ArcosGPT at EXIST 2025: Identifying Sexism in Memes with Multimodal Deep Learning: Fusing Text and Visual Cues

Notebook for the EXIST Lab at CLEF 2025

Iván Arcos[1]

[1]*Universitat Politècnica de València, Valencia, Spain*

Abstract

Sexism persists as a pervasive issue in society, particularly evident on social media platforms. This phenomenon encompasses a spectrum of expressions, ranging from subtle biases to explicit misogyny, posing unique challenges for detection and analysis. While previous research has predominantly focused on textual analysis, the dynamic nature of some social networks demands a more comprehensive approach. Multimodal analysis surpasses text-only methods, particularly in understanding sexism. Our results demonstrate that adding BLIP-generated image captions to OCR text raises F1-Macro from 0.6367 to 0.7298 (+9.3 points), and further including a GPT-4o description boosts it to 0.8114 (+8.2 points). The ViT+RoBERTa fusion model achieves the best overall performance (F1-Macro = 0.8308, +19.4 points over text-only), confirming that joint visual–text representations substantially enhance sexism detection. A similar pattern holds under soft-label training and across downstream tasks of intent classification and category categorization. These findings underscore the value of multimodal integration for robust, real-world sexism identification on social media.

Keywords

Multimodal Sexism Identification, Memes, Artificial Intelligence

1. Introduction

Sexism refers to multifaceted, encompassing subtle expressions that can be as insidious as explicit misogyny. Whether presented as seemingly positive remarks, jokes, or offensive comments, sexism permeates various aspects of individuals' lives, influencing domestic and parenting roles, career opportunities, sexual image, and life expectations. Recognizing the diverse forms of sexism is crucial to understanding its impact on society.

Social media platforms have become conduits for the dissemination of sexist content, perpetuating and even normalizing gender differences and biased attitudes. The Internet, with its vast reach, reflects and amplifies societal inequalities and discrimination against women. This study is particularly crucial given the significant presence of teenagers on social media platforms, urging the need for urgent investigation and societal dialogue, especially from an educational standpoint.

The rest of the paper is structured as follows. Section 2 presents some related work. Section 3 introduces the tasks of sexism detection, source intention classification and sexism categorization, as well as the the dataset. Section 4 describes the models for text and multimodal data. Section 5 and 6 presents the results and, finally, Section 7 draws some conclusions and discusses future work.

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

*Corresponding author.

†These authors contributed equally.

iarcgab@etsinf.upv.es (I. Arcos)

0009-0005-5290-1753 (I. Arcos)

2. Related Work

Hate Speech (HS) is generally described as any form of communication that belittles a person or a group based on attributes such as race, ethnicity, gender, sexual orientation, nationality, religion, among others [1]. When the target of hate speech is women, it manifests as a form of misogyny. However, *misogyny*, as defined by the Oxford English Dictionary [2], refers to feelings of hatred or dislike towards women, or beliefs that devalue women compared to men. Misogyny can exist in behaviors, attitudes, or beliefs that demean women or see them as inferior to men, without the need for overt hate speech. On the other hand, *sexism* is defined as prejudice, stereotyping, or discrimination, often against women, based on sex. Unlike misogyny, sexism can manifest subtly, such as through gender stereotypes, traditional gender roles or unequal access to opportunities [3].

The field of NLP has increasingly focused on detecting hate speech and sexism, driven by their growing societal impacts, especially on social platforms. Notable efforts include SemEval-2019 Task 5, which targeted hate speech against immigrants and women [4], and SemEval-2023 Task 10, which developed a hierarchical taxonomy of sexist content and a dataset of 20,000 social media comments to enhance detection explainability [5]

Since 2021, the EXIST task addresses the problem of sexism identification in social networks [6, 7, 8]. Recent advancements in multimodal analysis have significantly enhanced the detection of hate speech in memes and images. The Multimodal Hate Speech Event Detection task organized in 2023 explored binary and target-specific detection strategies in text-embedded images, demonstrating the effectiveness of multimodal approaches in identifying hate speech [9]. A novel method introduced in 2023 utilizes pre-trained vision-language models (PVLMs) for hateful meme detection [10]. Additionally, a study conducted in 2018 demonstrated the superiority of a multimodal approach over unimodal methods in detecting sexist content in advertisements [11]. The Multimedia Automatic Misogyny Identification (MAMI) task at SemEval-2022 focused on identifying misogynous content in memes [12]. In [13], the authors investigated both unimodal and multimodal approaches for recognizing misogynous memes and proposed a bias estimation and mitigation strategy—based on Bayesian Optimization—that corrected model predictions toward the true class in up to 61.43% of cases.

3. Tasks and Datasets

3.1. Tasks

Following EXIST [14], our aim is to address sexism identification in the following three tasks:

1. **Sexism Detection.** Determine if the meme contain sexist content. This is a binary classification:
    - **Not Sexist.** Memes not focusing on gender-related themes.
    - **Sexist.** Memes discussing or portraying gender-related stereotypes or issues.
2. **Source Intention Classification.** Categorize sexist memes based on the creation intent:
    - **Judgmental Sexist.** Memes sharing experiences of encountering sexism.
    - **Direct Sexist.** Memes explicitly promoting sexist beliefs.
3. **Sexism Categorization.** Classify sexist memes by the aspect of sexism they exhibit:
    - **Ideological and Inequality.** Memes undermining women's rights or contributions.
    - **Role Stereotyping and Dominance.** Memes perpetuating gender role stereotypes.
    - **Objectification.** Memes portraying women solely as objects of desire.
    - **Sexual Violence.** Memes containing or promoting sexual harassment or assault.
    - **Misogyny and Non-sexual Violence.** Memes expressing hostility or violence towards women.

3.2. Dataset

In this work, we partitioned the dataset into a 90% training split and a 10% test split. All models are trained on the training split and evaluated on the held-out test split.

Table 1
Label distribution across Train and Test splits

Task	Label	Train ($n = 3639$)				Test ($n = 405$)			
		Hard	Soft	Hard %	Soft %	Hard	Soft	Hard %	Soft %
Detection	NO	1549	9689	42.6%	44.4%	171	1066	42.2%	43.9%
	YES	2090	12145	57.4%	55.6%	234	1364	57.8%	56.1%
Intention	DIRECT	2207	7722	66.9%	63.6%	233	871	63.3%	63.9%
	JUDGEMENTAL	1081	4309	32.8%	35.5%	132	480	35.9%	35.2%
Category	IDEOLOGICAL-INEQUALITY	764	3751	25.2%	23.2%	97	439	29.8%	24.2%
	MISOGYNY-NON-SEXUAL-VIOLENCE	186	1846	6.1%	11.4%	20	227	6.1%	12.5%
	OBJECTIFICATION	860	4161	28.4%	25.7%	86	429	26.4%	23.7%
	SEXUAL-VIOLENCE	326	2017	10.8%	12.5%	39	227	12.0%	12.5%
	STEREOTYPING-DOMINANCE	889	4357	29.3%	26.9%	84	485	25.8%	26.8%

Table 1 shows that sexist memes outnumber non-sexist ones (57.4 % vs. 42.6 %), direct intent is more common than judgmental (66.9 % vs. 32.8 %), and stereotyping and objectification are the most frequent sexism categories, while misogyny and sexual violence remain rare.

4. Models & Methodology

To systematically evaluate how different levels of multimodal information affect sexism detection in memes, we experiment with four progressively richer model variants. We start with a strong text-only baseline built on RoBERTa and OCR-extracted meme text, then incrementally add visual cues: first by appending BLIP-generated captions to the text, next by incorporating a GPT-4o–derived high-level description, and finally by fusing raw image embeddings from ViT with text embeddings from RoBERTa. Below we describe each variant in turn, along with the data augmentations and training details designed to improve robustness and capture complementary visual–textual signals.

4.1. Model Variants

- **Text-only baseline**: RoBERTa [15] fine-tuned on OCR-detected meme text.
- **Text + BLIP** [16]: OCR text concatenated with BLIP-generated image captions.
- **Text + BLIP + GPT Description**: additionally include a GPT-4o explanation. Inputs: OCR text, BLIP captions, GPT-4o description.

  GPT-4o Prompt

  Describe the meme, identify whether it addresses sexist topics, and explain the intent (humor, critique, normalization, etc.). Justify your analysis.

- **ViT + RoBERTa** [17]: Fuse ViT image embeddings with RoBERTa text embeddings. For this variant, the following data augmentations were applied:
  - *Images*:
    - Random horizontal and vertical flips.
    - Random rotations up to 30°.
    - Random perspective distortions.
    - Random adjustments of brightness, contrast, saturation, and hue.
  - *Text*:

* Random token masking with a 10% probability per token.

These augmentations aim to improve model robustness by exposing it to various visual perturbations and textual variations during training.

4.2. Labeling Schemes

- **Task 1 (Detection).**
  - *Hard labels*: set $y = 1$ if at least three annotators marked the meme sexist, otherwise $y = 0$.
  - *Soft labels*: set $y = \frac{\#\text{sexist votes}}{6}$, e.g. $3/6 = 0.5$. (Used only in Task 1 to capture annotator uncertainty).
- **Task 2 (Intention).** Models struggled when using soft labels, so we train only on those memes with hard label $y = 1$.
- **Task 3 (Category Union).** For each category c, merge the six annotator labels by

$$y_c = \max_{j=1,\ldots,6} y_{j,c},$$

i.e. category c is positive if any annotator flagged it.

5. Results

5.1. Task 1

5.1.1. Hard Labels

In Table 2 we observe a clear progression from the text-only OCR baseline (F1-Macro = 0.6367) through multimodal enhancements. Adding BLIP captions raises F1-Macro by +0.0931 (to 0.7298), and the GPT-4o description contributes another +0.0816 (to 0.8114). The ViT+RoBERTa fusion achieves the highest F1-Macro of 0.8308, representing a +0.1941 gain over the baseline and +0.0194 over Text+BLIP+GPT. Across languages, English posts consistently outperform Spanish (e.g. ViT+RoBERTa: EN 0.8657 vs. ES 0.7897), indicating stronger model alignment with English meme content.

Table 2
Detection (Task 1) — Hard labels

Model	Lang	F1-Macro	F1(+)	ICM-Soft	ICM-Soft-Norm	ICM	ICM-Norm
Baseline	All	0.6367	0.7838	-0.6673	0.3773	-0.0941	0.4493
	EN	0.6660	0.7791	-0.6821	0.3741	-0.0260	0.4861
	ES	0.6056	0.7875	-0.6598	0.3791	-0.1563	0.4149
Text + BLIP	All	0.7298	0.8092	-0.4926	0.4094	0.1411	0.5760
	EN	0.7350	0.8034	-0.3798	0.4299	0.1632	0.5869
	ES	0.7245	0.8138	-0.5964	0.3907	0.1203	0.5655
Text + BLIP + GPT	All	0.8114	0.8775	0.4117	0.5757	0.3771	0.7032
	EN	0.8548	0.9032	0.6702	0.6236	0.5081	0.7707
	ES	0.7734	0.8562	0.1809	0.5332	0.2632	0.6433
ViT + RoBERTa	All	0.8308	0.8830	1.3312	0.7445	0.4230	0.7266
	EN	0.8657	0.8927	1.6064	0.7998	0.5732	0.7948
	ES	0.7897	0.8754	1.0561	0.6892	0.2729	0.6584

5.1.2. Soft Labels

When we switch to soft labels (Table 3), the OCR baseline improves slightly by +1.00 point to 0.6466; Text+BLIP gains +5.78 points over that soft baseline (to 0.7044); Text+BLIP+GPT adds +9.49 points (to 0.7993); and ViT+RoBERTa attains +11.07 points (to 0.7573) relative to the soft baseline. Comparing hard vs. soft labels for ViT+RoBERTa, F1-Macro drops from 0.8308 to 0.7573 (−7.35 points). This shows that soft labeling takes into account diverse annotator perspectives, but comes at the cost of some discriminative strength in our top multimodal model.

Table 3
Detection (Task 1) — Soft labels

Model	Lang	F1-Macro	F1(+)	ICM-Soft	ICM-Soft-Norm	ICM	ICM-Norm
Baseline	All	0.6466	0.7652	-0.6179	0.4028	-0.0935	0.4496
	EN	0.6497	0.7382	-0.9258	0.3292	-0.0918	0.4511
	ES	0.6390	0.7857	-0.8866	0.3376	-0.0992	0.4460
Text + BLIP	All	0.7044	0.8080	-0.9909	0.3442	0.0747	0.5402
	EN	0.7439	0.8320	-0.2021	0.4627	0.1947	0.6037
	ES	0.6705	0.7879	-0.5540	0.3985	-0.0295	0.4840
Text + BLIP + GPT	All	0.7993	0.8741	-0.4081	0.4358	0.3429	0.6848
	EN	0.8346	0.8924	0.7985	0.6473	0.4497	0.7396
	ES	0.7679	0.8590	0.3850	0.5705	0.2495	0.6359
ViT + RoBERTa	All	0.7573	0.8358	0.3749	0.5584	0.2058	0.6095
	EN	0.7739	0.8205	0.4379	0.5679	0.3014	0.6550
	ES	0.7322	0.8477	0.3118	0.5489	0.1102	0.5640

5.2. Task 2

Table 4 illustrates the challenge of intention classification. The text-only baseline achieves F1-Macro = 0.5934, BLIP alone degrades it to 0.4831, and Text+BLIP+GPT recovers to 0.6127 (+0.0193 over baseline). We omitted ViT+RoBERTa here due to inconsistent gains. Notably, Spanish examples outperform English in Task 2 (0.6627 vs. 0.5556 for Text+BLIP+GPT), suggesting that cultural or linguistic cues in Spanish data facilitate intent detection.

Table 4
Intention (Task 2) — Hard labels

Model	Lang	F1-Macro	F1(+)	ICM-Soft	ICM-Soft-Norm	ICM	ICM-Norm
Baseline	All	0.5934	0.7076	-1.4823	0.2328	-0.3155	0.3199
Baseline	EN	0.5607	0.6452	-1.5987	0.2074	-0.3906	0.2867
Baseline	ES	0.6211	0.7594	-1.4133	0.2501	-0.2543	0.3470
Text + BLIP	All	0.4831	0.4773	-2.0559	0.1294	-0.6159	0.1485
Text + BLIP	EN	0.3726	0.5000	-2.1920	0.0999	-0.7738	0.0739
Text + BLIP	ES	0.5528	0.4727	-1.9656	0.1511	-0.4798	0.2143
Text + BLIP + GPT	All	0.6127	0.7012	-1.4451	0.2395	-0.2573	0.3532
Text + BLIP + GPT	EN	0.5556	0.6069	-1.6002	0.2071	-0.4017	0.2807
Text + BLIP + GPT	ES	0.6627	0.7760	-1.3361	0.2638	-0.1289	0.4225

5.3. Task 3

Table 5 confirms that multi-label categorization benefits from multimodal fusion. The baseline starts at F1-Macro = 0.6597; Text+BLIP adds +0.0350 (to 0.6947); Text+BLIP+GPT adds +0.0783 (to 0.7380);

and ViT+RoBERTa peaks at 0.7562 (+0.0965 over baseline, +0.0182 over Text+BLIP+GPT). Spanish posts slightly outperform English here (0.7670 vs. 0.7562 for ViT+RoBERTa), demonstrating robust cross-lingual generalization in category detection. Within the ViT+RoBERTa model, the Stereotyping category is predicted best (F1 = 0.8517), followed by Objectification (0.8072) and Ideological-Inequality (0.8023), whereas Sexual Violence (0.6437) and Misogyny (0.6763) remain the most challenging categories.

Table 5
Task 3 (Category Union) — Complete results by model and language

Model	Lang	F1-Macro	Ideol. Ineq.	Stereotype	Misog.	Objectif.	Sex-Viol.	ICM	ICM-Norm
Baseline (Text-only)	All	0.6597	0.8297	0.8710	0.2690	0.7556	0.5730	-2.1477	0.3450
	EN	0.6352	0.8432	0.8831	0.1951	0.7543	0.5000	-2.3557	0.3324
	ES	0.6793	0.8156	0.8595	0.3371	0.7568	0.6275	-2.1146	0.3459
Text + BLIP	All	0.6947	0.8207	0.8710	0.4885	0.7521	0.5412	-2.0969	0.3486
	EN	0.6927	0.8043	0.8831	0.5217	0.7229	0.5316	-2.3225	0.3348
	ES	0.6951	0.8370	0.8595	0.4510	0.7784	0.5495	2.0395	0.3513
Text + BLIP + GPT	All	0.7380	0.8164	0.8710	0.5983	0.7762	0.6279	-1.4771	0.3933
	EN	0.7222	0.8108	0.8831	0.6000	0.7765	0.5405	-1.8350	0.3695
	ES	0.7495	0.8222	0.8595	0.5962	0.7760	0.6939	-1.2920	0.4058
ViT + RoBERTa	All	0.7562	0.8023	0.8517	0.6763	0.8072	0.6437	0.0383	0.5029
	EN	0.7562	0.8023	0.8517	0.6763	0.8072	0.6437	-0.1041	0.4926
	ES	0.7670	0.8171	0.8571	0.6286	0.7865	0.7455	0.1807	0.5132

6. Competition Ranking

Participation in the EXIST 2025 shared task was registered under the team name *ArcosGPT*. According to the official task overview [18], third place was achieved in the hard–hard evaluations of all three subtasks, using the full multimodal fusion model (ViT+RoBERTa) in each case. Tables 6–8 present the performance alongside the top five systems in each hard–hard evaluation.

Table 6
Task 2.1: Sexism Identification in Memes (Hard–Hard Evaluation)

System	Team	ICM-Hard	ICM-Hard Norm	F1(YES)	Rank
CogniCIC_1	CogniCIC	0.3691	0.6877	0.7810	1
GrootWatch_3	GrootWatch	0.3589	0.6825	0.7740	2
ArcosGPT_1	ArcosGPT	0.3200	0.6627	0.7571	3
TrankilTwice_2	TrankilTwice	0.1667	0.5848	0.7508	5
I2C-UHU-Altair_2	I2C-UHU-Altair	-0.0134	0.4932	0.7125	9

Table 7
Task 2.2: Source Intention in Memes (Hard–Hard Evaluation)

System	Team	ICM-Hard	ICM-Hard Norm	Macro F1	Rank
CogniCIC_1	CogniCIC	0.2254	0.5784	0.5634	1
GrootWatch_3	GrootWatch	0.1868	0.5649	0.5513	2
ArcosGPT_1	ArcosGPT	0.0597	0.5208	0.5109	3
NaturalThinker_1	NaturalThinker	-0.5429	0.3113	0.3762	6
I2C-UHU-Altair_1	I2C-UHU-Altair	-0.6519	0.2734	0.2685	7

Table 8
Task 2.3: Sexism Categorization in Memes (Hard–Hard Evaluation)

System	Team	ICM-Hard Norm	Macro F1	Rank
CogniCIC_1	CogniCIC	0.5051	0.5763	1
GrootWatch_3	GrootWatch	0.4834	0.5472	2
ArcosGPT_1	ArcosGPT	-0.4187	0.5501	3
I2C-UHU-Altair_1	I2C-UHU-Altair	-0.9958	0.4223	4
UMUTeam_1	UMUTeam	-1.5624	0.3582	5

7. Conclusions and Future Work

This study demonstrates the effectiveness of multimodal approaches combining textual and visual information. By integrating OCR-extracted text, BLIP-generated captions, GPT-4o contextual descriptions, and visual embeddings from ViT models, we achieved significant performance improvements across all tasks, particularly in fine-grained classification scenarios. The ViT+RoBERTa fusion model achieved the highest performance, validating the strength of joint visual-textual representations.

Despite these advancements, several avenues remain open for further research. First, **dataset expansion** is critical to improve model generalizability, especially in capturing underrepresented forms of sexism and ensuring cultural diversity.

We also plan to investigate **new multimodal features**, including sociolinguistic cues, and user interaction patterns such as comments and reactions, to enhance contextual understanding. In addition, an analysis of content diffusion may help reveal the mechanisms of propagation and offer insights for effective mitigation strategies. Finally, adopting **advanced model architectures**—such as cross-modal attention mechanisms—could further enhance the synergy between modalities and elevate performance across all subtasks.

Acknowledgments

This work was done in the framework of Malicious Actors Profiling and Detection in Online Social Networks Through Artificial Intelligence (MARTINI) research project, funded by MCIN/AEI/10.13039/501100011033 and by NextGenerationEU/PRTR (Grant PCI2022-135008-2).

Declaration on Generative AI

During the preparation of this work, the authors used OpenAI's ChatGPT and GPT-4o models in order to: (i) generate multimodal descriptions of memes used as additional model input features, (ii) assist with grammar and spelling checking, and (iii) rephrase and improve the clarity of some sentences.

After using these tools, the authors carefully reviewed and edited the content as needed and take full responsibility for the final text. No generative AI tool was used for developing the scientific insights, analysis, or conclusions of this paper.

References

[1] J. T. Nockleby, Hate Speech, 2 ed., Macmillan, New York, 2000.
[2] Oxford English Dictionary, Misogyny, https://www.oed.com/view/Entry/misogyny, 2025. Definition of misogyny. Accessed: 2025-06-05.
[3] Oxford English Dictionary, Sexism, https://www.oed.com/view/Entry/sexism, 2025. Definition of sexism. Accessed: 2025-06-05.

[4] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. Rangel Pardo, P. Rosso, M. Sanguinetti, SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter, 2019, pp. 54–63.

[5] H. Kirk, W. Yin, B. Vidgen, P. Röttger, SemEval-2023 task 10: Explainable detection of online sexism, in: A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2193–2210.

[6] F. Rodríguez-Sánchez, et al., Overview of EXIST 2021: Sexism identification in social networks, Procesamiento del Lenguaje Natural 67 (2021) 195–207.

[7] F. Rodríguez-Sánchez, et al., Overview of EXIST 2022: Sexism identification in social networks, Procesamiento del Lenguaje Natural 69 (2022) 229–240.

[8] L. Plaza, J. Carrillo-de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2023 – learning with disagreement for sexism identification and characterization, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer Nature Switzerland, Cham, 2023, pp. 316–342.

[9] S. Thapa, F. A. Jafri, A. Hürriyetoğlu, F. Vargas, R. K. W. Lee, U. Naseem, Multimodal Hate Speech Event Detection - shared task 4, case 2023, in: Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text, Incoma Ltd, 2023, pp. 151–159.

[10] R. Cao, M. S. Hee, A. Kuek, W.-H. Chong, R. K.-W. Lee, J. Jiang, Pro-Cap: Leveraging a Frozen Vision-Language Model for Hateful Meme Detection, in: Proceedings of the 31st ACM International Conference on Multimedia (MM '23), 2023, pp. 5244–5252. doi:10.1145/3581783.3612498.

[11] F. Gasparini, I. Erba, E. Fersini, S. Corchs, Multimodal Classification of Sexist Advertisements, in: Proceedings of the 15th International Joint Conference on e-Business and Telecommunications (ICETE), 2018, pp. 565–572. doi:10.5220/0006859405650572.

[12] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, J. Sorensen, Semeval-2022 task 5: Multimedia automatic misogyny identification, in: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, 2022, pp. 533–549.

[13] G. Rizzi, F. Gasparini, A. Saibene, P. Rosso, E. Fersini, Recognizing misogynous memes: Biased models and tricky archetypes, Information Processing Management 60 (2023) 103474. doi:10.1016/j.ipm.2023.103474.

[14] L. Plaza, J. Carrillo-de Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, EXIST 2025: Sexism Identification in Social Networks, http://nlp.uned.es/exist2025/, 2025. Accessed: 2025-06-05.

[15] Facebook AI, FacebookAI/xlm-roberta-large, https://huggingface.co/FacebookAI/xlm-roberta-large, 2021. Accessed: 2025-06-05.

[16] Salesforce AI Research, salesforce/blip-image-captioning-base, https://huggingface.co/salesforce/blip-image-captioning-base, 2022. Accessed: 2025-06-05.

[17] Google Research, google/vit-base-patch16-224, https://huggingface.co/google/vit-base-patch16-224, 2022. Accessed: 2025-06-05.

[18] L. Plaza, J. Carrillo-de Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, Overview of EXIST 2025: Learning with Disagreement for Sexism Identification and Characterization in Tweets, Memes, and TikTok Videos (Extended Overview), in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.