# DiTana-PV at sEXism Identification in Social neTworks (EXIST) Tasks 4 and 6: The Effect of Translation in Sexism Identification

Notebook for the sEXism Identification in Social neTworks (EXIST) Lab at CLEF 2024

Aitana Menárguez-Box[1,*], Diego Torres-Bertomeu[2,*]

[1]Pattern Recognition and Human Language Technology (PRHLT) Research Center, Spain
[2]Valencian Research Institute for Artificial Intelligence (VRAIN), Spain

**Abstract**

This paper details the participation of DiTana-PV team in the sEXism Identification in Social neTworks (EXIST) task at CLEF 2024. Specifically, we focused on Tasks 4 and 6, which involved identifying and categorizing sexism in memes. Our primary objective was to evaluate the effect of machine translation on model performance, as well as to explore data augmentation techniques and task combination strategies. By translating Spanish data to English and leveraging a pretrained BERTweet model fine-tuned for sexism detection, we aimed to improve classification accuracy. This work highlights the potential of translation and data handling techniques to enhance multilingual NLP tasks, contributing to more inclusive and effective AI applications in social media analysis.

**Keywords**

Sexim Identification, Data Augmentation through Machine Translation, Automatic Analysis of Memes, Pretrained Models Usage, BERTweet

## 1. Introduction

Sexism identification in social networks is an increasingly critical task given the proliferation of user-generated content that often contains harmful and discriminatory language. The Conference and Labs of the Evaluation Forum (CLEF) 2024 has organized the sEXism Identification in Social neTworks (EXIST) lab, which focuses on the automated detection and categorization of sexist content. This paper presents the efforts of the DiTana-PV team in addressing two specific tasks within this lab: Task 4 (sexism identification in memes) and Task 6 (categorization of sexism types in memes).

### 1.1. Task Description

The proposed lab of CLEF 2024 was sEXism Identification in Social neTworks (EXIST) [1, 2, 3]. Between all the different tasks proposed in EXIST, this paper details our team's (DiTana-PV) participation in Tasks 4 and 6: sexism identification and categorization in memes, respectively. Given an image (a meme), these aim to try and classify it as sexist or not sexist as well as which kinds of sexism, if any, are present from the following: (i) IDEOLOGICAL AND INEQUALITY, (ii) STEREOTYPING AND DOMINANCE, (iii) OBJECTIFICATION, (iv) SEXUAL VIOLENCE and (v) MISOGYNY AND NON-SEXUAL VIOLENCE.

### 1.2. Data Distribution

The information provided includes the memes images and their transcriptions into texts, along with some information about the annotators. We divided the dataset into three different partitions: `train`, `validation` and `test`. Tab. 1 shows the sample distribution of each language.

*CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France*
*These authors contributed equally.
amenbox@prhlt.upv.es (A. Menárguez-Box); dtorber@etsinf.upv.es (D. Torres-Bertomeu)
0009-0000-5957-0698 (A. Menárguez-Box); 0009-0009-2179-5942 (D. Torres-Bertomeu)

**Table 1**
Dataset samples distribution per partition. Additionally to the number of samples, the table also shows the percentage over the language.

| | train | val | test |
|---|---|---|---|
| **English** | 1809 (71.70%) | 201 (7.97%) | 513 (20.33%) |
| **Spanish** | 1830 (71.09%) | 204 (7.93%) | 540 (20.97%) |

In Tab. 2 we can see how each class for Task 4 is distributed along both languages for the train and validation partitions. As we trained the models with the hard labels there were some samples that half of the annotators labeled as sexist and the other half as not sexist (`tie`). After some experimentation, the best solution found to this situation was to discard the ambiguous samples, considering them noisy samples.

**Table 2**
Dataset classes distribution per language for `train` and `validation` partitions. Additionally to the number of samples, the table also shows the percentage over the language.

| | sexist | not sexist | tie |
|---|---|---|---|
| **English** | 965 (48.00%) | 743 (36.96%) | 302 (15.02%) |
| **Spanish** | 1073 (52.75%) | 639 (31.42%) | 322 (15.83%) |

### 1.3. Performance Measures

To measure the performance of our proposed systems, the Intermediate Concept Measure (ICM) [4] has been chosen as the official metric, although F1-score is also provided. ICM is a similarity function that generalizes Pointwise Mutual Information (PMI), and can be used to evaluate system outputs in classification problems by computing their similarity to the ground truth categories. These metrics were computed for the Hard-Hard evaluation as we trained our models with hard labels.

## 2. Main Objectives of Experiments

Here we described our main objectives and experimental findings for our participation in this competition. All this objectives focused on text processing, omitting the part of image processing for the memes.

### 2.1. The Effect of Machine Translation

The main objective of participating in this competition was to develop models that could detect sexism in memes and categorize it. More specifically, we were interested in how Machine Translation could affect the model's performance.

As Machine Translation has advanced enormously and there are far more resources in English than in any other language, some minority languages that may not have enough resources to train this kind of models could benefit from this: the idea of previously translating the content to English and using any model that was trained with all the available data in English.

This is an open line of research that we think has lots of future because it could mean a democratization of the benefits of Machine Learning for all languages, regardless of their available resources.

### 2.2. The Power of Data

We also aimed to evaluate how the use of Data Augmentation affects the performance of the models. Since it is a very widespread technique in the world of Machine Learning, especially in Computer Vision, but it has been proven that NLP tasks can also take advantage from it.

As the dataset was unbalanced, as we could see on Tab. 2, another of our objectives was to work with this type of datasets in which not all classes have the same presence and therefore a series of measures must be taken to prevent the model from being biased.

### 2.3. Tasks Combination

The last of our objectives, specially useful for Task 6, was to see what effect caused the combination of inferences for both tasks. We wanted to try to *help* a model used for Task 6 with the predictions of a model used for Task 4. This will be further explained in the following sections.

## 3. Approaches Used

In this section we will explain in detail the different approaches that best worked for us.

### 3.1. Using Machine Translation

After some preliminary experiments we discovered that the results in the English dataset were always some points higher than in the Spanish dataset, so we decided to automatically translate the Spanish samples into English samples, assuming the loss of quality that such automatic translation could entail. This approach was interesting because there are lots of resources in English that could improve our results.

### 3.2. The Base Model

For both tasks and languages (as we translated the Spanish samples to English), we used as pretrained model the *BERTweet-large-sexism-detector* [5] that was presented to the SemEval-2023 Task 10. It is a fine-tuned model of BERTweet-large [6] on the Explainable Detection of Online Sexism (EDOS) dataset [7]. This is intended to be used as a classification model for identifying tweets as `sexist` or `not sexist`. In spite of that, the model me all our needs for the competition because the text in a meme is quite short, as it is in tweets, so using a model that is trained with shorter texts should be beneficial. It also gives us a great advantage, not starting from a model that has simply been pretrained on general tasks, but has also been fine-tuned for tasks in the same domain of the ones in this competition.

### 3.3. Managing the Data

As seen in Tab. 2 there were two main problems with hard labels: there were some ambiguous ones and the amount of `sexist` and `not sexist` samples was unbalanced. We decided to discard the ambiguous samples, because taking them into account in the training process just added noise and led to worse performance. To try and solve the class-imbalance, for Task 4, we applied a weighted-loss function in order to give more weight to the `not sexist` class.

In some of the models, we also performed Data Augmentation to increase the amount of training data. This technique has proven quite good results in the vast majority of Machine Learning tasks and NLP is not an exception [8]. In particular, we used BERT contextual embeddings for paraphrasing the words in the original text. From each sample we generated three new augmented ones and a 30% of the words were substituted using the *nlpaug* library [9].

### 3.4. Combining Inferences

For Task 6, we trained the models for recognising six different labels for each meme: the ones for the type of sexism detected plus an extra one for not sexism detected at all. As one of our purposes was to try to combine the information from both tasks, we also focused on training a model for predicting just the main five labels for the type of sexism inside the meme.

We would use our inferences from Task 4 to detect if the meme was sexist or not, previously, and then in case the meme was classified as `sexist`, the second model would predict what type(s) of sexism were inside the given meme. This approach, as will be seen further in this paper, has been proven to work finely.

# 4. Models

In this section there is a description of each one of the six models that were presented to the competition. All these were fine-tuned for 10 epochs with an NVIDIA RTX 3090 with 24GB. The relevant hyper-parameters are: 8 samples per device, and a learning rate of $5 \cdot 10^{-5}$ with a linear scheduler.

## 4.1. Task 4 – Classification Models

For Task 4 we developed 3 different models following the different approaches explained. For all the scenarios described, one separate model was trained for each language.

The first pair of models ($\mathbf{M1}_{t4}$) were trained with all the samples from the train partition in English and in translated Spanish, for the English model we also added the validation partition in translated Spanish and vice versa. The second set of models ($\mathbf{M2}_{t4}$) were trained exactly as $\mathbf{M1}_{t4}$ but we added a weighted-loss function to correct the explained class-imbalanced. The last pair of models ($\mathbf{M3}_{t4}$) were trained as $\mathbf{M2}_{t4}$ but increasing the training dataset applying Data Augmentation.

## 4.2. Task 6 – Categorization Models

For this task, we developed 5 different models, according to the approaches in the previous sections: a pair for predicting 6 labels, a pair for predicting 5 labels and a single model for predicting six labels for both languages together. For the first two pairs, there is a model for each language (English and translated Spanish).

For the first pair $\mathbf{M1}_{t6}$, the Spanish model was trained with translated Spanish and the English one was trained with the English samples. The second pair, $\mathbf{M2}_{t6}$, follows the same logic but for making five label predictions. The last one $\mathbf{M3}_{t6}$ is not a pair but a single model which was trained with both translated Spanish and English samples.

# 5. Analysis of the Results

In this section we will discuss the results obtained for each of the models presented.

## 5.1. Task 4 – Classification Models

For the classification models, we can see that just by translating the Spanish samples to English, combining all of the samples in the training process and using as checkpoint the BERTweet model already fine-tuned in a sexism detection task, was enough to achieve a good result. Although, as we can see in the Annex A, it is lower than what was obtained during validation.

Adding weights to the loss function to gives *more importance* to the `not sexist` samples as they appear fewer times, allowing us to enhance a little bit the model's performance, as was already proved during validation.

When we applied Data Augmentation techniques we got worse results. This was surprising because during validation for Spanish it improved the results, and for English it got lower results but it was far more what we won in Spanish than what we lost in English. We hypothesize the reason why Data Augmentation this is because we already inserted noise in the Spanish samples through Machine Translation, thus the addition of more noise through paraphrasing with BERT was too much.

**Table 3**
Official final results (*Task 4 Hard-Hard ALL*) of the competition for the inferences presented to Task 4.

| Model | Run ID | Rank | ICM-Hard | ICM-Hard Norm | F1_YES |
|---|---|---|---|---|---|
| $M1_{t4}$ | *DiTana-PV_1.json* | 18 | 0.0337 | 0.5171 | 0.6908 |
| **$M2_{t4}$** | **DiTana-PV_2.json** | **6** | **0.1150** | **0.5585** | **0.7122** |
| $M3_{t4}$ | *DiTana-PV_3.json* | 10 | 0.0888 | 0.5451 | 0.7082 |

### 5.2. Task 6 – Categorization Models

For the categorization models in Task 6, the results indicate a varied performance across different approaches. As shown in Table 4, the models trained to predict six labels for each language separately ($M1_{t6}$) achieved the highest performance in terms of ICM-Hard and ICM-Hard Norm metrics, although its Macro F1 score was slightly lower than that of the M2$t6$ models.

The $M1_{t6}$ models, which predicts six labels, performed the best in terms of ICM-Hard and ICM-Hard Norm. This suggests that having a separate model for each language and focusing on the six distinct labels allowed the model to better capture the nuances of sexism categorization.

Interestingly, the $M2_{t6}$ models, which was trained to predict five labels, did not outperform the first models, indicating that removing the `not sexist` label from the categorization task, together with the possible error the previous model (as it was a joint prediction) may have introduced, might have led to a loss of valuable context needed for accurate categorization.

The single model trained on both English and translated Spanish samples ($M3_{t6}$) performed the worst. This may be due to the increased complexity and noise introduced by combining data from two languages, even after translation. The challenges of handling translated text, which may not perfectly capture the original sentiment or nuances, likely contributed to the poorer performance.

**Table 4**
Official final results (*Task 6 Hard-Hard ALL*) of the competition for the three models presented to Task 6.

| Model | Run ID | Rank | ICM-Hard | ICM-Hard Norm | Macro F1 |
|---|---|---|---|---|---|
| **$M1_{t6}$** | **DiTana-PV_1.json** | **1** | **-0.6996** | **0.3549** | **0.4319** |
| $M2_{t6}$ | *DiTana-PV_2.json* | 2 | -0.8450 | 0.3247 | 0.4430 |
| $M3_{t6}$ | *DiTana-PV_3.json* | 9 | -1.3691 | 0.2160 | 0.3255 |

## 6. Conclusions and Future Work

The results obtained from our participation in the sEXism Identification in Social neTworks (EXIST) competition demonstrate the potential impact of machine translation and data augmentation on improving model performance for sexism detection and categorization tasks.

Our experiments highlighted the benefits of translating minority language datasets into English, utilizing the wealth of available resources and pretrained models in English to enhance performance. This approach showed significant improvements in classification accuracy, suggesting a promising direction for future research aimed at democratizing the benefits of machine learning across languages with fewer resources.

The implementation of a weighted-loss function effectively addressed class imbalance, further improving model performance. However, the addition of data augmentation techniques, while beneficial during validation, did not consistently enhance results in the final evaluation, indicating the need for careful consideration of noise introduced by such methods, especially when combined with machine-translated data.

In Task 6, combining inferences from Task 4 to aid categorization proved to be a viable strategy, though the overall performance varied across different model configurations. This underscores the complexity of multi-label classification tasks and the importance of tailored model training approaches.

Overall, this highlights the value of leveraging translation and sophisticated data handling techniques to improve model accuracy in NLP tasks involving multiple languages. Future work could focus on further refining these methods, exploring visual-textual integration for meme analysis, and investigating more robust data augmentation strategies to mitigate noise.

These contributions provide a foundation for advancing sexism detection in multilingual contexts, paving the way for more inclusive and effective AI applications in social network analysis.

## Acknowledgement

This work is partially supported by the *Valencian Graduate School and Research Network of Artifficial Intelligence* (ValgrAI).

## References

[1] L. Plaza, J. Carrillo-de Albornoz, E. Amigó, J. Gonzalo, R. Morante, P. Rosso, D. Spina, B. Chulvi, A. Maeso, V. Ruiz, Exist 2024: sexism identification in social networks and memes, in: N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2024, pp. 498–504.

[2] L. Plaza, J. Carrillo-de-Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of EXIST 2024 – Learning with Disagreement for Sexism Identification and Characterization in Social Networks and Memes, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), 2024.

[3] L. Plaza, J. Carrillo-de-Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of EXIST 2024 – Learning with Disagreement for Sexism Identification and Characterization in Social Networks and Memes (Extended Overview), in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[4] E. Amigo, A. Delgado, Evaluating extreme hierarchical multi-label classification, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 5809–5819. URL: https://aclanthology.org/2022.acl-long.399. doi:10.18653/v1/2022.acl-long.399.

[5] A. Rydelek, D. Dementieva, G. Groh, AdamR at SemEval-2023 task 10: Solving the class imbalance problem in sexism detection with ensemble learning, 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. 1371–1381. URL: https://aclanthology.org/2023.semeval-1.190. doi:10.18653/v1/2023.semeval-1.190.

[6] D. Q. Nguyen, T. Vu, A. Tuan Nguyen, BERTweet: A pre-trained language model for English tweets, in: Q. Liu, D. Schlangen (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 9–14. URL: https://aclanthology.org/2020.emnlp-demos.2. doi:10.18653/v1/2020.emnlp-demos.2.

[7] 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. URL: https://aclanthology.org/2023.semeval-1.305. doi:10.18653/v1/2023.semeval-1.305.

[8] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, E. Hovy, A survey of data augmentation approaches for NLP, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Findings of the

Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 968–988. URL: https://aclanthology.org/2021.findings-acl.84. doi:10.18653/v1/2021.findings-acl.84.
[9] E. Ma, Nlp augmentation, https://github.com/makcedward/nlpaug, 2019.
[10] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.
[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

# A. Validation Results

In this appendix section illustrate the results that each of the models obtained for our validation partition of the dataset for each task, distributed by language.

## A.1. Task 4 – Classification Models

As we can see in Tab. 5 the starting pint was not brilliant, although we could improve our results for the Spanish split by four points just by translating the samples to English and using the BERTweet fine-tuned model for sexism detection. The weighted-loss function has also allowed to increase some points and the Data Augmentation has also enhanced the model performance.

In Tab. 6 we can see that, just comparing the baselines using BERT and BETO, the English partition gets a better performance. As in the Spanish validation split, just with the combination of the English and translated Spanish datasets and using the *BERTweet-large-sexism-detector* model as checkpoint, we already obtained good results. Adding the weighted-loss improved them, as happened with the Spanish dataset. In this case, differently from what happened in the Spanish dataset, the Data Augmentation did not improve model's performance.

**Table 5**
Table of results for Task 4 in the **Spanish** partition for our validation set.
[1]Baseline model is achieved fine-tuning BETO [10] pretrained model with the Spanish samples without translation.

| Model | ICM-Hard | ICM-Hard Norm | F1_YES |
|---|---|---|---|
| Baseline[1] | 0.069 | 0.535 | 0.687 |
| $M1_{t4}$ | 0.174 | 0.588 | 0.727 |
| $M2_{t4}$ | 0.222 | 0.613 | 0.744 |
| **$M3_{t4}$** | **0.260** | **0.632** | **0.756** |

**Table 6**
Table of results for Task 4 in the **English** partition.
[2]Baseline model is achieved fine-tuning BERT [11] pretrained model with the English samples.

| Model name | ICM-Hard | ICM-Hard Norm | F1_YES |
|---|---|---|---|
| Baseline[2] | 0.148 | 0.575 | 0.719 |
| $M1_{t4}$ | 0.343 | 0.674 | 0.784 |
| **$M2_{t4}$** | **0.381** | **0.694** | **0.797** |
| $M3_{t4}$ | 0.374 | 0.690 | 0.795 |

## A.2. Task 6 – Categorization models

Model $M1_{t6}$, which trained to predict six labels separately for each language, achieved a moderate performance. However, its F-Measure indicates room for improvement, suggesting that the model may struggle with certain categories or nuances in sexism categorization. The model trained to predict five labels $M1_{t6}$ exhibited lower performance across all metrics compared to $M1_{t6}$. This could be attributed to the removal of the `not sexist` label from the categorization task, as well as mixing its predictions with the ones from the previous model, potentially leading to loss of valuable context for accurate classification. The results can be seen in Tab. 7

For the Engish partition, as shown in Tab. 8 the first model achieved moderate performance, with comparable metrics to its counterpart on the Spanish partition. While the ICM-Hard and ICM-Hard Norm scores indicate reasonable categorization accuracy.

The single model trained on both English and translated Spanish samples demonstrated the best performance among the three models on both languages. Despite the differences in validation data composition, this model achieved significantly higher ICM-Hard and ICM-Hard Norm scores, indicating better categorization accuracy and similarity to ground truth labels. However, it's important to note that the validation data composition for M3$t6$ differed from that of M1$t6$ and M2$t6$, incorporating both English and translated Spanish samples. As such, the higher performance of M3$t6$ cannot be directly compared to the other models due to this difference in validation data composition.

**Table 7**
Table of results for Task 6 in the **Spanish** partition for our validation set.
[3]This results from $M3_{t6}$ are not comparable with the others in the table as the validation data is different (it contains English together with translated Spanish samples).

| Model | ICM-Hard | ICM-Hard Norm | F-Measure |
|---|---|---|---|
| $M1_{t6}$ | -1.431 | 0.220 | 0.292 |
| $M2_{t6}$ | -1.750 | 0.143 | 0.286 |
| **$M3_{t6}$**[3] | **-0.667** | **0.370** | **0.446** |

**Table 8**
Table of results for Task 6 in the **English** partition for our validation set.
[4]This results from $M3_{t6}$ are not comparable with the others in the table as the validation data is different (it contains English together with translated Spanish samples).

| Model | ICM-Hard | ICM-Hard Norm | F-Measure |
|---|---|---|---|
| $M1_{t6}$ | -0.511 | 0.394 | 0.417 |
| $M2_{t6}$ | -1.450 | 0.188 | 0.277 |
| **$M3_{t6}$**[4] | **1.762** | **0.867** | **0.841** |