

# Enhancing Sexism Detection in Tweets with Annotator-Integrated Ensemble Methods and Multimodal Embeddings for Memes

Notebook for the EXIST Lab at CLEF 2024

Martha Paola Jimenez-Martinez<sup>1\*</sup>, Joan Manuel Raygoza-Romero<sup>1</sup>,  
Carlos Eduardo Sánchez-Torres<sup>3</sup>, Irvin Hussein Lopez-Nava<sup>1,3</sup> and Manuel Montes-y-Gómez<sup>2</sup>

<sup>1</sup>Centro de Investigación Científica y de Educación Superior de Ensenada, Mexico

<sup>2</sup>Instituto Nacional de Astrofísica, Óptica y Electrónica, Mexico

<sup>3</sup>Universidad Autónoma de Baja California, Ensenada, Baja California, Mexico

## Abstract

This paper details MMICI's participation in the EXIST challenge at CLEF 2024, focusing on the identification and categorization of sexism in social media and memes. For tweets, we employed pre-trained transformer models and ensemble voting approaches. For memes, we utilized CLIP embeddings using a Vision Transformer (ViT) model and two types of classifiers: feed-forward neural networks and factorization machines. The tasks encompassed detecting sexism in tweets and memes, as well as categorizing their type and the author's intention. Our methodology for tweets integrates annotator profiles, such as gender and age, to enhance the accuracy of sexism identification, source intention, and sexism categorization. For memes, we utilized all annotator features (gender, age, ethnicity, study level, and country) for the same tasks. The results demonstrate the effectiveness of our models across various tasks, emphasizing the integration of diverse perspectives. Notably, our best performances include a 10th place ranking in Task 1, a 15th place ranking in Task 2, and a 13th place ranking in Task 3 for Spanish tweets. For memes, we achieved a 3rd place ranking in Task 4 for English memes, two 1st place rankings in Task 5 for both English and Spanish memes, and a 2nd place ranking in Task 6 for English memes. These results underscore the importance of incorporating the demographic factors of annotators and taking advantage of multimodal embeddings for robust performance in sexism detection.

## Keywords

Sexism detection, Sexism identification, Sexism classification, Social media, Transformer models

## 1. Introduction

According to the Cambridge Dictionary, sexism is defined as "(actions based on) the belief that the members of one sex are less intelligent, able, skilful, etc. than the members of the other sex, especially that women are less able than men [1]". In contrast, the Royal Spanish Academy defines sexism as "discrimination against individuals based on their sex" (in spanish: *discriminación de las personas por razón de sexo*) [2]. Both interpretations, based on the meaning and expression in both languages, agree that sexism not only reflects but also communicates and perpetuates the stereotypes and roles historically assigned to women and men in society. This perpetuation of stereotypes is a significant factor in the struggle for gender equity [3].

Research on gender ideologies employs the Ambivalent Sexism Inventory and the Ambivalence toward Men Inventory. The Ambivalent Sexism Inventory measures hostile sexism, which reflects antagonistic attitudes towards women, and benevolent sexism, which consists of subjectively favorable but patronizing beliefs about women. The Ambivalence toward Men Inventory assesses hostility

---

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

\*Corresponding author.

✉ jimenezmp@cicese.edu.mx (M. P. Jimenez-Martinez); jraygoza@cicese.edu.mx (J. M. Raygoza-Romero); a361075@uabc.edu.mx (C. E. Sánchez-Torres); hussein@cicese.edu.mx (I. H. Lopez-Nava); mmontesg@inaoep.mx (M. Montes-y-Gómez)

ORCID 0009-0005-8701-9875 (M. P. Jimenez-Martinez); 0000-0003-3085-5678 (J. M. Raygoza-Romero); 0000-0001-5799-4067 (C. E. Sánchez-Torres); 0000-0003-3979-9465 (I. H. Lopez-Nava); 0000-0002-7601-501X (M. Montes-y-Gómez)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

toward men, rooted in the resentment of men’s perceived greater power, and benevolence toward men, which involves favorable views of men as protectors and providers. Ambivalent sexism theory posits that hostile sexism and benevolent sexism arise due to social and biological factors common across cultures, such as patriarchy, gender differentiation, and heterosexuality. Systemically, hostile sexism and benevolent sexism function as complementary ideologies that justify and perpetuate gender inequality, showing a strong correlation across cultures. This underscores the necessity of addressing both hostile and benevolent forms of sexism in the pursuit of gender equality [4].

This paper details MMICI’s participation in the "sEXism Identification in Social neTworks" (EXIST) shared task at CLEF 2024. EXIST aims to broadly capture instances of sexism, ranging from overt misogyny to subtler expressions of implicit sexist behavior, a task it has been undertaking since 2021. The goal of utilizing automatic tools is not only to detect and alert against sexist behaviors and discourses but also to estimate the prevalence of sexist and abusive situations on social media platforms, identify the most common forms of sexism, and understand how sexism manifests in these media [3].

Over the years, EXIST has evolved significantly. In 2021 and 2022, it provided a dataset with definitive (hard) labels for each tweet. However, starting from 2023 and continuing into 2024, the task expanded to generate six different labels per tweet, each derived from six distinct annotator profiles. These profiles include three women and three men from distinct age groups: 18-22, 23-45, and 46+. Furthermore, the most recent edition incorporates the demographic parameters of the annotators, such as gender, age, level of education, ethnicity, and country of residence.

## 2. Dataset EXIST 2024

In its fourth edition [5], the task has incorporated new challenges involving images, specifically memes. The six tasks are as follows:

- **Task 1: Sexism Identification in Tweets** involves identifying whether a tweet is sexist or not.
- **Task 2: Source Intention in Tweets** follows, where once a tweet is classified as sexist, it involves categorizing the intention of the author—whether the intention is direct, reported, or judgmental.
- **Task 3: Sexism Categorization in Tweets** focuses on classifying sexist tweets into specific categories such as ideological and inequality, stereotyping and dominance, objectification, sexual violence, misogyny, and non-sexual violence.
- **Task 4: Sexism Identification in Memes** is similar to Task 1 but applied to memes, determining whether a meme is sexist.
- **Task 5: Source Intention in Memes** mirrors Task 2 but for memes, categorizing them based on the author’s intention, either direct or judgmental.
- **Task 6: Sexism Categorization in Memes** parallels Task 3, classifying sexist memes into the same categories as tweets.








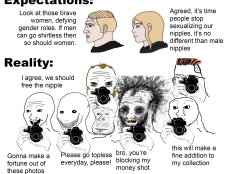
These tasks aim to enhance the understanding and detection of sexism across various forms of social media content in both English and Spanish, ultimately supporting efforts to combat sexism online. Given that information is provided from expressions in different languages, it cannot be assumed that models for detecting sexism in one language can be applied directly to another. This is due to the syntax and semantic differences in the manifestations of sexism across various countries and contexts [6]. To better understand the differences between the expressions in both languages, Table 1 provides some examples of the labels given for the different Dataset tasks where all annotators reached a consensus on that label.

Table 1: Examples of Tweets and Memes from the dataset EXIST 2024

Task	Label	Example 1 (Spanish)	Example 2 (English)
<b>TASK 1: Sexism Identification in Tweets</b>	Sexist	Mujer al volante, tenga cuidado!	People really try to convince women with little to no ass that they should go out and buy a body. Like bih, I don’t need a fat ass to get a man. Never have.






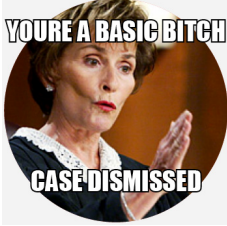
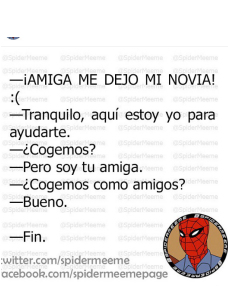



*Continued on next page*

Table 1 – Continued from previous page

Task	Label	Example 1 (ES)	Example 2 (EN)
	Not Sexist	Alguien me explica que zorra hace la gente en el cajero que se demora tanto.	@messyworldorder it's honestly so embarrassing to watch and they'll be like "not all white women are like that"
<b>TASK 2: Source Intention in Tweets</b>	Direct	Una mujer necesita amor, llenar la nevera, si un hombre puede darle esto a cambio de sus servicios (tareas domésticas, cocinar, etc.), no veo qué más necesita.	Women shouldn't code...perhaps be influencer/creator instead...it's their natural strength.
	Reported	Me duermo en el metro, abro los ojos sintiendo algo raro: la mano del hombre sentado a mi lado en mi pierna #SquealOnYourPig.	Today, one of my year 1 class pupils could not believe he'd lost a race against a girl.
	Judgemental	Como de costumbre, la mujer fue la que dejó su trabajo por el bienestar de la familia..."	21st century and we are still earning 25% less than men #Idonotrenounce.
<b>TASK 3: Sexism Categorization in Tweets</b>	Ideological and Inequality	Mi hermana y mi madre se burlan de mí por defender todo el tiempo los derechos de todos y me acaban de decir feminazi, la completaron.	I think the whole equality thing is getting out of hand. We are different, thats how were made!
	Stereotyping and Dominance	@Paula2R @faber_acuria A las mujeres hay que amarlas...solo eso... Nunca las entenderás.	Most women no longer have the desire or the knowledge to develop a high quality character, even if they wanted to.
	Objectification	"Pareces una puta con ese pantalón" - Mi hermano de 13 cuando me vio con un pantalón de cuero.	Don't get married than blame all woman for your poor investment. You should of got a hooker but instead you choose to go get a wedding ring.
	Sexual Violence	#MeToo Estas 4 no han conseguido su objetivo. El juez estima que se abrieron de patas	Fuck that cunt, I would with my fist.
	Misogyny and Non-Sexual Violence	Las mujeres de hoy en dia te enseñar a querer... estar soltero	Some woman are so toxic they don't even know they are draining everyone around them in poison. If you lack self awareness you won't even notice how toxic you really are.
<b>TASK 4: Sexism Identification in Memes</b>	Sexist		
	Not Sexist		
<b>TASK 5: Source Intention in Memes</b>	Direct		
	Judgemental		

Continued on next page

Table 1 – Continued from previous page

Task	Label	Example 1 (ES)	Example 2 (EN)
TASK 6: Sexism Categorization in Memes	Ideological and Inequality		
	Stereotyping and Dominance		Women logic... 
	Objectification		
	Sexual Violence		
	Misogyny and Non-Sexual Violence		

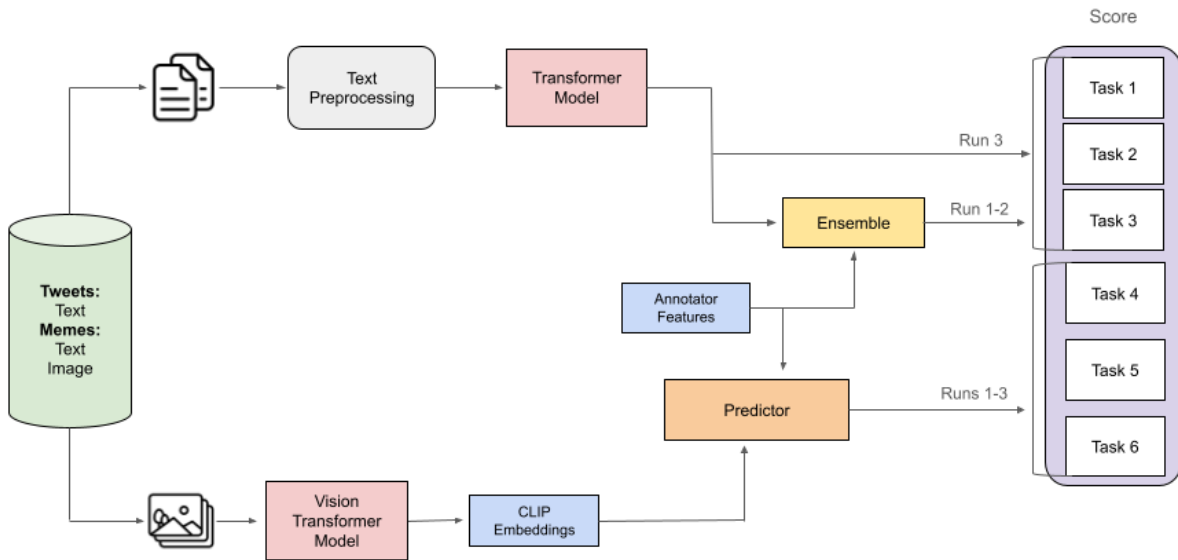
### 3. Overview of the proposal

Previous research has developed methods to model annotators in subjective tasks, allowing for the prediction of personalized labels for each annotator. For instance, Akhtar et al. [7] conducted an exhaustive search to classify annotators into two groups based on their annotation patterns. Their study demonstrated that an ensemble model, composed of two distinct classifiers representing the perspectives of each group, outperformed the traditional single-task model that only considers aggregated labels.

Additionally, traditional classification methods typically aggregate labels through majority voting or averaging before training. However, this approach has been found to potentially “silence the voices” of socio-demographic minority groups [7]. One of the objectives of this study is to leverage the individual opinions of annotators, or group them based on specific demographic characteristics, to ensure that their “voices” are effectively integrated into the sexism detection models.

Building on previous concepts, our approach to addressing the EXIST task encompasses multiple strategies across various runs and tasks:



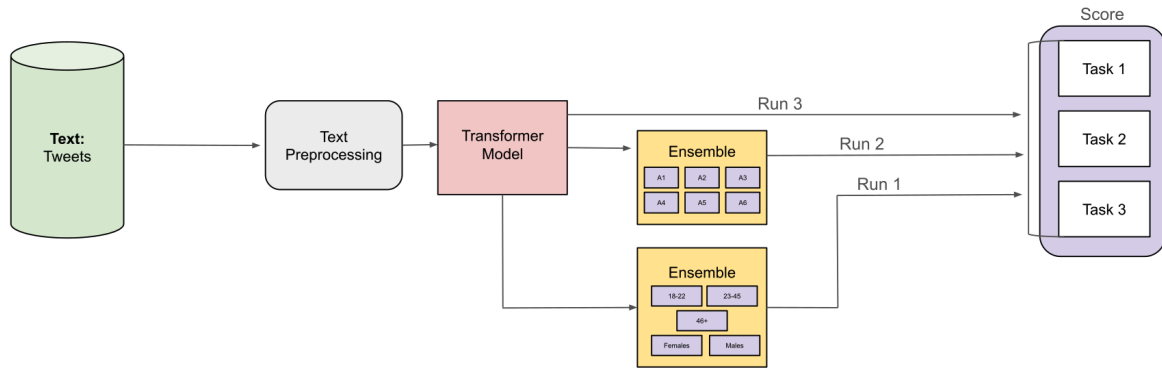


**Figure 1:** Overview of the proposal for Sexism Detection in EXIST 2024.

- **Run 1 for Tasks 1, 2, and 3:** The model predicts labels by employing an ensemble method that combines outputs based on different age groups and gender.
- **Run 2 for Tasks 1, 2, and 3:** The model predicts labels by employing an ensemble method that integrates outputs from various profiles of the six annotators.
- **Run 3 for Tasks 1, 2, and 3:** The model predicts labels using a majority vote approach, where the final prediction is based on the consensus among all annotators.
- **Runs 1 and 2 for Tasks 4 and 5:** For these tasks, our approach involves using embeddings for both the text and the image of each meme. These embeddings represent deep features of the meme. Additionally, annotator attributes are incorporated to develop a model capable of predicting labels for each annotator. The final label is determined by a voting mechanism among the predictions of the annotators.
- **Runs 1 and 2 for Task 6:** A specialized model is trained for each label using only sexism data, with the data balanced for each class. Embeddings for the meme text and image are utilized. The final output combines the model's prediction for non-sexist cases (from Task 4) with the outputs of the specialized models for each sexism category to produce a single prediction.
- **Run 3 for Tasks 4, 5, and 6:** The system predicts labels by concatenating the profile annotator's embedding with an image embedding in the same space. A multimodal embedding model assesses the relationship between annotators and items, and a voting mechanism is then applied to determine the final score.

Our general approach is presented in Figure 1 and integrates text and visual processing using transformer models to extract features and perform classifications. Texts (tweets) are preprocessed and fed into a transformer model to generate text embeddings, while images (memes) are processed through a vision transformer model to produce visual embeddings. Annotator features are extracted from the text embeddings, and a classifier is trained using these features along with the text and visual embeddings. An ensemble technique is applied to combine the outputs of the models, enhancing the accuracy of the classifier. The performance is then evaluated across several specific tasks to ensure comprehensive assessment and optimization of the results.

In the domain of text analysis in Spanish, our dataset was meticulously constructed, comprising 2526 samples for training purposes, with an additional 639 samples reserved for validation, and a final set of 490 samples designated for comprehensive testing. Conversely, for the English language domain, our dataset consisted of 1832 meticulously curated samples for training. An additional 574 samples were allocated for validation. Furthermore, our local test set, comprising 978 meticulously selected samples, served as a crucial benchmark for evaluating the generalization capabilities of our models in real-world applications. The metrics used for each task are as follows: For Task 1 and Task 4, we employed the ICM-Hard Norm F1-score for the positive class (sexism). For Task 2, Task 3, Task 5, and Task 6, we used the ICM-Hard Norm F1-score macro, which is the average of the F1-scores for all classes.



**Figure 2:** Leveraging Annotator Consensus and Profiles for Sexism Detection in Tweets.

## 4. Sexism Detection in Tweets

Firstly, for the detection of sexism in tweets, we focus on integrating annotator information, particularly considering their profiles such as gender and age, as summarized in Figure 2.

- **Text Preprocessing:** Mentions within the tweets were substituted with '@USER,' while any URLs were replaced with 'HTTPURL.'
- **Transformer Model:** We decided to use pre-trained models specifically for tweets: "cardiffnlp/twitter-roberta-base-sentiment" for English and "pysentimiento/robertuito-base-uncased" for Spanish, both from Hugging Face, since these models were trained with data in the respective languages.
- **Ensemble:** We employed two different ensembles. The first ensemble used a majority vote from the outputs of six different models, one for each annotator. The second ensemble used a majority vote from the outputs of five different models, focusing on gender and age (females, males, 18-22, 23-45, and 46+).

As mentioned in the previous Section, our runs for the first three text-focused tasks were:

1. **Run 1:** An ensemble was created from the outputs of five different models, focusing on gender and age. A majority vote was taken from the outputs of these five models, with the label being assigned if three or more groups agreed.
2. **Run 2:** An ensemble was created from the outputs of six different models, one for each annotator. A majority vote was taken from the outputs of these six models, with the label being assigned if four or more annotators agreed.
3. **Run 3:** A majority vote was taken initially from the six annotators' inputs, serving as the baseline. Similarly, the label was assigned if four or more annotators agreed.

To ensure a decision was always made in each ensemble without ties, we used probabilistic voting rather than hard voting from each model. This means that even if three models classify a tweet as sexist and three do not, the probabilities are compared, and the decision is made based on the highest probability, ensuring a definitive decision for all predictions in the ensembles.

For Task 2, which requires determining the intention of the tweets (single label), the label was assigned based on the highest probability prediction among the types of intentions if the tweet was sexist. To achieve this, a binary model was trained for each label. This approach ensures that the classification is both precise and comprehensive, taking into account the nuanced nature of the intentions expressed in the tweets.

For Task 3, which involves identifying the types of sexism in a tweet (multi-label), the ensemble takes into account all types of sexism indicated by the annotators. For example, if one annotator labels a tweet

as objectification and another labels it as misogynistic and sexual violence, all three types of sexism are included in our ensemble prediction. Furthermore, we employ multiple binary classification models for each label, allowing us to address each facet of identified sexism with specificity and precision.

To analyze in depth the impact of considering individualized annotators' opinions (A1-A6), grouped opinions before the classification models (All), opinions by demographic group (Females, Males, 18-22, 23-45, 46+), or assembled at the end (Ensemble Annotators, Ensemble Groups), Figure 3 presents the results based on the performance of the sexism identification, intention, and categorization models, respectively.

The selection of group ensemble and annotator ensemble approaches as Run 1 and Run 2, respectively, is grounded in their ability to integrate a wide range of perspectives and individual judgments. The group ensemble, by combining different demographics, offers an enriched and balanced overview, which is crucial for tackling the complexity of the tasks at hand. On the other hand, the annotator ensemble capitalizes on the diversity of individual judgments, ensuring a robust and competitive performance. Finally, the direct majority vote of annotators is established as the baseline (Run 3) due to its simplicity and effectiveness, providing a clear reference for evaluating ensemble methods. These choices are backed by the best results obtained in each task, where the group ensemble consistently outperforms others in terms of performance and ability to capture the inherent complexity in the datasets.

## 5. Sexism Detection in Memes

We chose a different path for tasks 4, 5, and 6 as shown in Figure 4. Although we leveraged the annotator data from the dataset, the text preprocessing steps depicted in Figure 1 were not applied. Instead, embeddings were extracted directly from the raw data using different approaches. We utilized CLIP embeddings for the memes and text along with annotator features. It was decided to address the tasks from the textual domain due to the high variability in the representation and graphic styles of the Memes (see examples in Table 1).

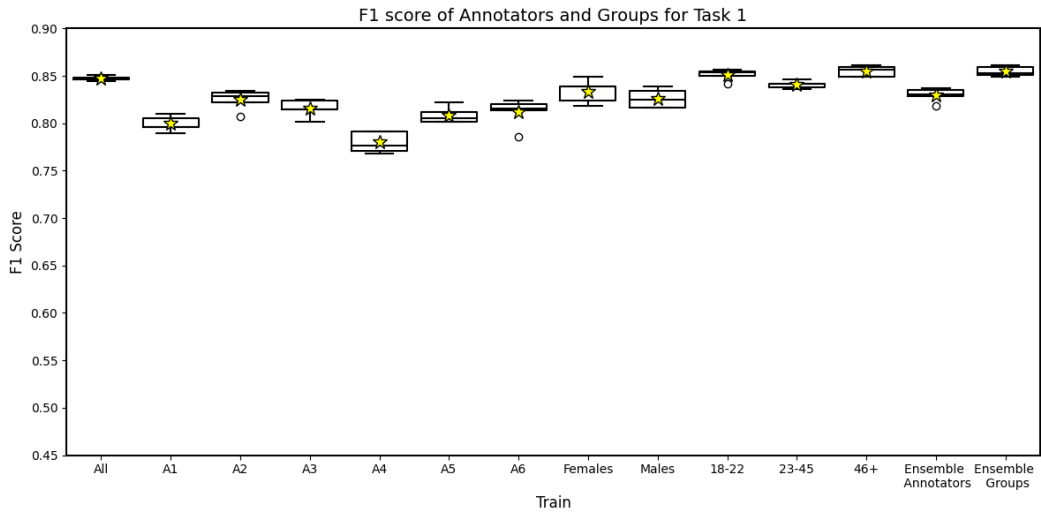
In Runs 1 and 2, annotator features were represented using one-hot encoding for gender, age range, ethnicity, study level, and country. The reader can explore this approach in the subsection 5.1. In Run 3, a descriptive text was created for the annotator features, from which embeddings were extracted. For Runs 2 and 3, the classifier used was a feed-forward neural network (FNN) with two hidden layers, containing 4096 and 512 neurons, respectively. Following these layers, a dropout layer with a dropout rate of 0.1 was applied. In Run 3, we further leverage the annotation and meme relationship and propose a Factorization Machine model, a Collaborative filtering technique, to predict the annotation based on annotator and meme CLIP embeddings. We explain more about this approach in the subsection 5.2.

### 5.1. Feed-forward neural network with CLIP embeddings

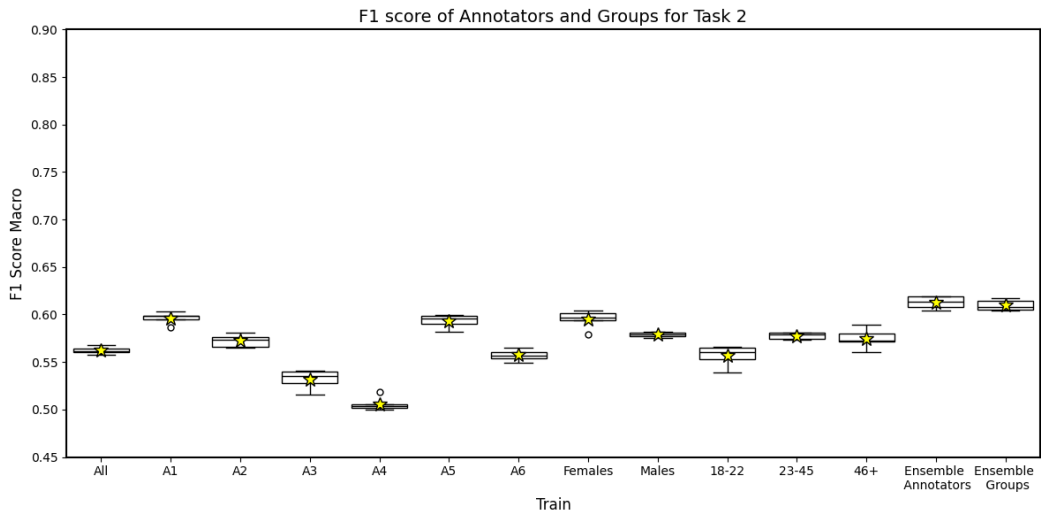
For Task 4, the output layer in the FNN consisted of a single neuron to produce the probability of the sexism class. We evaluated various approaches for the model: using only text embeddings, using only image embeddings, using both text and image embeddings, utilizing a general model (without annotator characteristics), and some combinations of the outputs of some of these models. Table 2 outlines the features of each evaluated model.

It is essential to define two concepts: early fusion and late fusion. In early fusion, the model simultaneously receives both text and image embeddings, meaning the model's input includes annotator features, text embeddings, and image embeddings (as seen in the "Text+Image" and "Text+Image General" models). In late fusion, the outputs of two models are combined. For example, in the "Text|Image" model, the outputs of the "Text" model (trained only with text embeddings) and the "Image" model (trained only with image embeddings) are combined by averaging their outputs. Similarly, the "Text|Image & Text|Image General" model combines the outputs of the "Text|Image" and "Text+Image General" models by averaging their outputs.

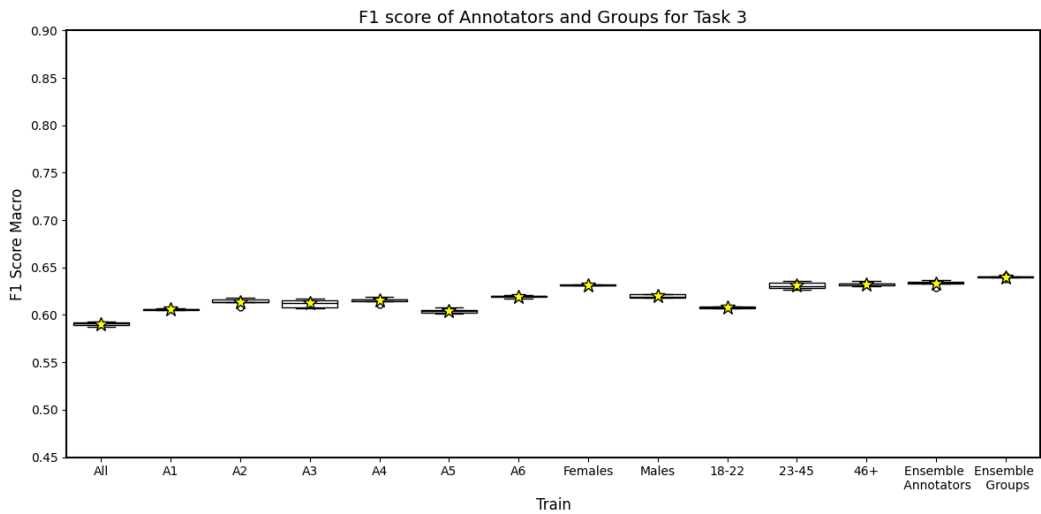
Figure 5 displays the F1 score for the positive case (sexism) across 10 runs for each model in Task 4. The results indicate that the "Text|Image" model and the "Text|Image & Text|Image General" model



(a) Task 1: Sexism Identification in Tweets.



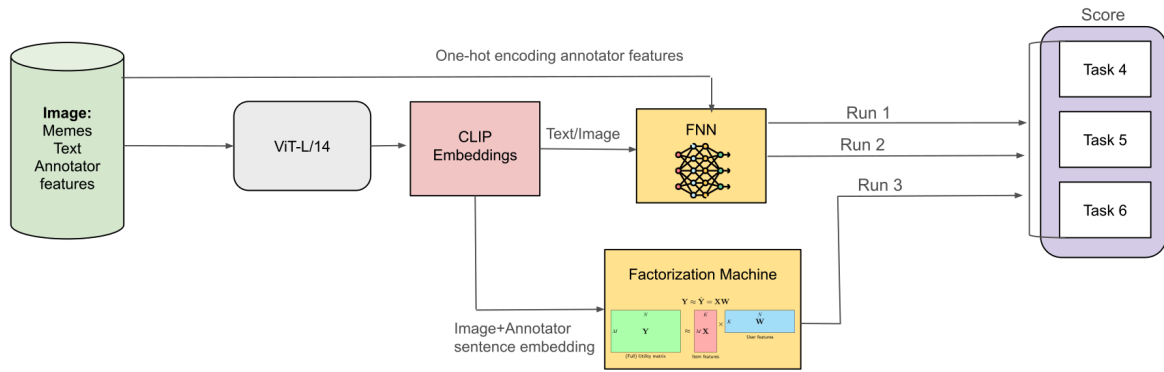
(b) Task 2: Source Intention in Tweets.



(c) Task 3: Sexism Categorization in Tweets.

Figure 3: Classification results for Sexism Detection in Tweets





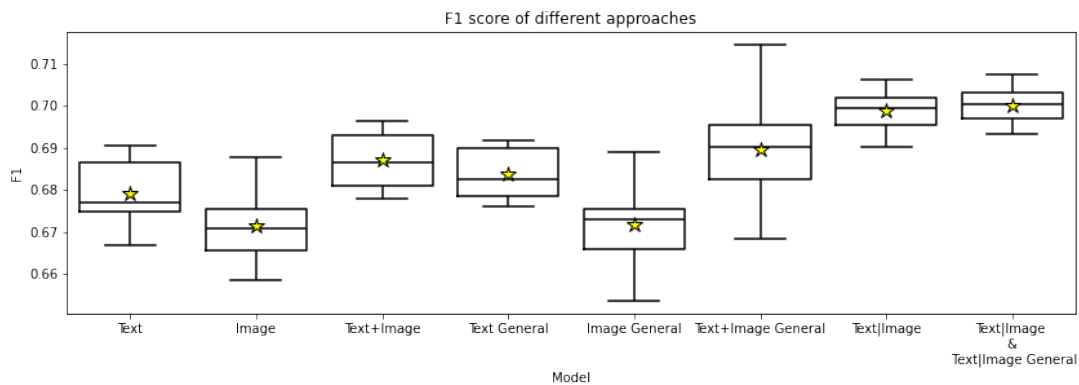
**Figure 4:** Leveraging Annotator Consensus and Profiles for Sexism Detection in Memes.

**Table 2**

Features of different models for the task 4.

Model Name	Annotator Features	Text Embeddings	Image Embeddings	Early Fusion	Late Fusion
Text	Yes	Yes	No	N/A	N/A
Image	Yes	No	Yes	N/A	N/A
Text+Image	Yes	Yes	Yes	Yes	No
Text General	No	Yes	No	N/A	N/A
Image General	No	No	Yes	N/A	N/A
Text+Image General	No	Yes	Yes	Yes	No
Text Image	Yes	Yes	Yes	No	Yes
Text Image & Text Image General	Yes&No	Yes	Yes	No	Yes

achieve higher mean F1 scores and low variations of the performance. These models correspond to Run 1 and Run 2, respectively.

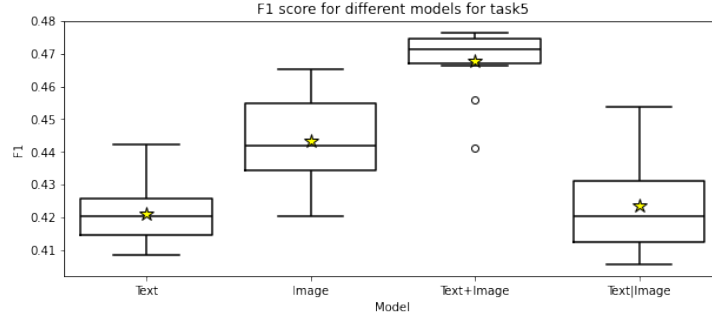


**Figure 5:** Classification results of different approaches for task 4.

For task 5, the output layer in the FNN consisted of a 3 neurons to yield the probability of every label. Similar to task 4, we evaluated some approaches for the model: using only text embeddings, using only image embeddings, using both text and image embeddings, and a combination of the outputs of “Text” and “Image” models by averaging their outputs. Figure 6 displays the F1 score macro across 10 runs for each model in Task 5.

The results in Figure 6 indicate the “Text+Image” model achieves higher mean F1 scores and low variations of the performance. These models correspond to Run 1. For Run 2, the “Text|Image” model was selected. Although it did not achieve the highest F1 score, it demonstrated a strong MSE score comparable to the “Text+Image” model.

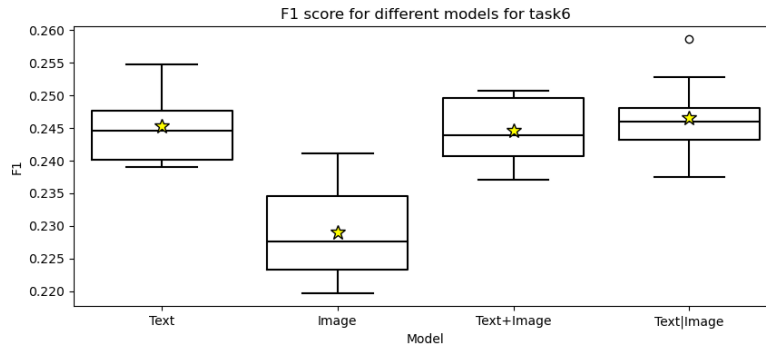
For Task 6, the output layer in the FNN consisted of one neuron to yield the probability for each



**Figure 6:** Classification results of different approaches for task 5.

label of sexism categorization. We created 5 models, each trained exclusively on data from sexism memes and with a random subset of training data for negative cases equal to the amount of training data for positive cases. Consequently, each model was trained on a balanced dataset. The probability output from the model for Task 4 was used to determine the probability of the not sexism label and then combined with the outputs from each of these 5 models to produce a final prediction.

There are two exceptional cases to consider: i) If the probability of not sexism is higher than 0.5, as well as one of the 5 categories of sexism, the final prediction is always not sexism. ii) If the probability of not sexism is lower than 0.5, as well as one of the 5 categories of sexism, the meme is classified as sexist, and the category of sexism with the highest probability is selected. Similar to Tasks 4 and 5, we evaluated various approaches for the model. Figure 7 presents the macro F1 scores for each model.



**Figure 7:** Classification results of different approaches for task 6.

We observed similar performance among the “Text”, “Text+Image”, and “Text|Image” models. Based on these results, we selected the “Text|Image” model for Run 1 and the “Text+Image” model for Run 2. The “Text” model was not chosen, as we believe that the combination of text and image embeddings yields better results.

## 5.2. Multimodal Collaborative Filtering employing CLIP embeddings and Factorization Machines

**Loni2018FactorizationMF** In this approach, we model similar to how to assign a score in a recommendation system or to predict links between nodes in a bipartite graph, leveraging the fact that we have the annotator and the item features. Given known subject-item preferences, predict new subject-item preferences. Formally, let  $U$  a set of all subjects and  $V$  a set of all items, our core task is to find a real-valued scalar function  $score(u, v)$  where  $u \in U$  and  $v \in V$ . To provide a hard label or multi-label,  $k$  subjects vote with their encoded scores. Hence, we’ve reduced our problem into a score prediction problem. For each user  $u \in U$ , let  $u \in \mathbb{R}^D$  for its  $D$ -dimensional embedding. For each item  $v \in V$ , let  $v \in \mathbb{R}^D$  be its  $D$ -dimensional embedding. So,  $score(u, v) \equiv f : \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}$ . In this approach,

memes and annotators are transformed into the same embedding space using CLIP. Specifically, user demographics such as age, gender, and interests are encoded with a phrase such as “A female aged 18-22, of Hispanic or Latino ethnicity, with a high school degree or equivalent, and located in Mexico” into one CLIP embedding. In contrast, the meme, which may include both image and text components, is encoded into another CLIP embedding. These embeddings capture the nuanced features of both the user and the meme content. We then concatenate these two embeddings into a single embedding that represents the combined features of the user and the meme.

For instance, the table 3 illustrates a complete utility matrix for Task 4 with known score entries  $f(u, v)$  where 0 represents the label “NO” and 1 represents the label “YES”. Our encoding method ignores "UNKNOWN" labels, but other encodings are possible. In this case, the voting policy is the selection of the class annotated by more than 3 subjects.

**Table 3**

An example of utility matrix for the task 4

	$V_1$	$V_2$	$V_3$
$U_1$	1	0	1
$U_2$	0	1	1
$U_3$	1	0	0
$U_4$	1	1	1
$U_5$	0	0	0
$U_6$	1	0	0
<i>Voting</i>	1	0	Undefined
<i>Label</i>	YES	NO	

For Task 5, the *score* function is encoded similarly to task 4, with the addition of a voting policy and a method to define similarity to hard labels. The voting policy is the arithmetic mean of votes  $\overline{score}$  which entails us into the encoding to predict the hard label as follows

$$\overline{score} \in [0, 0.67] \implies \text{No}$$

$$\overline{score} \in (0.67, 1.34] \implies \text{Direct}$$

$$\overline{score} \in (1.34, 2] \implies \text{Judgemental}$$

We apply softmax over the votes to find the probabilities, thus solving the soft-soft task.

For Task 6, the different combinations are encoded into a compact bit set as follows: each  $label_i$  is a bit  $2^i$  where  $i \geq 0$ . The union gives us the bit set across the different combinations. We provide an example below:

$$score(u, v) = 0b000001 \implies -$$

$$score(u, v) = 0b000010 \implies \text{IDEOLOGICAL-INEQUALITY}$$

$$score(u, v) = 0b000100 \implies \text{MISOGYNY-NON-SEXUAL-VIOLENCE}$$

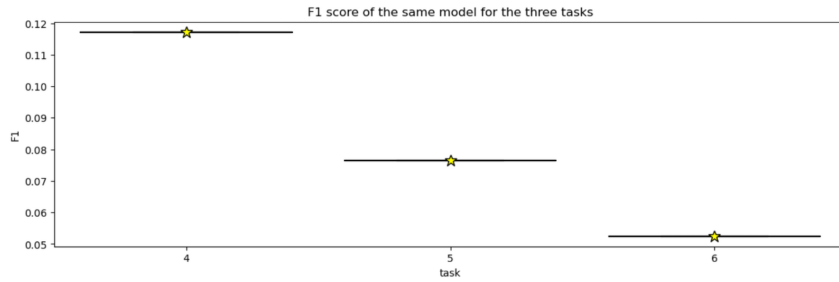
$$score(u, v) = 0b00010|0b00001 = 0b00011 \implies -, \text{IDEOLOGICAL-INEQUALITY}$$

Similarly to Task 5, we count the number of common bits and apply softmax to find the probability distribution.

We’ve defined how to decode *score* to solve the tasks, but how can we learn *score* from annotator and memes CLIP embeddings? Different embedding-based models include memory-based CF, model-based CF, Neighborhood methods, Neural Graph Collaborative Filtering, Factorization Machines [8], and GCN-based CF. Among those, the Factorization Machine models stands out for being efficient and accurate, enabling it to effectively predict the score [9] from concatenated embedding. Figure 8 shows how well this approach performs on our validation dataset after 10 runs.

## 6. Outcomes of the Evaluation Phase

Table 4 presents the combined results for both English and Spanish submissions in the sexism detection challenge across six different tasks. Each task involves several runs evaluated using two metrics: Hard-Hard and Soft-Soft. Below, we describe the results, focusing on the best runs for each task.



**Figure 8:** F1 score of hard-hard task 4, 5, 6 employing Collaborative Filtering.

For Task 1 (Tweets), the best performance was achieved by running MMICI\_3, which ranked 17th in the Hard-Hard metric with an ICM-Hard Norm of 0.7676, and an F1 score of 0.7637. In the Soft-Soft metric, this run ranked 21st with an ICM-Soft Norm of 0.5736, indicating it was the most effective in both metrics for this task.

For Task 4 (Memes), run MMICI\_2 excelled, ranking 8th in the Hard-Hard metric with an ICM-Hard Norm of 0.5515, and an F1 score of 0.7261. For Task 5, the top run was MMICI\_1, which ranked 7th in the Hard-Hard metric with an ICM-Hard Norm of 0.3934, and an F1 score of 0.4179. In the Soft-Soft metric, this run performed even better, ranking 2nd with an ICM-Soft Norm of 0.3654, making it the most effective in both categories. Lastly, for Task 6, the best run was MMICI\_1, which ranked 3rd in the Hard-Hard metric with an ICM-Hard Norm of 0.2954, and an F1 score of 0.4342.

**Table 4**

Results of Submission on Leaderboard for both Spanish and English (ALL)

Task	Run	Hard-Hard				Soft-Soft		
		Ranking	ICM-Hard	ICM-Hard Norm	F1	Ranking	ICM-Soft	ICM-Soft Norm
Task1	MMICI_1	31	0.4705	0.7365	0.7455	29	-0.3394	0.4456
Task1	MMICI_2	28	0.4780	0.7402	0.7460	30	-0.3622	0.4419
Task1	MMICI_3	<b>17</b>	<b>0.5324</b>	<b>0.7676</b>	<b>0.7637</b>	21	0.4589	0.5736
Task2	MMICI_1	27	-0.0987	0.4679	0.4548	24	-4.5753	0.1314
Task2	MMICI_2	32	-0.2406	0.4218	0.4383	25	-4.6285	0.1271
Task2	MMICI_3	28	-0.1076	0.4650	0.4525	20	-3.6350	0.2071
Task3	MMICI_1	27	-1.4509	0.1631	0.4026	22	-7.9356	0.0809
Task3	MMICI_2	28	-1.5003	0.1516	0.4017	23	-7.9380	0.0808
Task3	MMICI_3	23	-0.8105	0.3118	0.4805	20	-7.6413	0.0965
Task4	MMICI_1	12	0.0751	0.5382	0.7202	17	-0.6189	0.4005
Task4	MMICI_2	<b>8</b>	<b>0.1014</b>	<b>0.5515</b>	<b>0.7261</b>	16	-0.6183	0.4006
Task4	MMICI_3	24	-0.0361	0.4816	0.6781	19	-0.6410	0.3970
Task5	MMICI_1	7	-0.3066	0.3934	0.4179	<b>2</b>	<b>-1.2660</b>	<b>0.3654</b>
Task5	MMICI_2	10	-0.3868	0.3655	0.3770	3	-1.3738	0.3539
Task5	MMICI_3	8	-0.3297	0.3854	0.3814	13	-3.4751	0.1304
Task6	MMICI_1	<b>3</b>	<b>-0.9863</b>	<b>0.2954</b>	<b>0.4342</b>	19	-16.1248	0.0000
Task6	MMICI_2	7	-1.3446	0.2210	0.4453	20	-19.3246	0.0000
Task6	MMICI_3	24	-3.8341	0.0000	0.2347	21	-45.0237	0.0000

The results for the Spanish submissions are showcased in Table 5. Hereafter, we delve into these outcomes, centering our attention on the most successful executions for each task. For Task 1, the best performance was achieved by running MMICI\_3, which ranked 10th in the Hard-Hard metric with an ICM-Hard Norm of 0.7802, and an F1 score of 0.7892. In Task 2, the best run was MMICI\_1, ranking 15th in the Hard-Hard metric with an ICM-Hard Norm of 0.5522, and an F1 score of 0.5133. For Task 3, the top run was MMICI\_1, ranking 13th in the Hard-Hard metric with an ICM-Hard Norm of 0.4586, and an F1 score of 0.5486. In Task 4, run MMICI\_2 excelled, ranking 14th in the Hard-Hard metric with an ICM-Hard Norm of 0.4900, and an F1 score of 0.6997. For Task 5, the top run was MMICI\_1, which

ranked 7th in the Hard-Hard metric with an ICM-Hard Norm of 0.3945, and an F1 score of 0.4198. In the Soft-Soft metric, this run performed even better, ranking 1st with an ICM-Soft Norm of 0.3461, making it the best-performing in both categories. Lastly, in Task 6, the best run was MMICI\_1, which ranked 4th in the Hard-Hard metric with ICM-Hard Norm of 0.2473, and an F1 score of 0.3868.

**Table 5**  
Results of Submission on Leaderboard for Spanish

Task	Run	Hard-Hard				Soft-Soft		
		Ranking	ICM-Hard	ICM-Hard Norm	F1	Ranking	ICM-Soft	ICM-Soft Norm
Task1	MMICI_1	16	0.5323	0.7662	0.7817	24	0.0894	0.5143
Task1	MMICI_2	22	0.5007	0.7504	0.7705	25	0.0170	0.5027
Task1	MMICI_3	<b>10</b>	<b>0.5603</b>	<b>0.7802</b>	<b>0.7892</b>	15	0.6706	0.6076
Task2	MMICI_1	<b>15</b>	<b>0.1670</b>	<b>0.5522</b>	<b>0.5133</b>	23	-4.1728	0.1658
Task2	MMICI_2	26	0.0064	0.5020	0.4933	24	-4.2127	0.1626
Task2	MMICI_3	29	-0.1146	0.4642	0.4779	20	-3.4962	0.2200
Task3	MMICI_1	<b>13</b>	<b>-0.1853</b>	<b>0.4586</b>	<b>0.5486</b>	24	-7.8261	0.0927
Task3	MMICI_2	14	-0.2269	0.4493	0.5446	25	-7.8356	0.0922
Task3	MMICI_3	22	-0.5870	0.3689	0.5165	22	-7.4291	0.1134
Task4	MMICI_1	17	-0.0591	0.4699	0.6906	14	-0.6655	0.3939
Task4	MMICI_2	<b>14</b>	<b>-0.0196</b>	<b>0.4900</b>	<b>0.6997</b>	15	-0.6689	0.3933
Task4	MMICI_3	26	-0.1848	0.4059	0.6470	18	-0.8361	0.3667
Task5	MMICI_1	7	-0.3028	0.3945	0.4198	<b>1</b>	<b>-1.4813</b>	<b>0.3461</b>
Task5	MMICI_2	9	-0.4077	0.3580	0.3728	2	-1.5486	0.3392
Task5	MMICI_3	10	-0.4875	0.3302	0.3545	13	-4.0400	0.0804
Task6	MMICI_1	<b>4</b>	<b>-1.2346</b>	<b>0.2473</b>	<b>0.3868</b>	18	-14.9495	0.0000
Task6	MMICI_2	13	-1.6925	0.1536	0.4141	20	-18.0902	0.0000
Task6	MMICI_3	24	-3.8686	0.0000	0.2225	21	-42.6540	0.0000

The outcomes for the English submissions are outlined in Table 6. We elaborate on these results, specifically highlighting the top performances for each task. In Task 4, run MMICI\_2 excelled, ranking 3rd in the Hard-Hard metric with an ICM-Hard Norm of 0.6129, and an F1 score of 0.7559. For Task 5, the top run was MMICI\_3, which ranked 1st in the Hard-Hard metric with an ICM-Hard Norm of 0.4413, and an F1 score of 0.4094. Lastly, in Task 6, the best run was MMICI\_1, which ranked 2nd in the Hard-Hard metric with an ICM-Hard Norm of 0.3419, and an F1 score of 0.4726.

The strong results achieved with memes can be attributed to the use of CLIP (Contrastive Language-Image Pre-training) embeddings. CLIP effectively learns visual concepts from natural language descriptions, aligning images and text within a shared embedding space. This alignment is achieved by training on a vast dataset of images paired with their corresponding textual descriptions, enabling the model to understand and relate visual and textual information seamlessly. Using CLIP, Vision Transformers can be employed for image encoding and Text Transformers for text encoding, resulting in a unified model that excels in multi-modal tasks. The Vision Transformer processes the image data, while the Text Transformer processes the text data. Both sets of embeddings are then projected into a common space where their similarities can be measured and aligned, allowing the model to leverage the strengths of both visual and textual information effectively. This approach enabled the extraction of sexist expressions from memes in the dataset across both languages. By transferring the representation to the textual domain, it became possible to adopt state-of-the-art techniques for the classification tasks.

In summary, the combined analysis of English and Spanish submissions in the sexism detection challenge illuminates diverse approaches and performances across tasks. Each language cohort showcased distinct strengths, with notable runs such as MMICI\_1 and MMICI\_3 consistently demonstrating effectiveness across multiple tasks. These results underscore the complexity of sexism detection and highlight the importance of multilingual evaluation frameworks. Further exploration and refinement of these methodologies promise continued advancements in combating bias and fostering inclusivity in online content.



**Table 6**  
Results of Submission on Leaderboard for English

Task	Run	Hard-Hard				Soft-Soft		
		Ranking	ICM-Hard	ICM-Hard Norm	F1	Ranking	ICM-Soft	ICM-Soft Norm
Task1	MMICI_1	40	0.3840	0.6960	0.6971	32	-0.8805	0.3586
Task1	MMICI_2	33	0.4402	0.7246	0.7141	31	-0.8349	0.3659
Task1	MMICI_3	25	0.4912	0.7507	0.7315	21	0.1413	0.5227
Task2	MMICI_1	33	-0.4572	0.3418	0.3680	23	-5.0641	0.0861
Task2	MMICI_2	36	-0.5728	0.3018	0.3570	24	-5.1264	0.0810
Task2	MMICI_3	30	-0.1384	0.4521	0.4087	19	-3.8024	0.1892
Task3	MMICI_1	32	-2.8962	0.0000	0.2357	22	-7.9094	0.0666
Task3	MMICI_2	34	-2.9573	0.0000	0.2373	21	-7.9059	0.0668
Task3	MMICI_3	26	-1.1024	0.2298	0.4287	19	-7.7476	0.0755
Task4	MMICI_1	5	0.2094	0.6063	0.7538	20	-0.5779	0.4062
Task4	MMICI_2	<b>3</b>	<b>0.2224</b>	<b>0.6129</b>	<b>0.7559</b>	19	-0.5735	0.4069
Task4	MMICI_3	18	0.1131	0.5574	0.7122	17	-0.4621	0.4250
Task5	MMICI_1	6	-0.3112	0.3920	0.4156	2	-1.1089	0.3790
Task5	MMICI_2	8	-0.3657	0.3731	0.3815	3	-1.2447	0.3642
Task5	MMICI_3	<b>1</b>	<b>-0.1691</b>	<b>0.4413</b>	<b>0.4094</b>	13	-2.9704	0.1760
Task6	MMICI_1	<b>2</b>	<b>-0.7441</b>	<b>0.3419</b>	<b>0.4726</b>	15	-18.3643	0.0000
Task6	MMICI_2	7	-1.0095	0.2855	0.4752	16	-21.6764	0.0000
Task6	MMICI_3	20	-3.8687	0.0000	0.2447	17	-49.2040	0.0000

## 7. Conclusion

This paper has detailed MMICI’s participation in the EXIST shared task at CLEF 2024, focusing on the detection and categorization of sexism in social media content. By leveraging various innovative methodologies, including ensemble approaches that incorporate diverse annotator profiles and multi-modal embeddings, our models have demonstrated substantial efficacy in identifying and understanding sexism in both tweets and memes.

The results of our evaluation phase reveal that our ensemble methods, particularly those combining annotator profiles with text and image embeddings, achieve robust performance across multiple tasks. Specifically, our runs have shown competitive results in detecting sexism, discerning the intent behind sexist content, and categorizing different types of sexism. For instance, the ensemble approaches used in Runs 1 and 2 consistently outperformed traditional majority voting methods, highlighting the value of integrating diverse perspectives in addressing complex subjective tasks like sexism detection. Our approach emphasizes the importance of considering individual annotator characteristics, such as gender and age, to ensure that our models capture a wide range of viewpoints and avoid silencing minority voices. In most tasks, our baseline strategy performed the best. However, for tasks 2 and 3 in Spanish, our ensembles surpassed the baseline by capturing a broader range of perspectives. This nuanced understanding of sexism, facilitated by advanced machine learning techniques and diverse data representation, is crucial for effectively combating sexist behaviors and discourses online.

In related work, there is significant potential in exploring additional data collected on annotators in the EXIST 2024 dataset, including their ethnicities, study levels, and countries of origin, to enhance the cross-lingual and cross-cultural analysis capabilities of sexism detection systems. Developing models that effectively handle multiple languages and cultural contexts, possibly through cross-lingual transfer learning and the creation of culturally nuanced models, would improve global applicability. Additionally, further exploration of Transformer-based models and the creation of ensembles can leverage their strengths to improve detection accuracy. Expanding the dataset to include more diverse and underrepresented demographic groups would also contribute to building more robust and generalizable models. This could involve collecting additional annotated data from various social media platforms and cultural contexts. Moreover, improving multimodal techniques by leveraging advanced neural network architectures and incorporating additional features can further enhance model performance in

detecting sexism.

Overall, our participation in the EXIST task underscores the potential of advanced ensemble methods and multimodal analysis in improving the detection and categorization of sexism in social media. These methods not only enhance the accuracy of automatic tools but also contribute to a deeper understanding of how sexism manifests in various forms, thereby supporting broader efforts to promote gender equity and reduce discrimination in digital spaces.

## Acknowledgments

This work has been partially supported by CONAHCYT (The National Council of Humanities, Sciences, and Technologies of Mexico), which promotes scientific and technological development in the country.

Additionally, we acknowledge the support provided through the following scholarships: Martha Paola Jimenez-Martinez (scholarship number 828539) and Joan Manuel Raygoza-Romero (scholarship number 806073).

## References

- [1] Cambridge, Sexism, 2024. <https://dictionary.cambridge.org/dictionary/english/sexism>.
- [2] Real Academia Española, Sexismo, 2024. <https://dle.rae.es/sexismo>.
- [3] Comisión Nacional para Prevenir y Erradicar la Violencia Contra las Mujeres , ¿qué es el lenguaje sexista y por qué es importante visibilizarlo?, 2016. <https://www.gob.mx/conavim/articulos/que-es-el-lenguaje-sexista-y-por-que-es-importante-visibilizarlo?idiom=es>.
- [4] P. Glick, S. T. Fiske, Ambivalent sexism, in: *Advances in experimental social psychology*, volume 33, Elsevier, 2001, pp. 115–188.
- [5] L. Plaza, J. Carrillo-de-Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of EXIST 2024 – Learning with Disagreement for Sexism Identification and Characterization in Social Networks and Memes, in: *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024)*, 2024.
- [6] J. A. García-Díaz, M. Cánovas-García, R. Colomo-Palacios, R. Valencia-García, Detecting misogyny in spanish tweets. an approach based on linguistics features and word embeddings, *Future Generation Computer Systems* 114 (2021) 506–518.
- [7] S. Akhtar, V. Basile, V. Patti, Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection, *arXiv preprint arXiv:2106.15896* (2021).
- [8] B. Loni, M. Larson, A. Hanjalic, Factorization machines for data with implicit feedback, 2018. URL: <https://api.semanticscholar.org/CorpusID:56517380>.
- [9] S. Rendle, Factorization machines, in: *2010 IEEE International Conference on Data Mining, 2010*, pp. 995–1000. doi:10.1109/ICDM.2010.127.