Identification of Racial and Sexist Stereotypes in Spanish

A Learning with Disagreements Approach

In: Procesamiento del Lenguaje Natural (SEPLN), num. 74 (accepted)

Elias Urios Alacreu [1] Paolo Rosso [1,2]

[1] PRHLT Research Center, Universitat Politècnica de València
[2] ValgrAI Valencian Graduate School and Research Network of Artificial Intelligence

March 13, 2025

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

PRHLT
Pattern Recognition and
Human Language Technology
Research Center

Index

1. Introduction
2. LeWiDi
3. DETESTS-Dis
4. EXIST
5. Conclusions and future work

Index

1 Introduction

2 LeWiDi

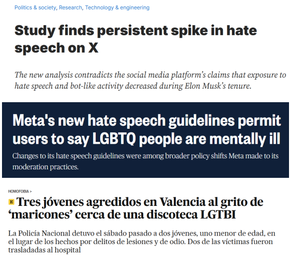
3 DETESTS-Dis

4 EXIST

5 Conclusions and future work

Introduction

- HS is an acknowledged phenomenon:
  - Social media platforms
  - Increasingly sheer volume of content
  - Change of policies
- Targets:
  - LGTBQ+
  - Black community
  - **Women**
  - **Immigrants**
- Hate speech incites violence and intolerance
- Transformers for addressing HS



Politics & society, Research, Technology & engineering

Study finds persistent spike in hate speech on X

The new analysis contradicts the social media platform's claims that exposure to hate speech and bot-like activity decreased during Elon Musk's tenure.

Meta's new hate speech guidelines permit users to say LGBTQ people are mentally ill

Changes to its hate speech guidelines were among broader policy shifts Meta made to its moderation practices.

HOMOFOBIA >

Tres jóvenes agredidos en Valencia al grito de 'maricones' cerca de una discoteca LGTBI

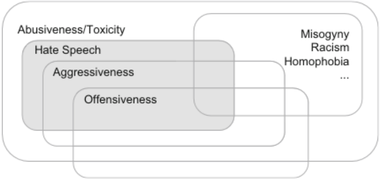
La Policía Nacional detuvo el sábado pasado a dos jóvenes, uno menor de edad, en el lugar de los hechos por delitos de lesiones y de odio. Dos de las víctimas fueron trasladadas al hospital

Observatorio Español del Racismo y la Xenofobia (OBERAXE). *Informe Anual de Monitorización del Discurso de Odio en Redes Sociales 2023*. Tech. rep. Ministerio de Inclusión, Seguridad Social y Migraciones, 2024

Challenges of addressing HS

- Lack of a universal definition
- Intersection with various fields/areas
- Lack of resources outside of English
- Forms of expression:
  - Aggressive: threats, violence
  - Non-aggressive: humor, irony, **stereotypes**
- Type of content:
  - Text (mostly addressed)
  - Multimodal (images, videos, **memes**)
- **Subjectivity**

> *[...] the lack of definitions in scholarship translates to uncertain definitions in law and social science research, and even more uncertain application of principles in on-line spaces*



Andrew Sellars. "Defining hate speech". In: *Berkman Klein Center Research Publication* 16-48.2016-20 (2016), pp. 16–48, Fabio Poletto et al. "Resources and benchmark corpora for hate speech detection: a systematic review". In: *Language Resources and Evaluation* 55.2

Deep learning and HS

DL limitations in HS

It is clear that automated methods, especially those based on DL, are necessary to address HS. However, the usage of black box methods involves ethical concerns.

Deep learning and HS

DL limitations in HS

It is clear that automated methods, especially those based on DL, are necessary to address HS. However, the usage of black box methods involves ethical concerns.

Addressing the limitations

What if we trained our models to see beyond black and white? What if we had more debiased datasets? What if we paid attention to all opinions?

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?
- Shared tasks:
    - **DETEST-Dis**: *DETEction and classification of racial STereotypes in Spanish - Learning with Disagreement*
    - **EXIST**: *sEXism Identification in Social neTworks*

Index

1 Introduction

2 LeWiDi

3 DETESTS-Dis

4 EXIST

5 Conclusions and future work

LeWiDi: Why?

- Supervised tasks require annotated data
- Data annotation is a time-consuming and expensive process
- Assumption: existence of a single, objective truth
- Reality: disagreements often arise
- Reasons
  - Mistakes/slips from the annotators
  - Poor annotation schemes
  - **Subjectivity**
  - ...
- Why would we ignore the minority over the majority?

Uma, Alexandra N. and Fornaciari, Tommaso and Hovy, Dirk and Paun, Silviu and Plank, Barbara and Poesio, Massimo. "Learning from Disagreement: A Survey". In: *J. Artif. Int. Res.* (2022)

LeWiDi: Beyond black and white

- Aggregate annotations into a gold truth (**hard label**)
  - Majority voting (traditional approach)
  - Probabilistic methods: MACE
- Ignore "difficult" labels by using disagreement
- Aggregate annotations into a probability distribution (**soft label**)
  - Probability or softmax
  - Soft loss function (**CE**, KL or MSE)
- Combine information from **hard** and **soft** labels
- Perspectivist: work directly with non-aggregated annotations

Uma, Alexandra N. and Fornaciari, Tommaso and Hovy, Dirk and Paun, Silviu and Plank, Barbara and Poesio, Massimo. "Learning from Disagreement: A Survey". In: *J. Artif. Int. Res.* (2022), Simona Frenda et al. "Perspectivist approaches to natural language processing: a survey". In: *Language Resources and Evaluation* (Aug. 2024). ISSN: 1574-0218. DOI: 10.1007/s10579-024-09766-4. URL: https://doi.org/10.1007/s10579-024-09766-4.

LeWiDi: Evaluation

- Hard evaluation: Traditional evaluation using hard labels
  - Common metrics: F1-Score, Accuracy, Information Contrast Metric (ICM)
  - Traditional approach usually works the best
  - Soft loss can be better under certain conditions
- Soft evaluation: How well does the model generalize
  - Common metrics: CE, JSD, KL, ICM Soft
  - Soft loss works the best

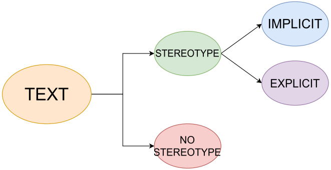
Index

1. Introduction
2. LeWiDi
3. DETESTS-Dis
4. EXIST
5. Conclusions and future work

Task descriptions

- Second edition of DETESTS
- Stereotype detection on text
- Framed within LeWiDi:
  - Aggregated annotations: majority voting and softmax
  - Non-aggregated annotations
- Two tasks:
  - **Stereotype detection**: binary classification task
  - **Stereotype implicitness detection**: novel hierarchy classification task

[a]



Task	Hard evaluation	Soft evaluation
Stereotype	F1	CE
Implicit	ICM	ICM Soft

[a] Wolfgang Schmeisseró-Nieto et al. "Overview of DETESTS-Dis at IberLEF 2024: DETEction and classification of racial STereotypes in Spanish - Learning with Disagreement". In: *Revista Procesamiento del*

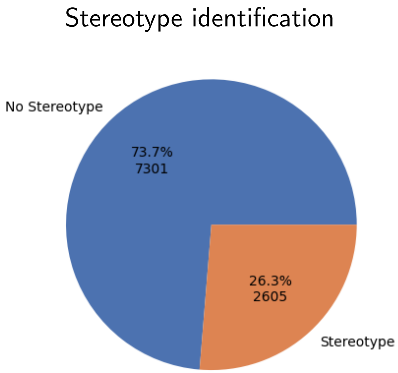
Task descriptions

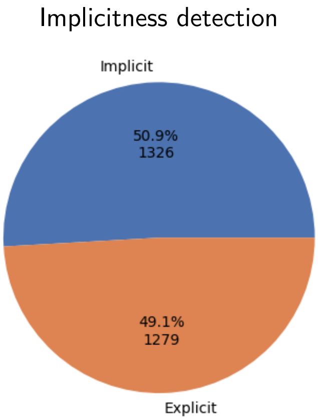
Text	Stereotype	Implicitness
The solution is to develop a critical and esceptical thinking.	X	-
Like it or not, one thing is clear: if there were no muslims in Europe, this wouldn't happen.	✓	X
Yesterday I was at the tax office, all Spaniards, in the afternoon I went to the health center, half of them Spaniards.	✓	✓

Dataset

- Comment threads from news articles
- Annotated by 3 expert annotators (2 linguistics and 1 researcher)
- Two corpora
- DETEST corpora:
  - Corpus from the first edition
  - Threads from online news forums
  - Annotations are provided on a sentence level
- StereoHOAX corpora:
  - New corpora for this edition
  - Threads from Twitter
  - Annotations are provided on a tweet level
- Different levels of context:
  - **Level 1**: Previous sentence (DETEST)
  - **Level 2**: Previous tweet/comment
  - **Level 3**: First tweet/comment
  - **Level 4**: News text

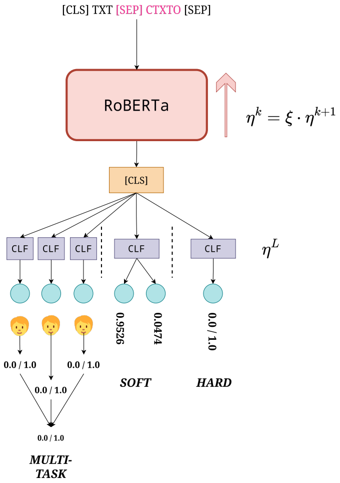
Dataset


Stereotype identification


Implicitness detection

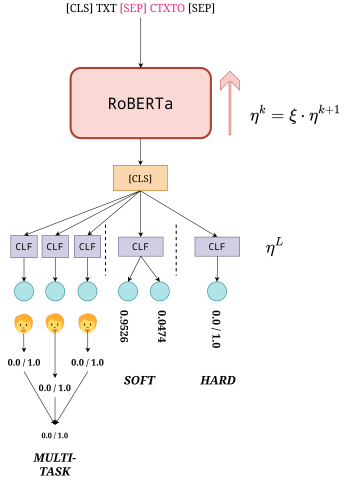
System proposals

- RoBERTa model as text encoder.
- **Hard label** approach
  - Classic approach (comparison purposes)
  - $\mathcal{L}(\hat{y}, y) = \mathrm{BCE}(\hat{y}, y)$
- **Soft label** approach
  - Train by soft label probablity distribution
  - $\mathcal{L}(\hat{y}, y) = \mathrm{CE}(\hat{y}, y)$
- **Perspectivist** approach
  - **Multi-task proposal**: three classification heads with one output neuron each
  - $\mathcal{L}(\hat{y}, y) = \sum_{a=1}^{3} \mathrm{CE}(\hat{y_a}, y_a)$
  - Aggregate outputs for predictions (majority voting and softmax)



System proposals

- Layer-Wise Learning Rate fine-tuning
  - Deeper encoder layers recieve larger updates, shallowers receive smaller ones
  - $\eta = \{1e-5, 2e-5, 5e-5, 1e-4\}$
  - $\xi = \{1'0, 0'97, 0'95, 0'90\}$
- Context inclusion:
  - Append context via [SEP] token
  - **DETEST** samples → Previous sentence
  - **StereoHOAX** samples → First tweet
- Back-translation on minority class (ES → EN → ES)



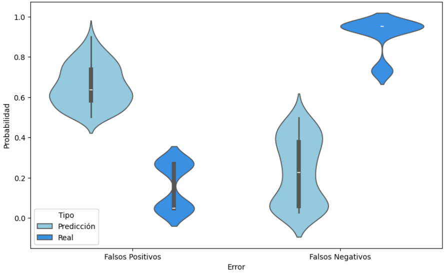
Stereotype detection: train

	Hiperparámetros		Hard Evaluation	Soft evaluation
Architecture	ξ	η	F1-Stereotype ↑	*Cross Entropy* ↓
Hard label	1.0	1e-5	0.7380 ± 0.0222	**0.6260** ± **0.0157**
Hard label	0.95	1e-5	**0.7410** ± **0.0184**	0.6308 ± 0.0125
Soft label	1.0	5e-5	0.7413 ± 0.0203	0.5979 ± 0.0250
Soft label	0.95	5e-5	**0.7488** ± **0.0175**	**0.5926** ± **0.0169** †
Multi-Task	1.0	2e-5	0.7342 ± 0.0308	0.8191 ± 0.0411
Multi-Task	0.90	2e-5	**0.7519** ± **0.0163** †	**0.8094** ± **0.0350**

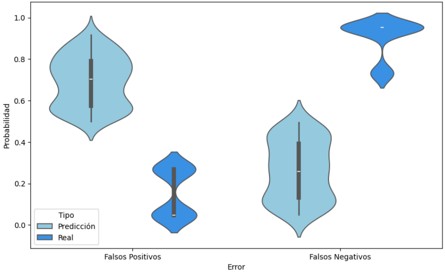
Table: Comparison of normal fine-tuning with the best parameters found on hyperparameter search.

Stereotype detection: train

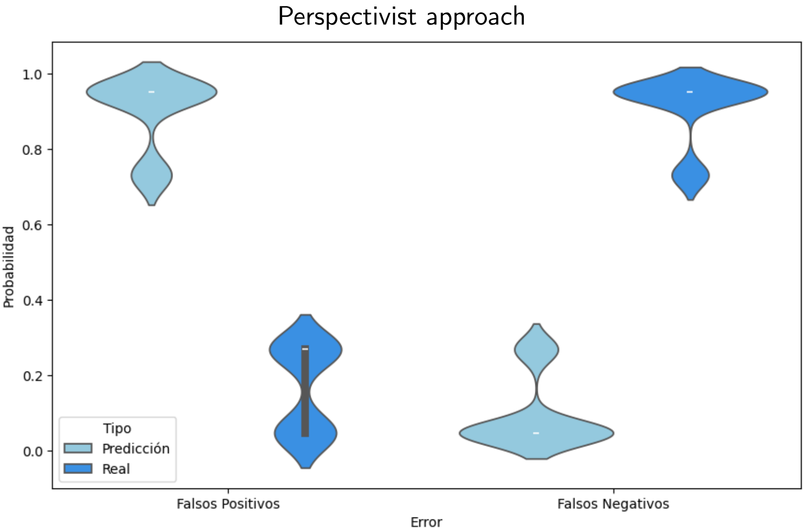
Hard label approach



Soft label approach



Stereotype detection: train



Stereotype detection: train

	Hard Evaluation	Soft evaluation
Architecture	F1-Stereotype ↑	*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: Back translation alongside context with each archicecture. † highlights the best results for each evaluation.

Stereotype detection results: test

	Hard evaluation		Soft evaluation	
Architecture	# ranking/21	F1 Score ↑	# ranking/8	*Cross Entropy* ↓
Hard label	7	0.653	8	1.409
Soft label	**4**	**0.691**	**2**	**0.850**
Multi-task	5	0.685	7	1.081
Gold baseline	0	1.000	0	0.255
Winners	1	0.720	1	0.841
Baseline BETO	6	0.663	4	0.893

Table: DETEST-Dis stereotype detection official test results. Our best results are highlighted in bold for each evaluation.

Wolfgang Schmeisseró-Nieto et al. "Overview of DETESTS-Dis at IberLEF 2024: DETEction and classification of racial STereotypes in Spanish - Learning with Disagreement". In: *Revista Procesamiento del Lenguaje Natural* 73 (2024)

Implicitness detection results: train

	Hard evaluation		Soft evaluation	
Arquitectura	ICM ↑	ICM Norm ↑	ICM Soft ↑	ICM Soft Norm ↑
Hard label	**0.0095** ± **0.0726**	**0.5049** ± **0.0533**	0.4277 ± 0.1586	0.5632 ± 0.0238
Soft label	-0.0459 ± 0.0748	0.4644 ± 0.0568	0.0870 ± 0.3862	0.5129 ± 0.0568
Multi-task	-0.0498 ± 0.1230	0.4649 ± 0.0891	**0.4981** ± **0.3719**	**0.5726** ± **0.0543**

Table: Train results for the implicitness detection task of DETEST-Dis. Our best results are highlighted in bold for each evaluation.

Implicitness detection results: train

Architecture	*Cross Entropy* ↓
Hard label	0.6639 ± 0.0208
Soft label	**0.6282 ± 0.0147**
Multi-task	0.8567 ± 0.0902

Table: Cross-entropy training results for the DETEST-Dis implicitness detection task. Our best results are highlighted in bold.

Implicitness detection results: test

	Hard evaluation			Soft evaluation		
Architecture	# ranking/14	ICM ↑	ICM Norm ↑	# ranking/6	ICM Soft ↑	ICM Soft Norm ↑
Hard label	4	0.045	0.516	2	-0.917	0.401
Soft label	**2**	**0.065**	**0.524**	3	-0.969	0.396
Multi-task	3	0.061	0.522	**1**	**-0.900**	**0.403**
Gold baseline	0	1.380	1.000	0	4.651	1.000
Baseline BETO	1	0.126	0.546	4	-1.124	0.379

Table: Test results for the implicitness detection task of DETEST-Dis. Our best results are highlighted in bold for each evaluation.

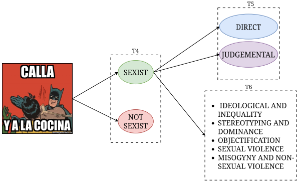
Wolfgang Schmeisseró-Nieto et al. "Overview of DETESTS-Dis at IberLEF 2024: DETEction and classification of racial STereotypes in Spanish - Learning with Disagreement". In: *Revista Procesamiento del Lenguaje Natural* 73 (2024)

Index

1. Introduction
2. LeWiDi
3. DETESTS-Dis
4. EXIST
5. Conclusions and future work

Task description

- 4th edition of EXIST
- Sexism identification and categorization in text/memes
- LeWiDi framework + **memes tasks**
- 6 tasks (*Tweets* + **Memes**) grouped in the same taxonomy:
  - Sexism Identification (1 and **4**)
  - Source Intention (2 and **5**)[a]
  - Sexism Categorization (3 and **6**)



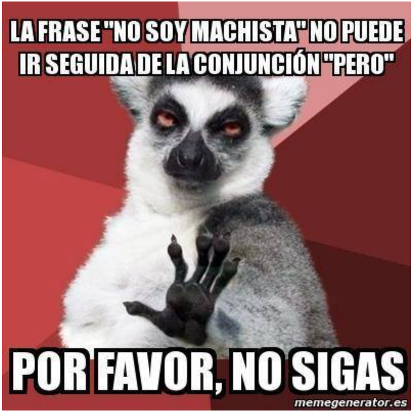
Task	Hard evaluation	Soft evaluation
Sexism Identification	**ICM, ICM Norm**, F1	**ICM Soft, ICM Soft Norm**, CE
Source Intention	**ICM, ICM Norm**, F1	**ICM Soft, ICM Soft Norm**, CE
Sexism Categorization	**ICM, ICM Norm**	**ICM Soft, ICM Soft Norm**

[a] Task 2 for tweets contains an extra category

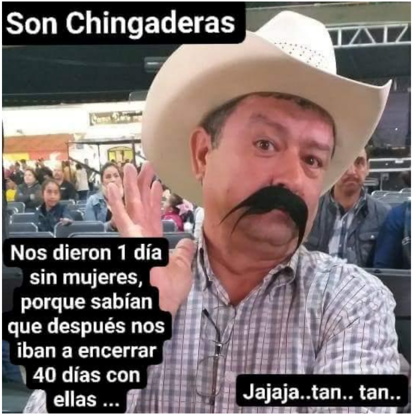
Laura Plaza et al. "Overview of EXIST 2024 - Learning with Disagreement for Sexism Identification and Characterization in Tweets and Memes (Extended Overview)". In: *Conference and Labs of the Evaluation Forum*. 2024

Sexism identification: examples

No stereotype

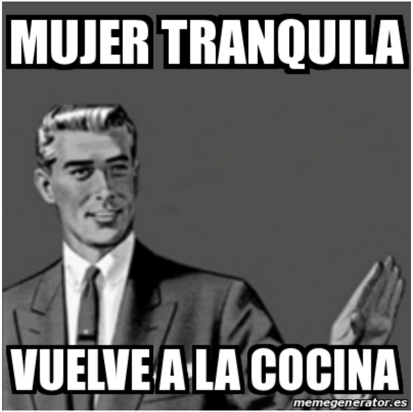

LA FRASE "NO SOY MACHISTA" NO PUEDE IR SEGUIDA DE LA CONJUNCIÓN "PERO"
POR FAVOR, NO SIGAS
memegenerator.es

Stereotype

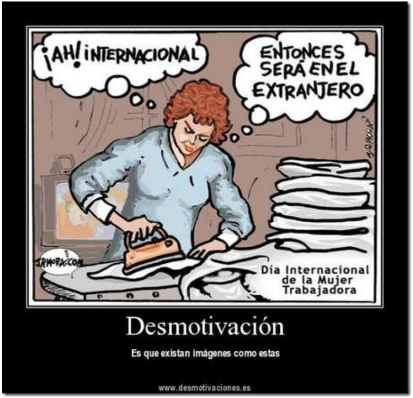

Son Chingaderas
Nos dieron 1 día sin mujeres, porque sabían que después nos iban a encerrar 40 días con ellas ...
Jajaja..tan.. tan..

Source intention: examples

Direct

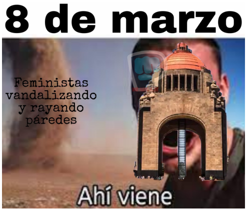


Judgmental

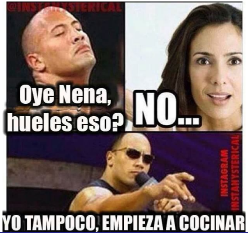

¡AH! INTERNACIONAL
ENTONCES SERÁ EN EL EXTRANJERO
JRMORA.COM
Día Internacional de la Mujer Trabajadora
Desmotivación
Es que existan imágenes como estas
www.desmotivaciones.es

Sexism categorization: examples

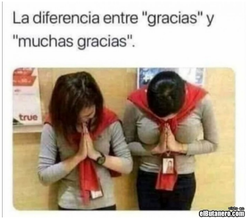
Ideological and inequality



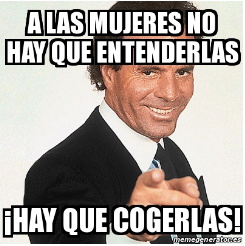
Stereotyping and dominance



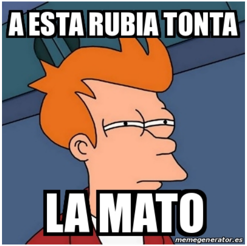
Objectification



Sexual violence



Misogyny and non-sexual violence



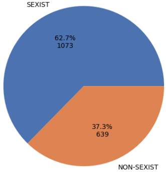
Dataset

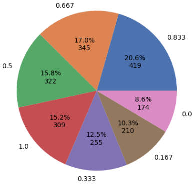
- Created by keyword search
- Various sources: Google, Twitter, Reddit and Forocoches
- English and Spanish
- Crowd annotation via Prolific
  - 450 annotators for each language
  - Each sample is annotated by 6 people
  - 26 annotations on average
  - Demographic information of each annotator

Laura Plaza et al. "Overview of EXIST 2024 - Learning with Disagreement for Sexism Identification and Characterization in Tweets and Memes (Extended Overview)". In: *Conference and Labs of the Evaluation Forum*. 2024

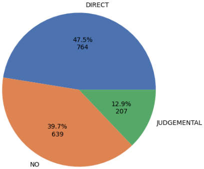
Dataset (ES)

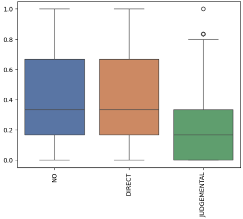
Task 4



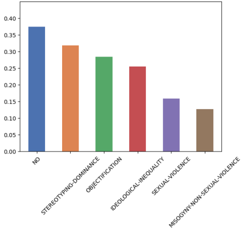


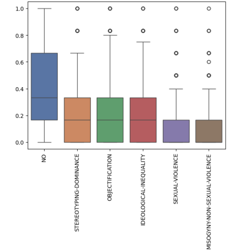
Task 5





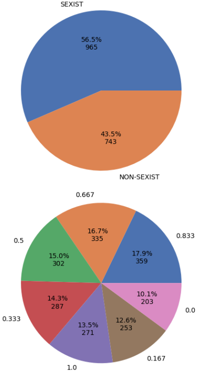
Task 6



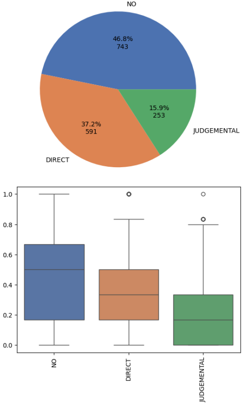


Dataset (EN)

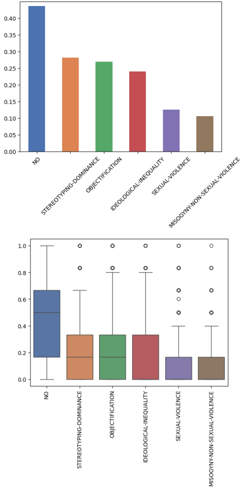
Task 4



Task 5

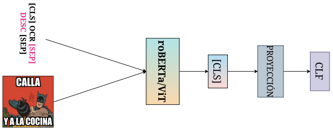


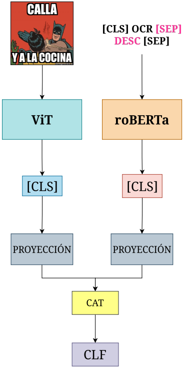
Task 6



System proposals

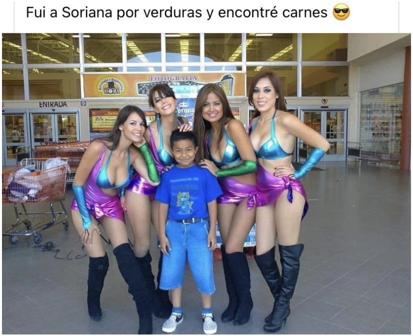
- Comparison of different modalities:
  - Unimodal architectures: *Transformer encoder* (RoBERTa and ViT)
  - Multimodal architecture: *Early fusion* by concatenating the [CLS] token
- LeWiDi approaches:
  - Hard label
  - Soft label
  - ***Why no multi-task approach?***
- One model for each language





Enhancing textual modality performance

- Textual preprocessing to remove noise (watermarks, emojis, URLS...)
- Avoid biases by masking *identity term*
- **Closing the visual and textual gap: meme descriptions (LLaVa)**
- **Data augmentation by incorporating tweets on the training dataset** [a]



Fui a Soriana por verduras y encontré carnes 😎

Figure: A group of women and a children pose for a photo.

[a] Available on text-only. Task 5 avoided.

Task 4: Sexism Identification in Memes

Architecture	Label	Ranking	ICM ↑	ICM Norm ↑	F1 - Sexist ↑
Text + CTXT	Hard	14	0.087	0.544	0.729
	Soft	13	0.088	0.545	0.697
Text + CTXT + Tweets	Hard	29	-0.093	0.453	0.684
	Soft	8	0.104	0.553	0.716
Image	Hard	43	-0.312	0.341	0.677
	Soft	45	-0.359	0.317	0.640
Early	Hard	4†	**0.166**	**0.584**	**0.736**
	Soft	34	-0.165	0.416	0.652
Gold Baseline	-	0	0.983	1.000	1.000
Ganadores	-	1	0.318	0.662	0.764

Table: Test results for the hard evaluation in task 4 of EXIST. In bold, our best results by metric. † denotes our best ranking.

Laura Plaza et al. "Overview of EXIST 2024 - Learning with Disagreement for Sexism Identification and Characterization in Tweets and Memes (Extended Overview)". In: *Conference and Labs of the Evaluation Forum*. 2024

Task 4: Sexism Identification in Memes

Architecture	Label	Ranking	ICM Soft ↑	ICM Soft Norm ↑	*Cross Entropy* ↓
Text + CTXT	Hard	2	-0.201	0.468	0.969
	Soft	25	-0.679	0.391	0.925
Text + CTXT + Tweets	Hard	17	-0.546	0.412	1.077
	Soft	11	-0.430	0.431	**0.918**
Image	Hard	27	-0.947	0.348	1.033
	Soft	34	-1.160	0.314	1.015
Early	Hard	1†	**-0.118**	**0.481**	1.081
	Soft	26	-0.869	0.360	0.980
Gold Baseline	-	0	3.111	1.000	0.585
Ganadores	-	1	-0.293	0.453	1.103

Table: Test results for the soft evaluation in task 5 of EXIST. In bold, our best results by metric. † denotes our best ranking.

Laura Plaza et al. "Overview of EXIST 2024 - Learning with Disagreement for Sexism Identification and Characterization in Tweets and Memes (Extended Overview)". In: *Conference and Labs of the Evaluation Forum*. 2024

Task 5: Source Intention in Memes

Architecture	Label	Ranking	ICM ↑	ICM Norm ↑	F1 - Sexist ↑
Text + CTXT	Hard	6	-0.272	0.406	0.382
	Soft	1†	**-0.207**	**0.428**	0.400
Image	Hard	16	-0.654	0.273	0.294
	Soft	20	-0.752	0.239	0.315
Early	Hard	13	-0.360	0.375	0.377
	Soft	2	-0.237	0.418	**0.411**
Gold Baseline	-	0	1.438	1.000	1.000
Ganadores	-	1	-0.240	0.417	0.387

Table: Test results for the hard evaluation in task 5 of EXIST. In bold, our best results by metric. † denotes our best ranking.

Laura Plaza et al. "Overview of EXIST 2024 - Learning with Disagreement for Sexism Identification and Characterization in Tweets and Memes (Extended Overview)". In: *Conference and Labs of the Evaluation Forum*. 2024

Task 5: Source Intention in Memes

Architecture	Label	Ranking	ICM Soft ↑	ICM Soft Norm ↑	*Cross Entropy* ↓
Text + CTXT	Hard	3†	**-1.323**	**0.359**	1.602
	Soft	8	-1.620	0.328	1.449
Image	Hard	10	-1.969	0.291	1.565
	Soft	13	-2.012	0.286	1.512
Early	Hard	9	-1.620	0.328	1.520
	Soft	5	-1.377	0.354	**1.434**
Gold Baseline	-	0	4.702	1.000	0.933
Ganadores	-	1	-1.245	0.368	1.624

Table: Test results for the soft evaluation in task 5 of EXIST. In bold, our best results by metric. † denotes our best ranking.

Laura Plaza et al. "Overview of EXIST 2024 - Learning with Disagreement for Sexism Identification and Characterization in Tweets and Memes (Extended Overview)". In: *Conference and Labs of the Evaluation Forum*. 2024

Task 6: Sexism Categorization in Memes

Architecture	Label	Ranking	ICM ↑	ICM Norm ↑	F1 - Sexist ↑
Text + CTXT	Hard	2†	**-0.783**	**0.338**	0.402
	Soft	5	-0.853	0.323	0.380
Text + CTXT + Tweets	Hard	8	-1.057	0.281	0.387
	Soft	3	-0.810	0.332	**0.434**
Image	Hard	20	-1.647	0.158	0.222
	Soft	21	-1.652	0.157	0.202
Early	Hard	11	-1.212	0.249	0.289
	Soft	12	-1.270	0.237	0.316
Gold Baseline	-	0	2.410	1.000	1.000
Ganadores	-	1	-0.700	0.355	0.432

Table: Test results of the hard evaluation in task 6 of EXIST. In bold, our best results by metric. † denotes our best ranking.

Laura Plaza et al. "Overview of EXIST 2024 - Learning with Disagreement for Sexism Identification and Characterization in Tweets and Memes (Extended Overview)". In: *Conference and Labs of the Evaluation Forum*. 2024

Task 6: Sexism Categorization in Memes

Architecture	Label	Ranking	ICM Soft ↑	ICM Soft Norm ↑
Text + CTXT	Hard	9	-5.737	0.196
	Soft	2	-4.609	0.256
Text + CTXT + Tweets	Hard	20	-8.080	0.072
	Soft	1†	**-4.310**	**0.272**
Image	Hard	11	-6.411	0.160
	Soft	14	-6.519	0.155
Early	Hard	7	-5.472	0.210
	Soft	8	-5.550	0.206
Gold Baseline	-	0	9.434	1.000
Ganadores	-	1	-4.904	0.245

Table: Test results of the soft evaluation in task 6 of EXIST. In bold, our best results by metric. † denotes our best ranking.

Laura Plaza et al. "Overview of EXIST 2024 - Learning with Disagreement for Sexism Identification and Characterization in Tweets and Memes (Extended Overview)". In: *Conference and Labs of the Evaluation Forum*. 2024

Index

1 Introduction

2 LeWiDi

3 DETESTS-Dis

4 EXIST

5 Conclusions and future work

Conclusions

- **RQ1**:
  - Competitive results
  - LeWiDi models achieve better results
  - LeWiDi models generalize better
  - Insights into the flaws of the multi-task approach
  - Context and data augmentation
  - Fine-tuning strategy
  - Difficulty of the implicit detection task
- **RQ2**:
  - Competitive results, even surpassing SOTA results
  - Text > Image
  - Importance of text description
  - LeWiDi models offer better generalization
  - Image-text relation in memes is very complex.

Future work (and some tips for EXIST 2025...)

- On BERT fine-tuning:
  - Intermediate tasks for boosting performance
  - Fine-tuned models on HF from similar tasks
  - Re-init BERT layers
- Stereotype on multimodal detection:
  - **External features are important, but so are the selected models**
  - Image/video description: **Qwen2.5 VL, GPT-like, SmolVLM2, Llama3.2 11B, Gemma 3**
  - Image features: **YOLO, Faster R-CNN, EfficientNet**
  - Improve text-image alignment
- LLM's
  - Data augmentation
  - Few-shot
  - Provide explanations
- Alternatives approaches for LeWiDi:
  - **Multi-task** with **hard and soft labels**
  - Perspectivist approach in EXIST with clustering and demographic information

Thanks for your attention!
Any questions?

PRHLT
Pattern Recognition and
Human Language Technology
Research Center

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

ICM

- Formula:

$$ICM(A,B) = 2IC(A) + 2IC(B) - 3IC(A \cup B)$$

- *Information Content* of one category:

$$\begin{aligned} IC(c_i) &= -\log_2(P(c_i)) \\ &\simeq -\log_2\left(\frac{|\bigcup_{c' \in \{c\} \cup \mathrm{Desc}(c')} \mathcal{I}_{c'}|}{|\bigcup_{c' \in \mathcal{C}} \mathcal{I}_{c'}|}\right) \end{aligned}$$

- IC of a set of categories:

$$\begin{aligned} IC(\{c_i\} \cup \{c_j\}) &= IC(c_i) + IC(c_j) - IC(\{c_i\} \cap \{c_j\}) \\ &= IC(c_i) + IC(c_j) - IC(\mathrm{lso}(c_i, c_j)) \end{aligned}$$

ICM

- Email:
  - Spam (S)
    - Scam (SC)
    - Viajes (TR)
  - NoSpam (NS)

ID	TRUTH	PRED
1	NS	NS
2	TR, SC	TR
3	SC	NS
4	NS	NS
5	TR	SC
6	NS	NS
7	NS	NS

ICM

- A prioris:

$$P(\text{NS}) = \frac{4}{7} \approx 0.571$$
$$P(\text{S}) = \frac{3}{7} \approx 0.429$$
$$P(\text{SC}) = P(\text{TR}) = \frac{2}{7} \approx 0.2857$$

- IC:

$$IC(\text{NS}) = -\log_2 P(\text{NS}) = -\log_2 0.571 \approx 0.80$$
$$IC(\text{S}) = -\log_2 P(\text{S}) = -\log_2 0.429 \approx 1.22$$
$$IC(\text{SC}) = IC(\text{TR}) = -\log_2 0.2857 \approx 1.80$$

ICM

- ID 1, 4, 6, 7:

$$\begin{aligned} ICM(\mathrm{NS}, \mathrm{NS}) &= 2IC(\mathrm{NS}) + 2IC(\mathrm{NS}) - 3IC(\{\mathrm{NS}\} \cup \{\mathrm{NS}\}) \\ &= 4IC(\mathrm{NS}) - 3IC(\mathrm{NS}) = IC(\mathrm{NS}) \\ &= 0.80 \end{aligned}$$

- ID 2:

$$\begin{aligned} IC(\{\text{TR, SC}\}, \{\text{TR}\}) &= 2IC(\{\text{TR, SC}\}) + 2IC(\{\text{TR}\}) - 3IC(\{\text{TR, SC}\}) \\ &= 2IC(\{\text{TR}\}) - IC(\{\text{TR, SC}\}) \\ &= 2 \cdot 1.8 - 2.38 = 1.22 \end{aligned}$$

$$\begin{aligned} IC(\{\text{TR, SC}\}) &= IC(\mathrm{TR}) + IC(\mathrm{SC}) - IC(\mathrm{lso}(\mathrm{TR}, \mathrm{SC})) \\ &= IC(\mathrm{TR}) + IC(\mathrm{SC}) - IC(\mathrm{S}) \\ &= 1.80 + 1.80 - 1.22 = 2.38 \end{aligned}$$

ICM

- ID 3:

$$\begin{aligned} ICM(\mathrm{SC},\mathrm{NS}) &= 2IC(\mathrm{SC}) + 2IC(\mathrm{NS}) - 3IC(\{\mathrm{SC}\} \cup \{\mathrm{NS}\}) \\ &= 2 \cdot 1.8 + 2 \cdot 0.8 - 3 \cdot 2.6 = -2.6 \end{aligned}$$

$$\begin{aligned} IC(\{\mathrm{SC},\mathrm{NS}\}) &= IC(\mathrm{SC}) + IC(\mathrm{NS}) - IC(\mathrm{lso}(\mathrm{SC},\mathrm{NS})) \\ &= 1.80 + 0.80 - 0 = 2.6 \end{aligned}$$

- ID 5:

$$\begin{aligned} ICM(\mathrm{SC},\mathrm{TR}) &= 2IC(\mathrm{SC}) + 2IC(\mathrm{TR}) - 3IC(\{\mathrm{SC},\mathrm{TR}\}) \\ &= 2 \cdot 1.80 + 2 \cdot 1.80 - 3 \cdot 2.38 = 0.06 \end{aligned}$$

ICM

Average ICM:

$$\frac{4 \cdot 0.8 + 1.22 - 2.6 + 0.06}{7} \approx 0.2685$$