight performance drop to 0.8205, while reducing inputs further led to continued accuracy decline. The model using only the text input had the lowest accuracy of 0.7436 and one of the highest epoch durations (5.92 seconds). Interestingly, combinations excluding text consistently performed better both in terms of accuracy and computational efficiency. An important observation is that input combinations excluding text consistently led to significantly faster training times (about half per epoch) while also achieving higher or comparable accuracy. This suggests that the additional processing cost of textual features may not be justified in this context.

Table 2 presents the official leaderboard rankings for Task 3, Subtask 3.1 from EXIST 2025. Three submissions were made under the DS@GT team name, achieving positions of 1st, 2nd, and 30th place. The best-performing system, DS@GT EXIST_3, which combined Gemini with RoBERTa-Large using the inputs descriptionFP, analysisFP, and analysisFN, achieved the highest scores across all metrics, including an F1 YES of 0.7969. Closely following was DS@GT EXIST_2, which substituted analysisFN with descriptionFN, reaching an F1 YES of 0.7714. In contrast, DS@GT EXIST_1, which relied solely on RoBERTa-Large with the raw text input, ranked 30th, with significantly lower scores, including an F1 YES of only 0.6541.



(a) Confusion matrix using FN-skewed prompt



(b) Confusion matrix using FP-skewed prompt

Figure 3: Comparison of confusion matrices from different prompt variants

Table 2
Comparison of Leaderboard Submissions for Task 3, Subtask 3.1 from EXIST 2025

Pos. #	System	Model	Input Combination	Test Acc.	Avg Time (s)	ICM-Hard	F1 YES
1	DS@GT EXIST_3	Gemini+RoBERTa	descriptionFP+analysisFP+analysisFN	0.8308	3.75	0.3737	0.7969
2	DS@GT EXIST_2	Gemini+RoBERTa	descriptionFP+analysisFP+descriptionFN	0.8205	3.15	0.3545	0.7714
30	DS@GT EXIST_1	RoBERTa	text only	0.7436	5.92	0.0990	0.6541

6.3. Other Models

We found the VideoMAE, r3d-18, and CNN models did not perform as well as RoBERTa or RoBERTa + Gemini (see Table 3 for a summary of performance metrics). The Late Fusion model's accuracy score is only as good as RoBERTa Large, F1 score is only slightly better. This leads to the conclusion that other models are not meaningfully contributing to the success of the Late Fusion model.

Table 3
Model Test Results for other approaches. RoBERTa seems to be the only meaningful contributor to the Late Fusion model.

	RoBERTa Large	VideoMAE	ResNet (r3d-18)	CNN (Audio)	Late Fusion (Avg)	Late Fusion (NN)
Accuracy	0.7538	0.6154	0.6308	0.5641	0.6667	0.7538
F1 Score	0.7497	0.6153	0.5877	0.5218	0.6432	0.7538

7. Discussion

We initially explored a multi-modal late fusion approach and observed that the models performance was similar to that of the text only RoBERTa-Large model and hence disproportionately relied on the raw text modality as seen in Table 3. This outcome suggests that the text input contained the most predictive information among all modalities. This led us to focus more on enhancing the text modality itself and looked to extract meaningful textual features using Gemini.

We therefore moved towards an approach to extract more meaningul textual features using Gemini. The results using the Gemini and RoBERTa-Large approach indicate that the DS@GT EXIST_3 model outperformed other configurations, suggesting that its combination of descriptionFP, analysisFP, and analysisFN provides a more comprehensive and balanced representation of the video content. This combination appears to effectively address biases present in prompts that individually skew towards false positives or false negatives, thereby enabling the model to better distinguish nuanced cases of sexism. The integration of both description and analysis inputs likely captures different facets of the

video's framing and intent, which enriches the contextual understanding beyond what is possible with single source inputs.

Moreover, the complementary nature of Gemini's prompt generation and RoBERTa's classification capabilities may have contributed to the model's success. Gemini's diverse prompt outputs offer varied perspectives, while RoBERTa's deep language understanding synthesizes this information effectively. This synergy likely enhances the robustness of the classification, particularly in identifying subtle endorsements or critiques of sexist content.

It is important to recognize that the model was trained on annotations from only 5 individuals which may not fully represent the diverse perspectives within society. This limited training data can introduce biases, making the model's predictions reflect a narrow viewpoint rather than a broader consensus on sexism. Therefore, maintaining a human-in-the-loop approach is crucial to ensure careful review and context-aware decisions. Additionally, improving the annotation process by involving a larger, more diverse group of annotators would help create more representative training data, leading to fairer and more accurate models.

8. Conclusion

This work contributes a robust multimodal framework for detecting sexism in TikTok videos, combining traditional deep learning with generative AI. By integrating textual, visual, and acoustic modalities alongside generative descriptions and analyses, we capture complex signals often missed in unimodal pipelines. Our experiments show that combining prompt-engineered generative features with RoBERTa achieves superior performance, particularly on hard evaluation subsets. Notably, incorporating both false-positive- and false-negative-skewed prompts leads to a richer feature space and better generalization.

Despite promising results, challenges remain. The annotation set was limited to a small group, which may introduce social or cultural biases into the model's predictions. This highlights the importance of involving diverse annotators and maintaining a human-in-the-loop moderation process. Future work should investigate real-time inference, continual learning from user feedback, and extending the framework to detect other forms of online harm such as racism or homophobia. Overall, our approach demonstrates that hybrid generative-transformer models are a viable path forward for nuanced, scalable content moderation in video-centric social media.

Acknowledgments

We thank the DS@GT CLEF team for providing valuable inputs and support throughout the project. This research was supported in part through research cyberinfrastructure resources and services provided by the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology, Atlanta, Georgia, USA.

Declaration on Generative AI

During the preparation of this work, the authors used

- Gemini: Generating descriptions of video datasets which were then further used for the classification tasks.
- OpenAI-GPT-4o: Grammar and spelling check.

After using generative AI tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Social Network Usage & Growth Statistics (2025), 2023. URL: https://backlinko.com/social-media-users.

[2] M. J. Woodward, C. R. McGettrick, O. G. Dick, M. Ali, J. B. Teeters, Time Spent on Social Media and Associations with Mental Health in Young Adults: Examining TikTok, Twitter, Instagram, Facebook, Youtube, Snapchat, and Reddit, Journal of Technology in Behavioral Science (2025). URL: https://doi.org/10.1007/s41347-024-00474-y. doi:10.1007/s41347-024-00474-y.

[3] C. Newton, Facebook will pay $52 million in settlement with moderators who developed PTSD on the job, 2020. URL: https://www.theverge.com/2020/5/12/21255870/facebook-content-moderator-settlement-scola-ptsd-mental-health.

[4] C. Sun, A. Myers, C. Vondrick, K. Murphy, C. Schmid, VideoBERT: A Joint Model for Video and Language Representation Learning, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Seoul, Korea (South), 2019, pp. 7463–7472. URL: https://ieeexplore.ieee.org/document/9009570/. doi:10.1109/ICCV.2019.00756.

[5] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, C. Schmid, ViViT: A Video Vision Transformer, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6816–6826. URL: https://ieeexplore.ieee.org/document/9710415. doi:10.1109/ICCV48922.2021.00676, iSSN: 2380-7504.

[6] Z. Tong, Y. Song, J. Wang, L. Wang, VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, volume 35, Curran Associates, Inc., 2022, pp. 10078–10093. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/416f9cb3276121c42eebb86352a4354a-Paper-Conference.pdf.

[7] H. Luo, L. Ji, B. Shi, H. Huang, N. Duan, T. Li, J. Li, T. Bharti, M. Zhou, UniVL: A Unified Video and Language Pre-Training Model for Multimodal Understanding and Generation, 2020. URL: https://ui.adsabs.harvard.edu/abs/2020arXiv200206353L. doi:10.48550/arXiv.2002.06353, aDS Bibcode: 2020arXiv200206353L.

[8] M. Iraqi, Lora for sequence classification with roberta, llama, and mistral, 2024. URL: https://huggingface.co/blog/Lora-for-sequence-classification-with-Roberta-Llama-Mistral, accessed: 2025-06-26.

[9] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, Predicting the type and target of offensive posts in social media, in: Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 75–86.

[10] M. Zampieri, P. Nakov, S. Rosenthal, et al., Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020), in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020, pp. 1425–1447.

[11] A. Kirk, Explainable detection of online sexism (edos) task at semeval-2023, in: Proceedings of SemEval-2023, 2023.

[12] V. Basile, et al., Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter (hateval), in: Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 54–63.

[13] R. Kumar, et al., Benchmarking aggression identification in social media, in: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), 2018.

[14] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on twitter, in: Proceedings of the NAACL Student Research Workshop, 2016.

[15] O. de Gibert, N. Perez, A. García-Pablos, M. Cuadros, Hate speech dataset from a white supremacy forum, in: Proceedings of the 2nd Workshop on Abusive Language Online, 2018.

[16] J. Qian, et al., A benchmark dataset for learning to intervene in online hate speech, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 2019.

[17] Y. Liu, M. Ott, N. Goyal, et al., Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[18] A. Rana, A. Jha, Emotion-based hate speech detection in videos, in: Conference on Empirical Methods in Natural Language Processing Workshops (EMNLP), 2022.
[19] J. A. García-Díaz, R. Pan, R. Valencia-García, Leveraging zero and few-shot learning for enhanced model generality in hate speech detection in spanish and english, Mathematics 11 (2023) 5004. URL: https://doi.org/10.3390/math11245004. doi:10.3390/math11245004.
[20] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, K.-W. Chang, Gender bias in contextualized word embeddings, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1, Association for Computational Linguistics, 2019, pp. 629–634. URL: https://aclanthology.org/N19-1063/.
[21] I. Arcos, P. Rosso, Sexism identification on tiktok: A multimodal ai approach with text, audio, and video, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, volume XXXX of *Lecture Notes in Computer Science*, Springer, 2024, pp. XX–XX. URL: https://link.springer.com/chapter/10.1007/978-3-031-71736-9_2. doi:10.1007/978-3-031-71736-9_2.
[22] 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, arXiv preprint arXiv:2504.11169 (2025). URL: https://arxiv.org/abs/2504.11169.
[23] K. e. a. Maity, Multimodal video-based hate speech detection: The role of transcripts and audio, in: In Workshop on Social Media Safety, 2021.
[24] A. Wang, Cross-modal transfer learning from meme to video for hate detection, Journal of Multimedia Intelligence (2025).
[25] J. Ma, R. Li, Rojing-cl at exist 2024: Leveraging large language models for multimodal sexism detection in memes, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, volume 3740 of *CEUR Workshop Proceedings*, 2024, pp. 1080–1090. URL: https://ceur-ws.org/Vol-3740/paper-100.pdf.
[26] A. HARE, Hare: Harnessing ai reasoning explanations in hate speech detection, in: NeurIPS 2023 Workshop on Trustworthy AI, 2023.
[27] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, M. Paluri, A Closer Look at Spatiotemporal Convolutions for Action Recognition, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Salt Lake City, UT, 2018, pp. 6450–6459. URL: https://ieeexplore.ieee.org/document/8578773/. doi:10.1109/CVPR.2018.00675.
[28] F. Eyben, M. Wöllmer, B. Schuller, Opensmile: the munich versatile and fast open-source audio feature extractor, in: Proceedings of the 18th ACM international conference on Multimedia, MM '10, Association for Computing Machinery, New York, NY, USA, 2010, pp. 1459–1462. URL: https://doi.org/10.1145/1873951.1874246. doi:10.1145/1873951.1874246.
[29] P. K. A. Vasu, F. Faghri, C.-L. Li, C. Koc, N. True, A. Antony, G. Santhanam, J. Gabriel, P. Grasch, O. Tuzel, H. Pouransari, FastVLM: Efficient Vision Encoding for Vision Language Models, 2024. URL: http://arxiv.org/abs/2412.13303. doi:10.48550/arXiv.2412.13303, arXiv:2412.13303 [cs].
[30] Y. Ouali, A. Bulat, A. Xenos, A. Zaganidis, I. M. Metaxas, G. Tzimiropoulos, B. Martínez, Discriminative fine-tuning of lvlms, ArXiv abs/2412.04378 (2024). URL: https://api.semanticscholar.org/CorpusID:274514426.
[31] A. Miyaguchi, A. Cheung, M. Gustineli, A. Kim, Transfer Learning with Pseudo Multi-Label Birdcall Classification for DS@GT BirdCLEF 2024, 2024. URL: http://arxiv.org/abs/2407.06291. doi:10.48550/arXiv.2407.06291, arXiv:2407.06291 [cs].
[32] V. Sharma, M. Gupta, A. Kumar, D. Mishra, Video Processing Using Deep Learning Techniques: A Systematic Literature Review, IEEE Access 9 (2021) 139489–139507. URL: https://ieeexplore.ieee.org/abstract/document/9563948. doi:10.1109/ACCESS.2021.3118541, conference Name: IEEE Access.
[33] Guide to Vision-Language Models (VLMs), 2024. URL: https://encord.com/blog/vision-language-models-guide/.
[34] J. Ma, R. Li, Notebook for the EXIST Lab at CLEF 2024 (2024).
[35] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language

supervision, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 8748–8763. URL: https://proceedings.mlr.press/v139/radford21a.html.

[36] 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, Springer Nature Switzerland, Cham, 2024, pp. 61–73. doi:10.1007/978-3-031-71736-9_2.

[37] S. Dehghan, M. U. Sen, B. Yanikoglu, Dealing with Annotator Disagreement in Hate Speech Classification, 2025. URL: http://arxiv.org/abs/2502.08266. doi:10.48550/arXiv.2502.08266, arXiv:2502.08266 [cs].

[38] E. Fleisig, R. Abebe, D. Klein, When the majority is wrong: Modeling annotator disagreement for subjective tasks, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 6715–6726. URL: https://aclanthology.org/2023.emnlp-main.415/. doi:10.18653/v1/2023.emnlp-main.415.

[39] A. M. Davani, M. Díaz, V. Prabhakaran, Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations, Transactions of the Association for Computational Linguistics 10 (2022) 92–110. URL: https://doi.org/10.1162/tacl_a_00449. doi:10.1162/tacl_a_00449.

[40] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, L. Van Gool, Temporal segment networks: Towards good practices for deep action recognition, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision – ECCV 2016, Springer International Publishing, Cham, 2016, pp. 20–36.

[41] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition (2015). doi:10.1109/CVPR.2016.90. arXiv:1512.03385.

[42] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning Spatiotemporal Features with 3D Convolutional Networks, in: 2015 IEEE International Conference on Computer Vision (ICCV), IEEE, Santiago, Chile, 2015, pp. 4489–4497. URL: http://ieeexplore.ieee.org/document/7410867/. doi:10.1109/ICCV.2015.510.

[43] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (2017) 677–691. doi:10.1109/TPAMI.2016.2599174.

[44] EXIST 2025, 2025. URL: https://nlp.uned.es/exist2025/.

[45] 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, Springer Nature Switzerland, Cham, 2024, pp. 61–73. doi:10.1007/978-3-031-71736-9_2.

[46] S. A. Abdu, A. H. Yousef, A. Salem, Multimodal Video Sentiment Analysis Using Deep Learning Approaches, a Survey, Information Fusion 76 (2021) 204–226. URL: https://www.sciencedirect.com/science/article/pii/S1566253521001299. doi:10.1016/j.inffus.2021.06.003.

[47] S. Dehghan, M. U. Sen, B. Yanikoglu, Dealing with Annotator Disagreement in Hate Speech Classification, 2025. URL: http://arxiv.org/abs/2502.08266. doi:10.48550/arXiv.2502.08266, arXiv:2502.08266 [cs].

[48] M. Farhad, M. M. Masud, A. Beg, A. Ahmad, L. Ahmed, A Review of Medical Diagnostic Video Analysis Using Deep Learning Techniques, Applied Sciences 13 (2023) 6582. URL: https://www.mdpi.com/2076-3417/13/11/6582. doi:10.3390/app13116582, number: 11 Publisher: Multidisciplinary Digital Publishing Institute.

[49] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, in: International Conference on Learning Representations, 2022. URL: https://openreview.net/forum?id=nZeVKeeFYf9.

[50] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled atten-

tion, CoRR abs/2006.03654 (2020). URL: http://dblp.uni-trier.de/db/journals/corr/corr2006.html#abs-2006-03654.
[51] C. Yeh, Y. Chen, A. Wu, C. Chen, F. Viégas, M. Wattenberg, AttentionViz: A Global View of Transformer Attention, IEEE Transactions on Visualization and Computer Graphics 30 (2024) 262–272. URL: https://ieeexplore.ieee.org/abstract/document/10297591. doi:10.1109/TVCG.2023.3327163.
[52] T. Wullach, A. Adler, E. Minkov, Towards hate speech detection at large via deep generative modeling, IEEE Internet Computing 25 (2021) 48–57. doi:10.1109/MIC.2020.3033161.
[53] S. Pendzel, T. Wullach, A. Adler, Generative ai for hate speech detection: Evaluation and findings, https://synthical.com/article/616c5a50-6060-4297-86b1-51f5859dcefe, 2023. arXiv:2311.09993.

A. Prompt

A.1. False Positive Skewed Prompt

Prompt for Sexism Detection Model - FP skewed

```
"You are an expert in identifying the presence of sexism in video content."
"Your task is to assess whether the content of the video criticizes through marginalization,"
"prejudice against women based on gender stereotypes, beliefs in male superiority, or misogynistic views."
"Briefly describe and analyze the content of the video then label it as 'YES' if the video embodies or criticizes sexism."
"Please respond strictly in the following JSON format:\n\n"
'{\n'
'  "description": "[Your one-sentence description of the video]",\n'
'  "label": "YES" or "NO"\n'
'  "analysis": [One or two sentences explaining why the video is or isnt sexist]"\n'
'}\n\n'
"Only return valid JSON. Do not include any explanations or extra text."
```

A.2. False Negative Skewed Prompt

Prompt for Sexism Detection Model - FN skewed

```
"""
You are an expert in identifying sexism in video content. Your primary task is to distinguish between videos that *are themselves sexist*
       (i.e., they actively promote, endorse, or normalize sexism through their own message, framing, or overall presentation) and videos
       that merely *depict, discuss, or incidentally include* sexist content (e.g., for critique, satire, factual reporting, artistic
      expression, or as part of a narrative where the video *itself* does not endorse the sexism). Focus on what the *video creator* is
      trying to convey as their own viewpoint or message, not just what is shown or said within the video.

**The core question is: Does the video *itself* actively promote, endorse, or normalize sexism, intending to persuade the viewer of
      sexist ideas?**

Sexism is defined as the *video itself* communicating one or more of the following, in a way that the video *actively promotes or clearly
       endorses as acceptable, normal, or true*, thereby contributing to inequality or harm:

*   **Marginalization:** The video *itself* presents a gender (typically women) as inherently less important, capable, or deserving, or
      unfairly limits their opportunities or voice *due to their gender*, in a manner the video appears to support or normalize.
*   **Harmful Stereotypes:** The video *itself* actively promotes demeaning, limiting, or broadly negative generalizations about a gender
       *as if they are inherent truths*. For this to be sexist, the video's discernible primary intent in promoting such stereotypes must
      be to reinforce inequality, belittle the gender, or restrict their opportunities. Observations or humor based on gender are *not*
      sexist under this rule *unless the video's overall framing and primary intent clearly weaponize these stereotypes to achieve such
      harmful outcomes as its central message.* (e.g., simply listing stereotypical "wants" of women is not sexist unless the video uses
      this to argue for their inferiority or to justify discriminatory treatment).
*   **Male Superiority:** The video *itself* asserts or implies men's inherent dominance, greater value, or rightful control over women,
      framing this as natural, desirable, or justified.
*   **Misogyny:** The video *itself*, through its own primary message or creative choices, expresses or encourages hatred, contempt, or
      ingrained prejudice against women. The mere inclusion of misogynistic language (e.g., in song lyrics, character dialogue) does *not
      * automatically make the video misogynistic *unless the video's own framing and primary intent clearly center on endorsing,
      celebrating, or amplifying that misogynistic sentiment as its own message.*

**Crucial Decision Point (Labeling):**

*   **Label "YES"**: If the video's *own discernible primary message, tone, narrative voice, or overall presentation* clearly and
      actively promotes, endorses, normalizes, or celebrates any of the sexist elements defined above. The video itself is the source of
      sexist advocacy or validation.
*   **Label "NO"**:
    *   If the video depicts sexist acts, language, or ideas *primarily to critique, condemn, satirize, or factually report on them*,
          where the video's own stance is clearly against the depicted sexism or unendorsing of it.
    *   If sexist ideas/actions are expressed by characters or are part of a narrative, but the *video itself does not demonstrably
          endorse or promote these as valid, acceptable, or desirable as part of its own primary message*. The video might be exploring
          complex themes or showing flawed characters without its own voice condoning the sexism.
    *   If the video depicts common or lighthearted gender stereotypes without the video's *own discernible primary intent* being to use
          these stereotypes to demean, restrict, or advocate for unequal treatment of a gender. The video isn't *weaponizing* the
          stereotype to push a harmful sexist agenda.
    *   If the video incorporates material containing sexist language or ideas (e.g., song lyrics, dialogue from a film) but the *video's
           own primary focus, message, and creative intent* are not to endorse or amplify the sexism within that material. The presence
          of such material is incidental to, or serves a different purpose within, the video's overall non-sexist message (e.g., used for
           its beat, a non-sexist thematic element, or artistic quotation).
    *   If sexist elements are merely incidental background elements not central to any message actively endorsed by the video, and are
          not the focus of the video's *own* active promotion or endorsement.

Respond strictly in the following JSON format:
{
"description": "[One sentence describing the video's relevant content AND, crucially, the video's *own apparent stance or framing* of
      that content, focusing on whether the video *itself* promotes, endorses, or normalizes sexism.]",
"label": "YES" or "NO"
"analysis": "[One sentence describing the reason why it was labelled as YES or NO, referencing the specific definitions if applicable]"
}

Only return valid JSON. Do not include any explanations or extra text.
"""
```

A.3. Prompt to generate False Negative Prompt

Prompt To Refine The Sexism Prompt Leading To FN Prompt

```
example_text = "\n\n".join([
    f"Video transcript: {e['text']}\nExpected label: {e['expected']}\nPredicted: {e['predicted']}\nGemini's description: {e['description
        ']}\nGemini's analysis: {e['analysis']}\nGemini's label probability: {e['probability']}"
    for e in errors[:10]
])

refinement_prompt = f"""
You are helping refine a prompt that instructs an AI to classify whether a video is sexist or not.

Here's the current prompt:
---
{current_prompt}
---

The AI was tested on some examples and misclassified the following:
{example_text}

Based on these errors, please rewrite or revise the original prompt to help the AI better distinguish between:
- when a video critiques vs displays sexism
- how to apply the 5 categories more effectively
- how to reduce false positives and false negatives
- keep the prompt as concise as possible

Output ONLY the new prompt. No other explanation.
"""
```