Procesamiento del Lenguaje Natural, Revista nº 74, marzo de 2025, pp. 15-31 recibido 20-12-2024 revisado 27-01-2025 aceptado 06-02-2025

# Identification of Racial and Sexist Stereotypes in Spanish: A Learning with Disagreements Approach

## *Identificación de Estereotipos Raciales y Sexistas en Español: Un Enfoque Basado en Aprendizaje con Desacuerdos*

**Elias Urios Alacreu,**[1] **Paolo Rosso**[1,2]
[1]PRHLT Research Center, Universitat Politècnica de València, Valencia, Spain
[2]ValgrAI Valencian Graduate School and Research Network of Artificial Intelligence, Spain
elural2@prhlt.upv.es, prosso@dsic.upv.es

**Abstract:** Hate speech has proliferated significantly in recent years, largely driven by the widespread adoption of social media platforms. Hate speech often operates implicitly, leveraging subtle stereotypes to propagate discriminatory views. These covert mechanisms allow harmful content to disguise itself, making detection increasingly complex. As a result, tackling hate speech has become an urgent priority, driving the widespread adoption of deep learning models to detect and combat harmful content. Given the inherently subjective nature of hate speech and its nuanced manifestations, there is a need to develop models that are as generalizable as possible. This has led to the emergence of the learning with disagreements paradigm, which aims to introduce disagreements within the task itself to enhance model generalizability. This paper investigates the latter paradigm through two shared tasks. The first task, DETEST-Dis, explores stereotypes against immigrants in online comments and was organized at IberLEF 2024. Our results are among the best of all participating teams, surpassing traditional approaches. The second task, EXIST, focuses on sexism in memes and was organized at CLEF 2024. Here, our performance is enhanced by adding features from an external model as well as data augmentation. Our source code can be found on `https://github.com/Buzzeitor30/DETESTS-DIS` and `https://github.com/Buzzeitor30/EXIST-TFM`.
**Keywords:** Learning with disagreements, racial and sexist stereotypes, LLMs, Transformers.

**Resumen:** El discurso del odio ha proliferado significativamente en los últimos años, en gran medida impulsado por la adopción generalizada de plataformas de redes sociales. El discurso de odio a menudo opera de manera implícita, aprovechando estereotipos sutiles para propagar pensamientos discriminatorios. Estos mecanismos encubiertos han permitido que el contenido odioso se oculte a si mismo, haciendo que su detección resulte cada vez más compleja. Como resultado, la lucha contra los discursos de odio se ha convertido en una prioridad urgente, impulsando la adopción generalizada de modelos de aprendizaje profundo para detectar y combatir contenidos nocivos. Dada la naturaleza inherentemente subjetiva de los discursos del odio, es necesario desarrollar modelos que sean lo más generalizables posible. Esto ha llevado a la aparición del paradigma de aprendizaje con desacuerdos, que tiene como objetivo introducir desacuerdos dentro de la propia tarea para mejorar la generalizabilidad del modelo. Este trabajo investiga este último paradigma a través de dos *shared tasks*. La primera tarea, DETEST-Dis, explora los estereotipos contra los inmigrantes en comentarios en línea y fue organizada en IberLEF 2024. Nuestros resultados se encuentran entre los mejores de todos los equipos participantes, superando los enfoques tradicionales. La segunda tarea, EXIST, se centra en el sexismo en los memes y fue organizada en CLEF 2024. En este caso, nuestro rendimiento presenta una mejoría añadiendo características de un modelo externo así como también *data augmentation*. Nuestro código fuente se puede encontrar en `https://github.com/Buzzeitor30/DETESTS-DIS` y `https://github.com/Buzzeitor30/EXIST-TFM`
**Palabras clave:** Aprendizaje con desacuerdos, detección de estereotipos raciales y sexistas, LLMs, Transformers.

ISSN 1135-5948 DOI 10.26342/2025-74-1 

## 1 Introduction

Hate Speech (HS) can be defined as any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics (Nockleby, 2000). Historically, it has been a persistent issue in traditional media, where HS has often shaped public discourse. However, with the emergence of social media platforms, characterized by anonymity and their global, instantaneous reach, the spread of HS has escalated dramatically, making its detection and mitigation more crucial than ever (Rawat, Kumar, and Samant, 2024; Gandhi et al., 2024).

In addition, since the sheer volume of data generated by these platforms cannot be handled manually, the usage of automatic HS detection methods, particularly those leveraging Large Language Models (LLMs) based on the Transformer architecture (Vaswani, 2017), have become the preferred approach for combating this issue (Subramanian et al., 2023).

Moreover, HS detection is often considered a challenging task, mainly due to the lack of a universally accepted definition of HS and the inherent difficulties in defining such a complex phenomenon, which can be conveyed through various forms of expression (Rawat, Kumar, and Samant, 2024; Gandhi et al., 2024).

One area of concern is the use of stereotypes, which are one example of how HS can be suggested in a implicit yet damaging way (Schmeisser-Nieto, Nofre, and Taulé, 2022; Schmeisser-Nieto et al., 2024b). Although stereotypes can target a wide variety of groups, two specific groups have been notably affected: women (FRA, 2023), who have historically faced stereotypes, and immigrants (OBERAXE, 2024), who have become a focus of negative rhetoric due to recent political events.

In addition, although HS has predominantly been studied in textual forms, there is growing recognition that HS now extends to multimodal communication. Hence, the detection of multimodal HS has emerged as a new challenge in the field (Rawat, Kumar, and Samant, 2024; Gandhi et al., 2024).

Within this multimodal landscape, memes represent a particularly concerning medium. Memes, often conceived as a source of positivity and entertainment, have become a form of expression to perpetuate HS (Kiela et al., 2020).

On the other hand, English is the predominant language used on most social media platforms and, therefore, HS detection research in this language is the most advanced. Nevertheless, as the phenomenon of HS spreads to other languages, such as Spanish, the need to collect resources to address this problem has become essential, especially in the multimodal area, where little to no work exists in languages other than English (Jahan and Oussalah, 2023).

Finally, a key challenge in HS detection is the subjectivity of the task (Subramanian et al., 2023), which leads to bias in datasets and models. To address this, researchers propose a new paradigm known as learning with disagreements (LeWiDi), which aims to create more generalized systems that reflect diverse perspectives and recognize that HS detection is not always binary (Uma et al., 2021).

Therefore, this work is aimed to study how the aforementioned learning paradigm affects HS detection tasks, specifically focusing on stereotypes against immigrants and women. Our experimentation encompasses both textual stereotypes analysis and explores the emerging phenomenon of stereotypes in memes. Although one of the shared tasks we participated in was held in both Spanish and English, our studies are primarily conducted in Spanish.

In addition, this paper aims to investigate the following research questions:

- **RQ1**: *How does the LeWiDi paradigm influence a classifier performance for detecting racial stereotypes in online comments and discussion forums?*
- **RQ2**: *How does the LeWiDi paradigm influence a classifier performance for detecting sexist stereotypes in memes?*

The rest of the paper is structured as follows. Firstly, some of the most relevant literature regarding HS detection in text and memes, as well as the LeWiDi paradigm, is reviewed. Next, we described the two selected tasks where the detection of racial and sexist stereotypes is addressed for our research. Following, we present the proposed model for each of the selected tasks. Next, we describe some key aspects of the conducted experimentation such as the hardware or the

training procedure. Afterwards, we present the results of each task and provide an analysis of them. Finally, we draw some conclusions and summarize the key findings of the study.

## *2 Related work*

### 2.1 Stereotype Identification in Text

One of the most early and influential work regarding HS detection in texts is presented in Waseem (2016), which focuses on sexism and racism identification. Subsequent work (Sánchez-Junquera et al., 2021; Chulvi et al., 2024) has analyzed the presence of stereotypes against immigrants in political speeches, whereas other research has focused on misogyny identification (Anzovino, Fersini, and Rosso, 2018).

Numerous shared tasks have been organized to tackle specific issues within HS, as well as sexist and racist stereotypes, detection in texts: HatEval, which focuses on the detection of HS against immigrants and women (Basile et al., 2019); EXIST, aimed at identifying sexism (Rodríguez-Sánchez et al., 2021; Rodríguez-Sánchez et al., 2022; Plaza et al., 2023; Plaza et al., 2024); DETEST, which addresses stereotypes against immigrants (Ariza-Casabona et al., 2022), and AMI, targeting the identification of misogyny (Fersini et al., 2018; Fersini, Nozza, and Rosso, 2020), among others.

### 2.2 Stereotype Identification in Memes

Memes, originally intended to convey humor, have become a new medium for spreading HS. Detecting HS in memes is particularly challenging, as it requires analyzing both the visual and textual elements to fully grasp the underlying content of the meme (Hermida and Santos, 2023). Despite the increasing prevalence of HS and the difficulty of the task, research in this area is still limited.

One of the first notable efforts in this field was the Hateful Memes Challenge (HMC) introduced by Facebook AI (Kiela et al., 2020). The challenge highlighted the importance of using pretrained Transformer architectures which combined both text and image features (Chen et al., 2020; Li et al., 2019) to tackle these tasks effectively. In addition, the winning team from the competition (Zhu, 2020), as well as subsequent research utilizing the HMC dataset, demonstrated the importance of adding extra features (such as image captioning or entity classification) from external models to boost model's performance on the task (Hermida and Santos, 2023).

Another important contribution is the Multimedia Automatic Misogyny Identification (MAMI) task (Fersini et al., 2022), which focuses on mysogyny. Akin to the HMC proposals, most participants utilized the Transformer architectures previously mentioned, although most of them integrated them into an ensemble approach (Fersini et al., 2022).

Among all the participants, only the DD-Tig team integrated extra features into their model input through image captioning (Zhou et al., 2022). Their results, however, did not improve by using this stand-alone technique, proving how these methods are strongly related to the external model ability to interpret the image (Zhou et al., 2022; Hermida and Santos, 2023).

Subsequent work in the MAMI dataset is presented in Rizzi et al. (2023), which proves that although the textual modality contains more useful information than the visual one, a multi-modal approach is required for this task.

### 2.3 LeWiDi: Learning With Disagreements

As stated previously, subjectiveness plays a key role in HS detection task (Subramanian et al., 2023), since it can influence both the quality of the resulting dataset as well as the model. In addition, subjectiveness is also associated with disagreements, since various annotators can offer different perspectives for the same sample, which is especially important in a subjective task as HS detection.

The classic approach for dealing with disagreements assumes the existence of a single, objective label, known as the **gold label**, which can be extracted through a majority voting scheme or statistical methods (Dawid and Skene, 1979). Although this methodology is simple and effective, it is also true that it ignores the opinion of the minority over the majority, hence neglecting other subjective points of view and the disagreements themselves (Uma et al., 2021).

Nevertheless, other researchers suggest that "disagreement is signal, not noise" (Aroyo and Welty, 2015). In other words,

disagreement provides useful information for learning, and should be leveraged into the the task (Uma et al., 2021).

Among all the approaches found in literature for dealing with disagreements (Uma et al., 2021), we highlight the following two.

On the one hand, we have the soft loss approach, considered the best for the LeWiDi paradigm. Although it also aggregates the annotations, it does it into a probability distribution, either via an empirical distribution or a softmax, known as **silver label**. The main goal behind this approach is to optimize our model distribution to resemble the original one produced by the annotators disagreement (Uma et al., 2020; Uma et al., 2021).

On the other hand, there is the perspectivist approach, which disregards aggregation and proposes to work directly with the individual annotations (Cabitza, Campagner, and Basile, 2023). According to Mostafazadeh Davani, Díaz, and Prabhakaran (2022), the multi-task architecture is the most effective one for this approach.

Finally, it is worth mentioning that LeWiDi tasks are often evaluated through two evaluation contexts:

1. The **hard evaluation**, which addresses how well does the model predict the gold label. The most common metrics used for this evaluation are F1-Score or Information Contrast Metric (ICM) (Amigo and Delgado, 2022).
2. The **soft evaluation**, which evaluates how well does the model generalize. The most common metric used for this type of evaluation is the Cross Entropy, although novel metrics such as ICM Soft have appeared.

## *3 Task descriptions*

### 3.1 Identification of racial stereotypes in text

DETESTS-Dis (DETEction and classification of racial STereotypes in Spanish - Learning with Disagreement) (Schmeisser-Nieto et al., 2024b) is the second edition of the DETEST shared task (Ariza-Casabona et al., 2022) organized at IberLEF 2024. Like the previous edition, this one is also focused on the detection and classification of racial stereotypes on comments on online news related content. Nevertheless, this new edition introduces two key changes by shifting the learning paradigm to LeWiDi in addition to the proposal of a novel task of stereotype classification based on its implicitness.

#### 3.1.1 Tasks

The DETEST-Dis shared task is divided itself into two sub-tasks:

1. **Stereotype detection**, a binary classification task whose goal is to asses whether a given text contains a stereotype against immigrants or not.
2. **Implicitness identification**, a novel hierarchical binary classification task based on determining whether the stereotype inside a text is implicit or not. An example of a implicit stereotype can be found on the following translated sentence: *Yesterday I was at the tax office, all Spaniards. In the afternoon, I went to the health center, half Spaniards.*[1]

#### 3.1.2 Evaluation metrics

Depending on the considered task as well as the evaluation context, the DETEST-Dis tasks are evaluated according to the metrics of the Table 1.

| Task | Hard Evaluation | Soft Evaluation |
|---|---|---|
| Stereotype | F1-Score | Cross Entropy |
| Implicitness | ICM, ICM Norm | ICM Soft, ICM Soft Norm |

Table 1: Official metrics depending on the given task along with the evaluation context in DETEST-Dis.

#### 3.1.3 Datasets

The DETEST-Dis dataset is composed of two datasets: sentences from online comments (DETEST corpus) and comments on news extracted from Twitter (Schmeisser-Nieto et al., 2024a). In addition, each sample provides different levels of context depending on the sample source.

On the other hand, the corpora has been annotated by two students in linguistics alongside one researcher. Furthermore, the dataset not only includes the aggregated labels in their gold (majority voting) and silver format, but also the non-aggregated format.

[1] Original: Ayer estuve en hacienda tributando, todos españoles, por la tarde fui al centro de salud, españoles, la mitad.

The following table provides a summary of the data distribution of the gold labels of each task. Note that for the second task the number of samples is reduced since it is a hierarchical task.

| TASK | YES | NO | TOTAL |
|---|---|---|---|
| Stereotype | 2605 | 7301 | 9906 |
| Implicit | 1326 | 1296 | 2622 |

Table 2: Summary of sample distributions for each one of the sub-tasks.

## 3.2 Identification of sexist stereotypes on memes

EXIST (sEXism Identification in Social neTworks) 2024 (Plaza et al., 2024) is the fourth consecutive edition of this task (Rodríguez-Sánchez et al., 2021; Rodríguez-Sánchez et al., 2022; Plaza et al., 2023), aimed at sexism identification in social networks. Similar to its previous edition, EXIST 2024 is built upon the LeWiDi learning framework as well as a multilingüal perspective with content from both Spanish and English.

Nevertheless, while past editions were centered around sexism identification on posts extracted from Twitter, the 2024 edition introduces novel tasks related to sexism identification on memes.

### 3.2.1 Tasks

Although EXIST 2024 tasks are divided into text only tasks (1-3) and meme only tasks (4-6), we can group them according to the following taxonomy:

1. **Sexism identification** (1 & 4), focused on detecting whether a given content contains sexism or not.
2. **Source intention** (2 & 5), a binary hierarchical task aimed at detecting the intention behind the sexist content.
3. **Sexism categorization** (3 & 6), a multi-label hierarchical task aimed at categorizing the sexist attack into five categories, including the **Stereotyping and Dominance** category, which expresses false ideas about how women are more suitable for certain roles and their lack of ability for others, or claiming male superiority.

As we mentioned earlier, one of the objectives of this work is to study the impact of the LeWiDi paradigm on the detection of sexist stereotypes in memes. Therefore, for EXIST we mainly address the sub-task 6 on identifying the aforementioned category.

### 3.2.2 Evaluation metrics

Depending on the considered task as well as the evaluation context, the EXIST tasks are officially evaluated according to the metrics of Table 3.

| Task | Hard Evaluation | Soft Evaluation |
|---|---|---|
| 4 | **ICM**, **ICM Norm**, F1-Score | **ICM Soft**, **ICM Soft Norm**, Cross Entropy |
| 5 | **ICM**, **ICM Norm**, F1-Score | **ICM Soft**, **ICM Soft Norm**, Cross Entropy |
| 6 | **ICM**, **ICM Norm** | **ICM Soft**, **ICM Soft Norm** |

Table 3: Official metrics depending on the given task along with the evaluation context on EXIST. Metrics on bold are the ones considered for ranking.

### 3.2.3 Datasets

The EXIST dataset for memes is composed of a total of 4044 samples, 2010 of them belonging to English while the rest are part of the Spanish.

The annotators have been selected and filtered through the crowd-sourcing Prolific [2] platform, which yields different perspectives on the annotation process. As a consequence, each sample has been annotated by 6 people, although each annotator has annotated an average of 27 memes.

In addition, samples are provided with both aggregated forms (gold and silver labels) as well as non-aggregated. However, since each sample is annotated by an even number of people, there are some cases which end up being a draw. Therefore, no gold label is provided for these memes.

# *4 System Proposals*

## 4.1 DETEST-Dis

Our system proposal is based on the RoBERTa (Liu et al., 2019) Transformer architecture as the central component. More specifically, we have used the one from the

[2] Available at https://www.prolific.com/. Visited on 29/10/2024

MarIA project (Gutiérrez-Fandiño et al., 2022) [3], which was pretrained in Spanish.

In order to explore how the LeWiDi paradigm impacts the model performance, we consider the following proposals, depicted on Figure 1, which modify the classification head (a Multi-Layer Perceptron) along with the training procedure:

1. **Hard label approach**. This methodology is based on the classic approach, by which we train our model using the aforementioned gold label. Therefore, the output layer of our classification head is composed by a single neuron and the model is trained by minimizing Binary Cross Entropy.

2. **Soft label approach**. This approach is based on training the model on the silver labels. Since the original distribution is provided with a softmax, we will train our model by adding an extra neuron to the output layer and minimizing Cross Entropy loss.

3. **Multi-task approach**. Following the perspectivist approach, we will try to predict each annotator opinion individually. In order to do so, we have based this approach on the multi-task architecture proposed by Mostafazadeh Davani, Díaz, and Prabhakaran (2022), where we had three different classification heads, one for each annotator. Since this architecture is intended to predict non-aggregated labels, we will aggregate them using the same procedures as the one used by the organizers -i.e., majority voting and softmax normalization.

Furthermore, we adopted a layer-wise decreasing learning rate fine-tuning strategy (Sun et al., 2019). This strategy ensures that the Transformer layers closer to the output, the most important for classification tasks, receive larger updates, while the layers closer to the input, which capture more abstract features, are updated less aggressively (Sun et al., 2019). This approach is reflected on Equation 1, where $\alpha_l$ is the learning rate for the Transformer layer $l$, whereas $\xi$ indicates the rate of decrease.

$$\alpha_l = \xi \cdot \alpha_{l+1} \quad (1)$$

[3] PlanTL-GOB-ES/roberta-base-bne

On the other hand, since the dataset is unbalanced for the first sub-task, we have applied back-translation (Siino, Lomonaco, and Rosso, 2024) from Spanish to English and vice-versa to all the stereotyped samples, which represent the minority class, using the NLP Augmentation toolkit (Ma, 2019) and the corresponding Opus-MT models[4,5] (Tiedemann et al., 2023; Tiedemann and Thottingal, 2020). The proposed data augmentation technique duplicates the number of stereotyped samples, resulting in a more balanced dataset for the first sub-task and a larger number of samples for the second sub-task.

Finally, we have decided to incorporate the context during training to provide more information for the given task. If the provided sample is a tweet, we introduce the first tweet of the thread as context, whereas if it is a comment from the forum, the previous sentence is used as context. The context has been appended by using the special separation token [SEP].

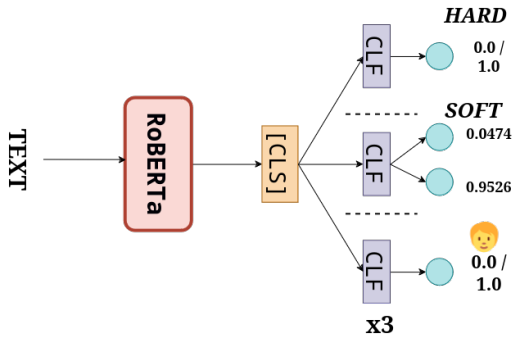


Figure 1: DETEST-Dis architecture proposals, where CLF represents a classifier head.

## 4.2 EXIST

Since memes are provided with both textual and visual information, we decided to make a comparison of each modality individually as well as both of them combined. Therefore, we have three different architectures:

1. **Text approach**, which uses a RoBERTa Transformer as the central component, feeding its representation of the special token of classification [CLS]

[4] Helsinki-NLP/opus-mt-es-en
[5] Helsinki-NLP/opus-mt-en-es

to a classification head. Depending on the source language of the meme, we will either use the one from Meta for English [6] and the one from the MarIa project for Spanish.

2. **Vision approach**. Since Transformers have achieved state-of-the-art results in many vision related tasks, we will be using the original Vision Transformer (Dosovitskiy et al., 2021) encoder. [7] Akin to text model, it also includes a classification token [CLS] whose representation will be fed to a classification head.

3. **Early fusion approach**. In order to combine both modalities in an effective way, we have decided to present an early fusion architecture which is based on the concatenation of the [CLS] tokens produced by each one the encoders, as depicted on Figure 2. Although some multimodal multilingual generative models can be found in literature (Yue et al., 2024; Geigle et al., 2024), this proposal is motivated by the lack of discriminative pretrained multimodal models in Spanish as well as the capability of choosing a specific text encoder depending on the origin language of the meme.

Regarding the LeWiDi paradigm, every architecture has been trained using both gold and silver labels, which provides a fair comparison. We have decided to discard the multi-task approach for this task, since the number of annotations per annotator is not enough for training a classification head, which would lead to poor performance.

Finally, in order to enhance the text encoder performance, we have decided to explore the following techniques:

1. **Text preprocessing**. In order to obtain a cleaner text, we have decided to apply a preprocessing step that consists of lower-casing the entire text and removal of URL's, usernames, emojis and the hashtag symbol.

2. **Randomly masking identity terms.** Identity terms are sensitive terms which lead to bias on misogyny detection tasks (Nozza, Volpetti, and Fersini, 2019). In order to address this, we have manually collected a list of these terms and randomly replaced them with the mask token [MASK] during training. The identity term list for each language can be found in Appendix B.1.

3. **Generating image captions**. As previously mentioned, adding extra features such as image captions can enhance the performance in this type of tasks. As a consequence, we will use image captions generated by LLaVa (Liu et al., 2023). The prompt used for generating captions can be found in Appendix B.2.

4. **Augmentation of the training dataset with tweets**. Since both the EXIST meme and tweet tasks belong to a common taxonomy and share similar characteristics—such as their short length and casual tone—we have decided to include the EXIST tweets in the meme training corpus.

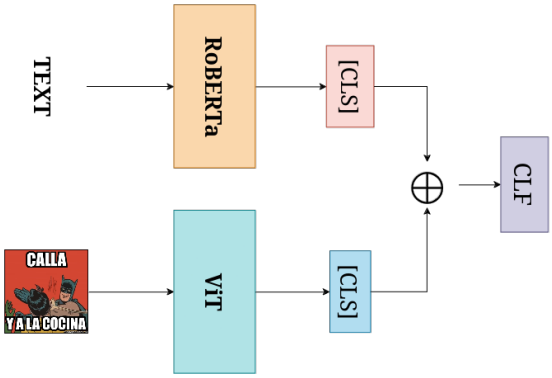


Figure 2: Multimodal EXIST architecture. $\bigoplus$ denotes concatenation operation, whereas CLF is a classifier head.

## 5 *Experimental setup*

All experiments were conducted on an RTX 2080 GPU (8 GB VRAM) with a fixed batch size of 16, employing gradient accumulation when memory constraints arose. Our training protocol utilized an AdamW optimizer (Loshchilov and Hutter, 2019) with a linear learning rate schedule and 10% warmup period over 10 epochs, using a default learning rate of 5e-5 unless otherwise specified. To mitigate potential overfitting during training, early stopping was implemented with a patience of 3 epochs. Furthermore, to establish

[6] FacebookAI/roberta-base
[7] google/vit-base-patch16-224

statistical robustness of our findings, we performed stratified 10-Fold Cross Validation on the training set.

On the other hand, we used the first sub-task of each shared task to determine the best hyper-parameters and techniques. These were then applied to the remaining sub-tasks within the same shared task, as they address similar topics and share common characteristics.

## *6 Results and Discussion*

### 6.1 DETEST-Dis

Before conducting any experimentation on the effects of back-translation and context in our model performance, we first performed a hyper parameter search of $\xi$ and $\alpha$ in order to adopt the most effective fine-tuning. The full results of this fine-tuning approach can be found in Appendix A.1.

Table 4 features the best hyper-parameters for each architecture as well as the results for a basic fine-tuning strategy when $\xi = 1$, which allows for a fair comparison between them. The results reflect that our selected fine-tuning strategy slightly improves the results for all architectures.

| | $\xi$ | $\alpha$ | F1 ↑ | Cross Entropy ↓ |
|---|---|---|---|---|
| Hard Label | 1.0 | 1e-5 | 0.738 ± 0.022 | 0.626 ± 0.016 |
| | 0.95 | 1e-5 | 0.741 ± 0.018 | 0.631 ± 0.012 |
| Soft label | 1.0 | 5e-5 | 0.741 ± 0.020 | 0.598 ± 0.025 |
| | 0.95 | 5e-5 | 0.749 ± 0.017 | 0.593 ± 0.017 |
| Multi-Task | 1.0 | 2e-5 | 0.734 ± 0.031 | 0.819 ± 0.041 |
| | 0.95 | 2e-5 | 0.752 ± 0.016 | 0.809 ± 0.035 |

Table 4: Comparison of best hyper-parameters for the proposed fine-tuning against a normal fine-tuning.

On the other hand, all architectures seem to show a similar performance in the F1 score, with the Multi-task approach being the most effective. Nevertheless, this latter architecture falls apart in comparison with the other two when computing Cross Entropy, with the Soft label approach showing the most promising results.

To investigate this phenomenon, we analyzed the error probability distribution predicted by the architecture against the original data, revealing in Figure 3 that most errors stem from high-confidence incorrect predictions, thus increasing Cross Entropy. In other words, most of the errors are due to the model predicting, for three annotators, the opposite of their original annotations.

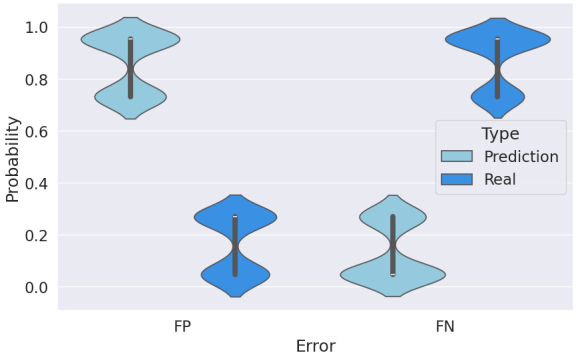


Figure 3: Probability distribution of False Positives (FP) and False Negatives (FN) in comparison with the real silver label for the Multi-task approach.

Table 5 shows the results of including context and back-translation during training, which further enhances the performance of each architecture, with the Soft label one showing the best results in both metrics.

| Architecture | F1 Score ↑ | Cross Entropy ↓ |
|---|---|---|
| Hard Label | 0.8713 ± 0.0081 | 0.6588 ± 0.0397 |
| Soft label | **0.8980 ± 0.0046** | **0.5177 ± 0.0076** |
| Multi-task | 0.8829 ± 0.0084 | 0.6369 ± 0.0289 |

Table 5: Results of using back-translation alongside context on each architecture. Best results are highlighted in bold.

Table 6 shows our results at the first task of the shared task. As it can be seen, our soft label proposal has achieved competitive results.

| Architecture | F1 Score ↑ | Cross Entropy ↓ |
|---|---|---|
| Hard label | 0.653 | 1.409 |
| Soft label | **0.691** | **0.850** |
| Multi-task | 0.685 | 1.081 |
| Gold baseline | 1.000 | 0.255 |
| Winning team | 0.720 | 0.841 |
| Organizers baseline BETO | 0.663 | 0.893 |

Table 6: Results on the test set of the first task of DETEST-Dis. Best results are highlighted in bold.

Whereas the performance on the soft evaluation is not surprising at all, the results of the F1 score are surprising since they surpass both the Multi-task proposal as well as the classical one. As pointed by on Uma et

al. (2020), this situation arises in scenarios in which the annotators are experts about the annotation subject.

On the other hand, whereas our Multi-task yields very similar results on the F1 score, its performance is the worst among all the considered teams for the Cross Entropy. In addition, our model trained with gold labels obtains the worst results overall, proving that a classical approach might not always be the most effective training.

For the implicitness detection task, since it closely resembles the racism detection one, we decided to train the models using the best hyper-parameters identified previously. As shown in Table 7, our proposals achieved the best performance among all the participants. However, it is worth mentioning that we are dealing with a very challenging task, since for the hard evaluation no participant was capable of improving the baseline obtained by BETO (Cañete et al., 2020) and established by the organizers, whereas for the soft evaluation only our proposals were capable of surpassing it.

| Architecture | ICM ↑ | ICM Norm ↑ | ICM Soft ↑ | ICM Soft Norm ↑ |
|---|---|---|---|---|
| Hard label | 0.045 | 0.516 | -0.917 | 0.401 |
| Soft label | **0.065** | **0.524** | -0.969 | 0.396 |
| Multi-Task | 0.061 | 0.522 | **-0.900** | **0.403** |
| Gold standard | 1.380 | 1.000 | 4.651 | 1.000 |
| Second team | -0.240 | 0.413 | -1.250 | 0.366 |
| Organizers baseline BETO | 0.126 | 0.546 | -1.124 | 0.379 |

Table 7: Results on the test set of the second task of DETEST-Dis. Best results are highlighted in bold.

## 6.2 EXIST

In our evaluation of the EXIST shared task, the sexism identification task on memes (task 4) allowed us to asses the impact of the various proposed text techniques. Our results indicate that only the usage of text captioning and incorporating tweets into the training dataset provided a significant improvement over the text baseline results. For a further explanation of these results, please refer to the Appendix B.3.

Given the significant performance improvement observed with image captioning, we incorporated it as a default preprocessing step for all text-based and multimodal models. In addition, since augmenting the training dataset with tweets also yielded to better results, we have also decided to include them in the text only approaches.

Table 8 presents the results of sexism categorization task and, more specifically, the **Stereotyping and Dominance** class, thus providing a fair comparison of the selected class against the macro average. As the results show, the model's performance on the stereotype identification aligns with the macro average, which suggests that the model generalizes well for the stereotype class.

| Architecture | Label | F1 - Stereotyping ↑ | F1 - Macro ↑ |
|---|---|---|---|
| Text + Image Captions | Gold | 0.2626 ± 0.2602 | 0.2974 ± 0.1786 |
| | Silver | 0.2152 ± 0.2449 | 0.2618 ± 0.1634 |
| Text + Image Captions + Tweets | Gold | 0.4073 ± 0.0691 | **0.4829 ± 0.0970** |
| | Silver | **0.4528 ± 0.0654** | 0.4570 ± 0.0332 |
| Image | Gold | 0.1366 ± 0.0707 | 0.2134 ± 0.0356 |
| | Silver | 0.0396 ± 0.0598 | 0.1535 ± 0.0206 |
| Early Fusion | Gold | 0.3871 ± 0.0861 | 0.3355 ± 0.0291 |
| | Silver | 0.1080 ± 0.1725 | 0.1814 ± 0.0904 |

Table 8: F1 scores on the sexism categorization (task 6) problem of EXIST in Spanish. Best results for each metric are highlighted in bold.

In order to evaluate the generalization capabilities of our models, we have calculated the Cross Entropy between our predictions and the original silver label of the aforementioned class. Note that we cannot compare the probability distribution of the overall performance since we are on a multi-label problem. As we can see in Table 9, the entropy decreases for every architecture by training with silver labels, which further proves that the LeWiDi paradigm results in obtaining more generalizable systems.

The results also present a comprehensive comparison of text-only, image-only, and multimodal approaches of the given task. Indeed, our results demonstrate that the text modality contains more useful information than the images.

Finally, the text-only model with data augmentation performed better than the multimodal approach. However, it is important to take into account that this data aug-

| Architecture | Label | Cross Entropy ↓ |
|---|---|---|
| Text + Image Captions | Gold | 2.4008 ± 0.1639 |
| | Silver | 1.4704 ± 0.1442 |
| Text + Image Captions + Tweets | Gold | 2.420 ± 0.1964 |
| | Silver | **1.3853 ± 0.1525** |
| Image | Gold | 2.4257 ± 0.0576 |
| | Silver | 1.67 ± 0.1191 |
| Early Fusion | Gold | 2.6143 ± 0.0947 |
| | Silver | 1.4126 ± 0.2376 |

Table 9: Cross Entropy of the stereotype class on the sexism categorization (task 6) problem of EXIST. Best result is highlighted in bold.

mentation technique was only possible due to the characteristics of the shared task. Under other circumstances were only memes are available, the Early Fusion model delivers better performance than the text modality, highlighting the importance of leveraging both image and text information when working on meme-related tasks.

## 7 Conclusions

In this work we have investigated the LeWiDi paradigm and how does it affect in two shared tasks.

On the one hand, for DETESTS-Dis, we developed three distinct architectures: one trained with gold labels using the classical approach for addressing disagreements, and two within the LeWiDi framework. The latter were trained either with silver labels or directly with the non-aggregated labels in order to predict each annotator point of view.

We also have performed a hyperparameter search for each architecture in order to apply an effective layer wise learning rate fine-tuning. Moreover, we have introduced context as well as data augmentation through back-translation in order to boost our models performance.

Our findings align with others, suggesting that systems trained using the soft loss approach produce more generalizable systems. Notably, when annotators are experts on the matter, the performance of the soft loss approach surpasses the classical one (**RQ1**).

We also conducted a comprehensive analysis of our perspectivist approach, providing insights into why our model struggles to achieve high generalizability, primarily due to the specific error patterns observed in our system.

On the other hand, for EXIST research we have developed three different architectures for processing image, text and combined modalities of memes in sexism stereotype identification. Our image-only architecture performed the worst among all the three, whereas our multimodal model achieved better results than the text-only if no data augmentation was applied.

Moreover, we have investigated different techniques to boost the performance of the text encoder in our architectures. Notably, the most effective techniques were adding an image caption generated by an external model as well as performing data augmentation, which aligns with prior findings pointed out by other researchers. In addition, we have shown that by training with silver labels we are capable of obtaining more generalizable systems for identifying sexist stereotypes in memes (**RQ2**).

In conclusion, our research has made significant strides in addressing annotation disagreements through the LeWiDi paradigm, especially in tasks as subjective as stereotype detection. We participated in two distinct shared tasks, notably dealing with the emerging challenge of stereotype detection in memes—a complex and evolving domain. By implementing the LeWiDi framework, we demonstrated that soft labels provide us with more generalizable systems capable of reflecting disagreements. In addition, we have also analyzed why the selected perspectivist approach might not be able to generalize as good as others.

## Acknowledgments

This work has been developed under the project FAKEnHATE-PdC (PDC2022-133118-I00) with founding from European NextGenerationEU/PRTR. EUR and PR carried out part of thier work also in the framework of the following projects respectively: MARTINI (Grant PCI2022-135008-2) and FairTransNLP-Stereotypes (Grant PID2021-124361OB-C31) both funded by MCIN/AEI/10.13039/501100011033 and by European Union NextGenerationEU/PRTR.

## References

Amigo, E. and A. Delgado. 2022. Evaluating Extreme Hierarchical Multi-label Classification. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*

*(Volume 1: Long Papers)*, pages 5809–5819, Dublin, Ireland, May. Association for Computational Linguistics.

Anzovino, M., E. Fersini, and P. Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *Natural Language Processing and Information Systems: 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings 23*, pages 57–64. Springer.

Ariza-Casabona, A., W. S. Schmeisser-Nieto, M. Nofre, M. Taulé, E. Amigó, B. Chulvi, and P. Rosso. 2022. Overview of DETESTS at IberLEF 2022: DETEction and classification of racial STereotypes in Spanish. *Procesamiento del Lenguaje Natural*, 69(0):217–228.

Aroyo, L. and C. Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Basile, V., C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, and M. Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, and S. M. Mohammad, editors, *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Cabitza, F., A. Campagner, and V. Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish Pre-Trained BERT Model and Evaluation Data. In *PML4DC at ICLR 2020*.

Chen, Y.-C., L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu. 2020. UNITER: UNiversal Image-TExt Representation Learning. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, page 104–120, Berlin, Heidelberg. Springer-Verlag.

Chulvi, B., M. Molpeceres, M. F. Rodrigo, A. H. Toselli, and P. Rosso. 2024. Politicization of Immigration and Language Use in Political Elites: A Study of Spanish Parliamentary Speeches. *Journal of Language and Social Psychology*, 43(2):164–194.

Dawid, A. P. and A. M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28.

Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Fersini, E., F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, and J. Sorensen. 2022. SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification. In G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, and S. Ratan, editors, *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States, July. Association for Computational Linguistics.

Fersini, E., D. Nozza, and P. Rosso. 2020. AMI @ EVALITA2020: Automatic Misogyny Identification. *EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020*.

Fersini, E., D. Nozza, P. Rosso, et al. 2018. Overview of the evalita 2018 task on automatic misogyny identification (ami). In *CEUR workshop proceedings*, volume 2263, pages 1–9. CEUR-WS.

FRA. 2023. Online content moderation – current challenges in detecting hate speech. Available at: `https://fra.europa.eu/en/publication/2023/online-content-moderation` Accessed: 2025-02-11.

Gandhi, A., P. Ahir, K. Adhvaryu, P. Shah, R. Lohiya, E. Cambria, S. Poria, and A. Hussain. 2024. Hate speech detection: A comprehensive review of recent works. *Expert Systems*, page e13562.

Geigle, G., A. Jain, R. Timofte, and G. Glavaš. 2024. mBLIP: Efficient bootstrapping of multilingual vision-LLMs. In J. Gu, T.-J. R. Fu, D. Hudson, A. Celikyilmaz, and W. Wang, editors, *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 7–25, Bangkok, Thailand, August. Association for Computational Linguistics.

Gutiérrez-Fandiño, A., J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, and M. Villegas. 2022. MarIA: Spanish Language Models. *Procesamiento del Lenguaje Natural*, 68:39–60.

Hermida, P. C. d. Q. and E. M. d. Santos. 2023. Detecting hate speech in memes: a review. *Artificial Intelligence Review*, 56(11):12833–12851, Nov.

Jahan, M. S. and M. Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232.

Kiela, D., H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624.

Li, L. H., M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang. 2019. VisualBERT: A simple and performant baseline for vision and language. *arXiv*.

Liu, H., C. Li, Y. Li, and Y. J. Lee. 2023. Improved Baselines with Visual Instruction Tuning.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Loshchilov, I. and F. Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Ma, E. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Mostafazadeh Davani, A., M. Díaz, and V. Prabhakaran. 2022. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Nockleby, J. T. 2000. Hate speech. *Encyclopedia of the American constitution*, 3(2):1277–1279.

Nozza, D., C. Volpetti, and E. Fersini. 2019. Unintended bias in misogyny detection. In *Ieee/wic/acm international conference on web intelligence*, pages 149–155.

OBERAXE. 2024. Informe Anual de Monitorización del Discurso de Odio en Redes Sociales 2023. Technical report, Ministerio de Inclusión, Seguridad Social y Migraciones.

Plaza, L., J. Carrillo-de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, and P. Rosso. 2023. Overview of EXIST 2023–learning with disagreement for sexism identification and characterization. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 316–342. Springer.

Plaza, L., J. Carrillo-de Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, and D. Spina. 2024. Overview of EXIST 2024—Learning with Disagreement for Sexism Identification and Characterization in Tweets and Memes. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 93–117. Springer.

Rawat, A., S. Kumar, and S. S. Samant. 2024. Hate speech detection in social media: Techniques, recent trends, and future challenges. *WIREs Computational Statistics*, 16(2):e1648.

Rizzi, G., F. Gasparini, A. Saibene, P. Rosso, and E. Fersini. 2023. Recognizing misogynous memes: Biased models and tricky archetypes. *Information Processing & Management*, 60(5):103474.

Rodríguez-Sánchez, F., J. Carrillo-de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, and T. Donoso. 2021. Overview of EXIST 2021: sEXism Identification in Social neTworks. *Procesamiento del Lenguaje Natural*, 67:195–207.

Rodríguez-Sánchez, F., J. Carrillo-de Albornoz, L. Plaza, A. Mendieta-Aragón, G. Marco-Remón, M. Makeienko, M. Plaza, J. Gonzalo, D. Spina, and P. Rosso. 2022. Overview of EXIST 2022: sEXism Identification in Social neTworks. *Procesamiento del Lenguaje Natural*, 69:229–240.

Schmeisser-Nieto, W., M. Nofre, and M. Taulé. 2022. Criteria for the annotation of implicit stereotypes. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 753–762, Marseille, France, June. European Language Resources Association.

Schmeisser-Nieto, W. S., A. T. Cignarella, T. Bourgeade, S. Frenda, A. Ariza-Casabona, M. Laurent, P. G. Cicirelli, A. Marra, G. Corbelli, F. Benamara, C. Bosco, V. Moriceau, M. Paciello, V. Patti, M. Taulé, and F. D'Errico. 2024a. Stereohoax: a multilingual corpus of racial hoaxes and social media reactions annotated for stereotypes. *Language Resources and Evaluation*, Dec.

Schmeisser-Nieto, W. S., P. Pastells, S. Frenda, A. Ariza-Casabona, M. Farrús, P. Rosso, and M. Taulé. 2024b. Overview of DETESTS-Dis at IberLEF 2024: DETEction and classification of racial STereotypes in Spanish-Learning with Disagreement. *Procesamiento del Lenguaje Natural*, 73:323–333.

Siino, M., F. Lomonaco, and P. Rosso. 2024. Backtranslate what you are saying and I will tell who you are. *Expert Systems*, page e13568.

Subramanian, M., V. Easwaramoorthy Sathiskumar, G. Deepalakshmi, J. Cho, and G. Manikandan. 2023. A survey on hate speech detection and sentiment analysis using machine learning and deep learning models. *Alexandria Engineering Journal*, 80:110–121.

Sun, C., X. Qiu, Y. Xu, and X. Huang. 2019. How to fine-tune BERT for text classification? In *Chinese computational linguistics: 18th China national conference, CCL 2019, Kunming, China, October 18–20, 2019, proceedings 18*, pages 194–206. Springer.

Sánchez-Junquera, J., B. Chulvi, P. Rosso, and S. P. Ponzetto. 2021. How do you speak about immigrants? taxonomy and stereoimmigrants dataset for identifying stereotypes about immigrants. *Applied Sciences*, 11(8).

Tiedemann, J., M. Aulamo, D. Bakshandaeva, M. Boggia, S.-A. Grönroos, T. Nieminen, A. Raganato Y. Scherrer, R. Vazquez, and S. Virpioja. 2023. Democratizing neural machine translation with OPUS-MT. *Language Resources and Evaluation*, (58):713–755.

Tiedemann, J. and S. Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Uma, A., T. Fornaciari, D. Hovy, S. Paun, B. Plank, and M. Poesio. 2020. A case for soft loss functions. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 173–177.

Uma, A. N., T. Fornaciari, D. Hovy, S. Paun, B. Plank, and M. Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Waseem, Z. 2016. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In D. Bamman, A. S. Doğruöz, J. Eisenstein, D. Hovy, D. Jurgens, B. O'Connor, A. Oh,

O. Tsur, and S. Volkova, editors, *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas, November. Association for Computational Linguistics.

Yue, X., Y. Song, A. Asai, S. Kim, J. de Dieu Nyandwi, S. Khanuja, A. Kantharuban, L. Sutawika, S. Ramamoorthy, and G. Neubig. 2024. Pangea: A fully open multilingual multimodal llm for 39 languages. *CoRR*.

Zhou, Z., H. Zhao, J. Dong, N. Ding, X. Liu, and K. Zhang. 2022. DD-TIG at SemEval-2022 Task 5: Investigating the Relationships Between Multimodal and Unimodal Information in Misogynous Memes Detection and Classification. In G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, and S. Ratan, editors, *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 563–570, Seattle, United States, July. Association for Computational Linguistics.

Zhu, R. 2020. Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution. *arXiv*.

# *A DETEST-Dis*

## A.1 Hyper-parameter search results

Tables 10, 11, 12 contain the results of the hyper-parameter search procedure carried out for the fine-tuning.

| $\xi$ | $\eta$ | F1_Stereotype | Cross Entropy |
|---|---|---|---|
| 1.0 | 1e-4 | 0.7145 ± 0.0195 | 0.7191 ± 0.0435 |
| 0.97 | 1e-4 | 0.7181 ± 0.0255 | 0.7117 ± 0.0476 |
| 0.95 | 1e-4 | 0.7115 ± 0.0317 | 0.6639 ± 0.0361 |
| 0.90 | 1e-4 | 0.7145 ± 0.0449 | 0.6913 ± 0.0277 |
| 1.0 | 5e-5 | 0.7023 ± 0.0270 | 0.6810 ± 0.0395 |
| 0.97 | 5e-5 | 0.7210 ± 0.0303 | 0.6754 ± 0.0531 |
| 0.95 | 5e-5 | 0.7106 ± 0.0366 | 0.6804 ± 0.0271 |
| 0.90 | 5e-5 | 0.7316 ± 0.0151 | 0.6484 ± 0.0117 |
| 1.0 | 2e-5 | 0.7322 ± 0.0334 | 0.6432 ± 0.0296 |
| 0.97 | 2e-5 | 0.7256 ± 0.0279 | 0.6436 ± 0.0281 |
| 0.95 | 2e-5 | 0.7350 ± 0.0170 | 0.6355 ± 0.0332 |
| 0.90 | 2e-5 | 0.7398 ± 0.0127 | 0.6343 ± 0.0302 |
| 1.0 | 1e-5 | 0.7380 ± 0.0222 | **0.6260 ± 0.0157** |
| 0.97 | 1e-5 | 0.7373 ± 0.0210 | 0.6331 ± 0.0175 |
| 0.95 | 1e-5 | **0.7410 ± 0.0184** | 0.6308 ± 0.0125 |
| 0.90 | 1e-5 | 0.7341 ± 0.0214 | 0.6340 ± 0.0124 |

Table 10: Best hyper-parameters $\xi$ and $\eta$ for the selected fine-tuning approach for the model trained with gold labels in the racist stereotype identification task. Best results are highlighted in bold.

| $\xi$ | $\eta$ | F1_Stereotype | Cross Entropy |
|---|---|---|---|
| 1.0 | 1e-4 | 0.5817 ± 0.3054 | 0.9401 ± 0.2293 |
| 0.97 | 1e-4 | 0.7361 ± 0.0185 | 0.8211 ± 0.0447 |
| 0.95 | 1e-4 | 0.7254 ± 0.0032 | 0.8274 ± 0.0499 |
| 0.90 | 1e-4 | 0.7375 ± 0.0290 | 0.8102 ± 0.0421 |
| 1.0 | 5e-5 | 0.7313 ± 0.0275 | 0.8193 ± 0.0300 |
| 0.97 | 5e-5 | 0.7443 ± 0.0203 | 0.7801 ± 0.0503 |
| 0.95 | 5e-5 | 0.7404 ± 0.0091 | 0.8093 ± 0.0531 |
| 0.90 | 5e-5 | 0.7504 ± 0.0179 | **0.7881 ± 0.0722** |
| 1.0 | 2e-5 | 0.7342 ± 0.0308 | 0.8191 ± 0.0411 |
| 0.97 | 2e-5 | 0.7498 ± 0.0196 | 0.8064 ± 0.0336 |
| 0.95 | 2e-5 | 0.7490 ± 0.0200 | 0.8070 ± 0.0249 |
| 0.90 | 2e-5 | **0.7519 ± 0.0163** | 0.8094 ± 0.0375 |
| 1.0 | 1e-5 | 0.7433 ± 0.0140 | 0.8154 ± 0.0282 |
| 0.97 | 1e-5 | 0.7434 ± 0.0202 | 0.8175 ± 0.0297 |
| 0.95 | 1e-5 | 0.7478 ± 0.0220 | 0.8175 ± 0.0399 |
| 0.90 | 1e-5 | 0.7424 ± 0.0234 | 0.8203 ± 0.0350 |

Table 11: Best hyper-parameters $\xi$ and $\eta$ for the selected fine-tuning approach for the model trained with the perspectivist approach in the racist stereotype identification task. Best results are highlighted in bold.

| $\xi$ | $\eta$ | F1_Stereotype | Cross Entropy |
|---|---|---|---|
| 1.00 | 1e-4 | 0.7311 ± 0.0315 | 0.61067 ± 0.0192 |
| 0.97 | 1e-4 | 0.7454 ± 0.0169 | 0.6061 ± 0.0214 |
| 0.95 | 1e-4 | 0.7430 ± 0.0249 | 0.6040 ± 0.0147 |
| 0.90 | 1e-4 | 0.7464 ± 0.0227 | 0.5994 ± 0.0207 |
| 1.00 | 5e-5 | 0.7413 ± 0.0203 | 0.5979 ± 0.0250 |
| 0.97 | 5e-5 | 0.7463 ± 0.0279 | 0.5931 ± 0.0191 |
| 0.95 | 5e-5 | 0.7467 ± 0.0284 | 0.6014 ± 0.0205 |
| 0.90 | 5e-5 | **0.7488 ± 0.0175** | **0.5926 ± 0.0169** |
| 1.00 | 2e-5 | 0.7363 ± 0.0287 | 0.5969 ± 0.0201 |
| 0.97 | 2e-5 | 0.7358 ± 0.0331 | 0.5995 ± 0.0185 |
| 0.95 | 2e-5 | 0.7323 ± 0.0345 | 0.5990 ± 0.0182 |
| 0.90 | 2e-5 | 0.7322 ± 0.0328 | 0.6018 ± 0.0179 |
| 1.0 | 1e-5 | 0.7276 ± 0.0224 | 0.6068 ± 0.0159 |
| 0.97 | 1e-5 | 0.7279 ± 0.0172 | 0.6052 ± 0.0137 |
| 0.95 | 1e-5 | 0.7282 ± 0.0144 | 0.6028 ± 0.0125 |
| 0.90 | 1e-5 | 0.7331 ± 0.0301 | 0.6036 ± 0.0153 |

Table 12: Best hyper-parameters $\xi$ and $\eta$ for the selected fine-tuning approach for the model trained with gold labels in the racist stereotype identification task. Best results are highlighted in bold.

# *B EXIST*

## B.1 Identity term list

### B.1.1 Spanish

| | | | | | |
|---|---|---|---|---|---|
| mujer | escote | hombre | hombres | mujeres | feminista |
| novia | misogino | misoginia | coche | patriarcado | sexismo |
| senora | sexual | inclusivo | machista | cocina | |
| nalgada | nalga | culo | acoso | minoria | |
| feminismo | marzo | papa | fregar | violacion | |
| varon | engañar | matrimonio | lenguaje | dieta | |
| senorita | hermana | mamá | gluteo | feminazi | |

### B.1.2 English

| | | | |
|---|---|---|---|
| women | woman | feminists | feminism |
| feminist | blonde | female | stepsister |
| girl | hot | kitchen | brother |
| stepmom | man | ass | patriarchy |
| suck | stepbrother | digger | trophy |
| lesbian | chick | skrit | mum |
| pig | cow | husband | girlfriend |
| misandry | blonde | boob | boy |

## B.2 LlaVa prompt

USER: <image>

Describe the content of the meme, but ignore the text caption of the meme. Provide a clear, concise and short answer.

ASSISTANT:

## B.3 Results of the text modality on the sexism identification task

As it can be seen in the results from Tables 13 and 14, neither the preprocessing nor identity term masking offer a significant improvement over the baselines results for the text modality.

Regarding text preprocessing, the modest improvements suggest that the text was less noisy than initially anticipated. Conversely, the identity term masking technique's limited performance indicates methodological constraints, potentially arising from mistakes in our manually curated term list or algorithmic implementation.

| Architecture | Language | Label | ICM ↑ | ICM Norm ↑ | F1 - Sexist ↑ |
|---|---|---|---|---|---|
| Text | ES | Gold | 0.0077 ± 0.0851 | 0.4960 ± 0.0446 | 0.7384 ± 0.0276 |
| | | Silver | -0.1676 ± 0.1808 | 0.4120 ± 0.0949 | 0.6702 ± 0.0595 |
| | EN | Gold | 0.1133 ± 0.1040 | 0.5574 ± 0.0527 | 0.7430 ± 0.0468 |
| | | Silver | -0.0339 ± 0.2765 | 0.4828 ± 0.1400 | 0.6738 ± 0.0231 |
| Text + Preprocessing | ES | Gold | -0.0397 ± 0.0568 | 0.4792 ± 0.0298 | 0.7406 ± 0.0585 |
| | | Silver | 0.0119 ± 0.1536 | 0.5062 ± 0.0806 | 0.6999 ± 0.0271 |
| | EN | Gold | 0.1336 ± 0.0933 | 0.5676 ± 0.0472 | 0.7388 ± 0.0216 |
| | | Silver | 0.0072 ± 0.2061 | 0.5036 ± 0.1044 | 0.6829 ± 0.0150 |
| Text + Word Masking | ES | Gold | -0.1247 ± 0.1731 | 0.4346 ± 0.0910 | 0.7326 ± 0.0303 |
| | | Silver | -0.0371 ± 0.0752 | 0.4806 ± 0.0395 | 0.6748 ± 0.0362 |
| | EN | Gold | -0.0637 ± 0.2616 | 0.4678 ± 0.1324 | 0.7051 ± 0.0800 |
| | | Silver | -0.1564 ± 0.3112 | 0.4209 ± 0.1575 | 0.6709 ± 0.0295 |
| Text + Context | ES | Gold | 0.0525 ± 0.0423 | 0.5275 ± 0.0222 | 0.7706 ± 0.0344 |
| | | Silver | -0.1463 ± 0.0991 | 0.4232 ± 0.0520 | 0.6875 ± 0.0239 |
| | EN | Gold | 0.1891 ± 0.0978 | 0.5957 ± 0.0495 | **0.7609 ± 0.0445** |
| | | Silver | 0.0668 ± 0.0440 | 0.5338 ± 0.0223 | 0.6817 ± 0.0156 |
| Text + Context + Tweets | ES | Gold | 0.2705 ± 0.0852 | 0.6358 ± 0.0428 | **0.7953 ± 0.0366** |
| | | Silver | **0.3384 ± 0.0490** | **0.6699 ± 0.0246** | 0.7334 ± 0.0185 |
| | EN | Gold | 0.1201 ± 0.3941 | 0.5604 ± 0.1980 | 0.5906 ± 0.4104 |
| | | Silver | **0.2192 ± 0.2145** | **0.6101 ± 0.1078** | 0.6487 ± 0.1164 |

Table 13: Results on the hard evaluation in the sexism identification task with only the text and the proposed techniques. Best results for each language are highlighted in bold.

| Architecture | Language | Label | ICM Soft ↑ | ICM Soft Norm ↑ | Cross Entropy ↓ |
|---|---|---|---|---|---|
| Text | ES | Gold | -0.4525 ± 0.1161 | 0.4270 ± 0.0187 | 0.9567 ± 0.0401 |
| | | Silver | -0.8858 ± 0.1098 | 0.3609 ± 0.0172 | 0.9520 ± 0.0078 |
| | EN | Gold | -0.1890 ± 0.1283 | 0.4690 ± 0.0211 | 0.9496 ± 0.0595 |
| | | Silver | -0.6526 ± 0.3085 | 0.3975 ± 0.0484 | 0.9547 ± 0.0242 |
| Text + Preprocessing | ES | Gold | -0.5563 ± 0.1531 | 0.4102 ± 0.0247 | 0.9389 ± 0.0067 |
| | | Silver | -0.6488 ± 0.1984 | 0.3981 ± 0.0312 | 0.9414 ± 0.0084 |
| | EN | Gold | -0.1325 ± 0.0990 | 0.4783 ± 0.0162 | 0.9677 ± 0.0627 |
| | | Silver | -0.6527 ± 0.1184 | 0.3975 ± 0.0186 | 0.9500 ± 0.0138 |
| Text + Word Masking | ES | Gold | -0.6645 ± 0.1736 | 0.3927 ± 0.0281 | 0.9453 ± 0.0150 |
| | | Silver | -0.6965 ± 0.1232 | 0.3907 ± 0.0193 | 0.9457 ± 0.0084 |
| | EN | Gold | -0.4129 ± 0.2436 | 0.4322 ± 0.0400 | 0.9528 ± 0.0121 |
| | | Silver | -0.8440 ± 0.2585 | 0.3674 ± 0.0406 | 0.9726 ± 0.0160 |
| Text + Context | ES | Gold | -0.3040 ± 0.0867 | 0.4509 ± 0.0141 | 0.9698 ± 0.0457 |
| | | Silver | -0.8296 ± 0.0766 | 0.3698 ± 0.0121 | 0.9491 ± 0.0090 |
| | EN | Gold | 0.0160 ± 0.1382 | 0.5026 ± 0.0227 | 0.9465 ± 0.0477 |
| | | Silver | -0.5813 ± 0.1280 | 0.4087 ± 0.0201 | 0.9392 ± 0.0134 |
| Text + Context + Tweets | ES | Gold | **0.3255 ± 0.1378** | **0.5541 ± 0.0229** | 0.8866 ± 0.0355 |
| | | Silver | 0.1729 ± 0.0984 | 0.5273 ± 0.0155 | **0.8574 ± 0.0146** |
| | EN | Gold | **0.0353 ± 0.6651** | **0.5060 ± 0.1102** | 0.9441 ± 0.0328 |
| | | Silver | -0.1252 ± 0.2768 | 0.4801 ± 0.0441 | **0.8844 ± 0.0375** |

Table 14: Results on the soft evaluation in the sexism identification task with only the text and the proposed techniques. Best results for each language are highlighted in bold.