

UMU Team at EXIST 2023: Sexism Identification and Categorisation Fine-tuning Multilingual Large Language Models

Notebook for the EXIST Lab at CLEF 2023

José Antonio García-Díaz¹, Ronghao Pan¹ and Rafael Valencia-García¹

¹Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100, Spain

Abstract

The third edition of the EXIST shared task focuses on the identification and categorisation of sexism in social networks. This edition will take place as a lab in CLEF 2023. There are two innovations in this edition. The main one is the perspective of learning with disagreements instead of classification with hard labels. The other novelty is the Source Intention task, which focuses on determining the author's intention. It distinguishes among direct messages whether the intention is to incite sexism, to report and share sexist situations suffered by women, or to judge sexist situations with the aim of condemning them. As in previous editions, there are documents in English and Spanish. Our proposal to solve all tasks in both languages is to combine sentence embeddings from several multilingual and Spanish Large Language Models with linguistic features. We achieve position 11 in the first task (sexism identification) and position 10 in the second task (source intention), both using the Soft vs. Soft paradigm for Spanish and English combined. For the third task (sexism categorisation) we achieve 18th position using the hard vs. hard paradigm.

Keywords

Sexism identification, Sexism categorisation, Source intention detection, Feature Engineering, Automatic Document Classification, Natural Language Processing

1. Introduction

In this paper we describe the UMU team's participation in the third edition of EXIST [1, 2], which focused on the identification and categorisation of sexism in social networks. Under the umbrella of anonymity that social networks provide, women suffer discrimination, abuse and other sexist behaviour. This behaviour is difficult to detect and remove automatically.

In this new edition, the organisers propose the latest challenges of sexism identification and categorisation with a new perspective: learning with disagreements. EXIST 2023 has three tasks. The first task is a binary classification, called sexism identification, in which the participants need to determine if a text is sexist. The second task is a multi-classification task called source


CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece


*García-Díaz et al.

†These authors contributed equally.

✉ joseantonio.garcia8@um.es (J. A. García-Díaz); ronghao.pan@um.es (R. Pan); valencia@um.es (R. Valencia-García)

ORCID 0000-0002-3651-2660 (J. A. García-Díaz); 0009-0008-7317-7145 (R. Pan); 0000-0003-2457-1791 (R. Valencia-García)

 © 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

Intention, which focuses on determining the author’s intention by distinguishing between (1) direct messages, if the intention is to write a sexist message; (2) reported, if the intention is to report a sexist situation; and (3) judged, if the intention is to judge. The third task is a multi-label classification called sexism categorisation, which focuses on identifying sexist characteristics. The labels are: (1) ideological and inequality, if the message downplays feminism or, equality between men and women; (2) stereotyping and dominance, if the message includes stereotypes about social roles; (3) objectification, if the message includes physical characteristics about beauty standards or hyper-sexualisation.

Our research group has experience in the detection of hate speech in general [3] and misogyny in particular [4] with the compilation and evaluation of several corpora in Spanish. In English, however, our experience is more limited, since we have participated in the previous editions of EXIST [5, 6] and other joint tasks that include English as one of their subtasks [7].

2. Dataset

The EXIST 2023 dataset followed the same methodology used in previous editions [8, 9]. The dataset includes texts in Spanish and English, and was crawled for specific expressions and terms that are commonly used to undervalue the role of women. The organisers also included seeds that are commonly used in both sexist and non-sexist contexts to avoid labelling bias. The first phase of the dataset compilation yielded more than 8 million tweets, with tweets written between 2021 and 2022. This dataset also includes other strategies and heuristics to mitigate temporal and terminological bias.

For the learning with disagreements paradigm, this edition does not include strict gold annotations, but the data from the 6 annotators separately. These annotators were selected to cover a wide range of demographic characteristics, including age range and gender. For this reason, we trained the first task as a regression task instead of a binary classification. In this sense, each comment is labelled with the number of annotators who indicated that a text is sexist. Using this approach, the training split of the EXIST 2023 shared tasks has an average of 2.733 *YES* with a standard deviation of 2.079. The final output considers a text to contain sexism if the regression model returns a value equal to or greater than 2.5. The soft labels are the output of the regression model normalised to a range of 1–10.

Table 1 shows the label distribution for subtasks 2 and 3. As can be seen, we extracted a split from the dataset provided for individual validation (in a ratio of 80-20) using label stratification. Task 2 has an important imbalance between labels, with direct sexism being the label with the most examples and judgement and reporting with a similar number of examples. For Task 3, the dataset has more balance between the characteristics. We have included the number of unknown responses.

To examine the dataset, we extract linguistic clues from UMUTextStats [10] and calculate the information gain for each label to observe the details of the language used in the corpus. In Figure 1 (Task 1) it can be seen that the documents labelled as sexist have more lexis about gender and social groups, concerning women and family groups. There is also strong offensive language and morphological features about proper nouns. Looking at the information gain for task 2 (see Figure 2), there are no relevant differences between the classes. The exception

is the use of offensive language in sexist messages labelled as *direct*. It should also be noted that messages containing moral judgements are more correlated with family social groups. Finally, regarding the categorisation of sexism (see Figure 3), we can observe: (1) a presence of adjectives in tweets labelled as ideological inequality, (2) anger and offensive language in misogyny documents but without sexual violence, (3) offensive language and lexis related to health in objectification, (4) lexis related to sex with sexual violence and (5) stereotyping and dominance.

3. System architecture

In this section we describe our pipeline for solving all three tasks of EXIST 2023.

The dataset contains texts in two languages: Spanish and English. In previous editions, we evaluated the two languages separately, but in this edition we decided to train all languages together due to the development of state-of-the-art multilingual LLMs. In a nutshell, we have fine-tuned several LLMs for each task, including multilingual models (multilingual BERT [11], MDeBERTa [12], XLM [13], and XLMTwitter), but we also have some Spanish models (BETO [14], MarIA [15], BERTIN [16], DistilBETO [17], ALBETO [17]). Linguistic features (LFs) from UMUTextStats were also used for classification [10]. It should be noted that this tool is designed for Spanish, but a subset of these features can be used for English.

First, for each LLM and subtask, we perform a hyperparameter tuning of 10 models to fine-tune them. We evaluate the learning rate, the number of epochs between 1 and 5, the batch size (8 and 16), the warm-up steps and the weight decay to adjust the learning rate in the early steps of training. The results of this process are shown in table 2. It can be observed that all models require 3 or more epochs for training and that Task 3 requires more complex models to achieve the best results for each LLM.

Once all the LLMs have been fine-tuned, we extract the classification token for each document, LLM and task. To do this, we proceed as indicated in SentenceBERT [18] and extract the [CLS] token. The result of this process is that each document is represented by a unique vector of fixed length. We use this vector to train a new multi-input neural network following a knowledge integration (KI) strategy. The training is performed by Keras in a deep neural network. This model has one input layer for each LLM and another input layer for the LFs. We also train other

Table 1
Datasets statistics for Tasks 2 (left) and 3 (right)

label	train	val	total	label	train	val	total
-	3175	2090	5265	ideological inequality	2013	1390	3403
direct	992	665	657	misogyny non sexual violence	1526	1053	2579
judgemental	296	225	521	objectification	1857	1173	3030
reported	309	206	515	sexual violence	1140	722	1862
total	4772	3186	7958	stereotyping dominance	2253	1475	3728
				unknown	114	67	181
				total	8903	5880	14783

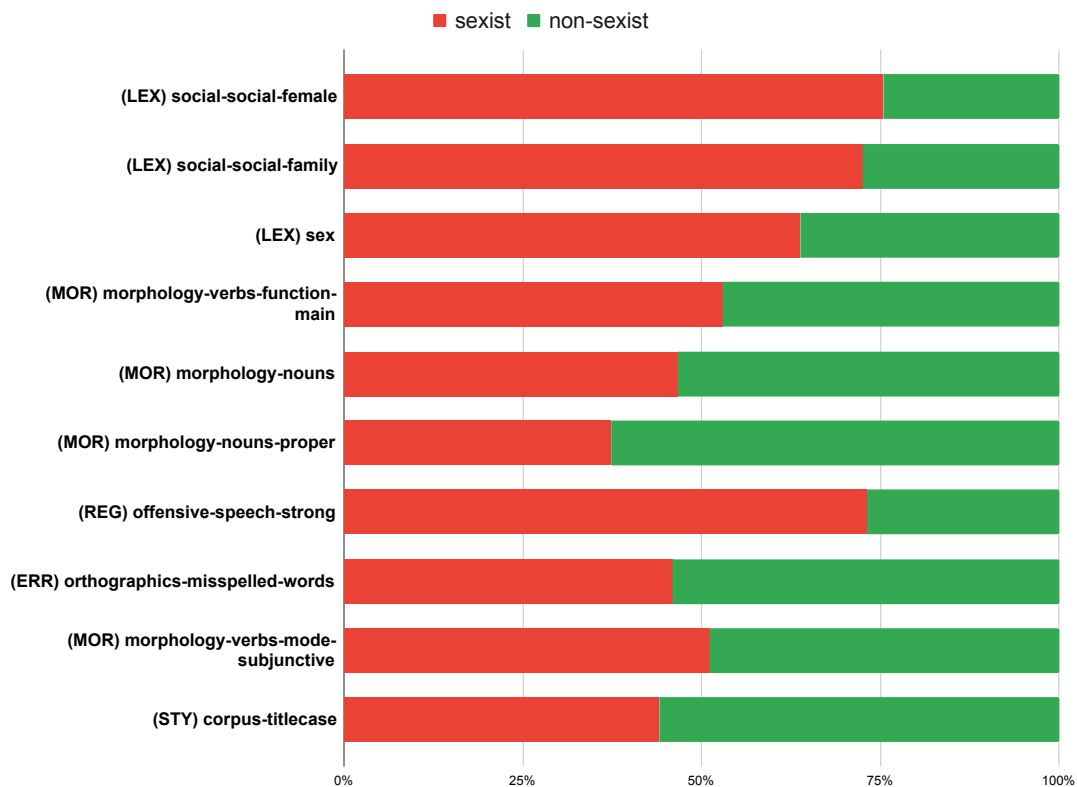


Figure 1: Information gain of the linguistic features on the labels for task 1. In this case, the labels have been transformed into a binary classification problem, where any document with three or more positive votes is considered sexist.

neural networks for each LLM and another for the LFs to build different ensemble learning strategies to combine the predictions and probabilities for each feature set. The result of this process is shown in table 3. Looking at the architecture of the neural networks, for task 1 most of the neural networks are flat, i.e. they have few hidden layers (except BETO) and a brick shape (i.e. all hidden layers have the same number of neurons). For tasks 2 and 3, however, the best results are usually achieved with deep neural networks and complex shapes such as triangles, diamonds or long funnels and a large number of neurons. For all tasks, it is usually better to use a dropout to avoid overfitting.

In addition to KI, we test another strategy for combining the features. The predictions of the models using only one LLM and the model of the LFs are combined using ensemble learning. For task 1, we only average the predictions of the model as we treat this task as a regression task. For the other two tasks, we evaluate three strategies: (1) mode of predictions, (2) averaging of probabilities, and (3) obtaining the highest probability.

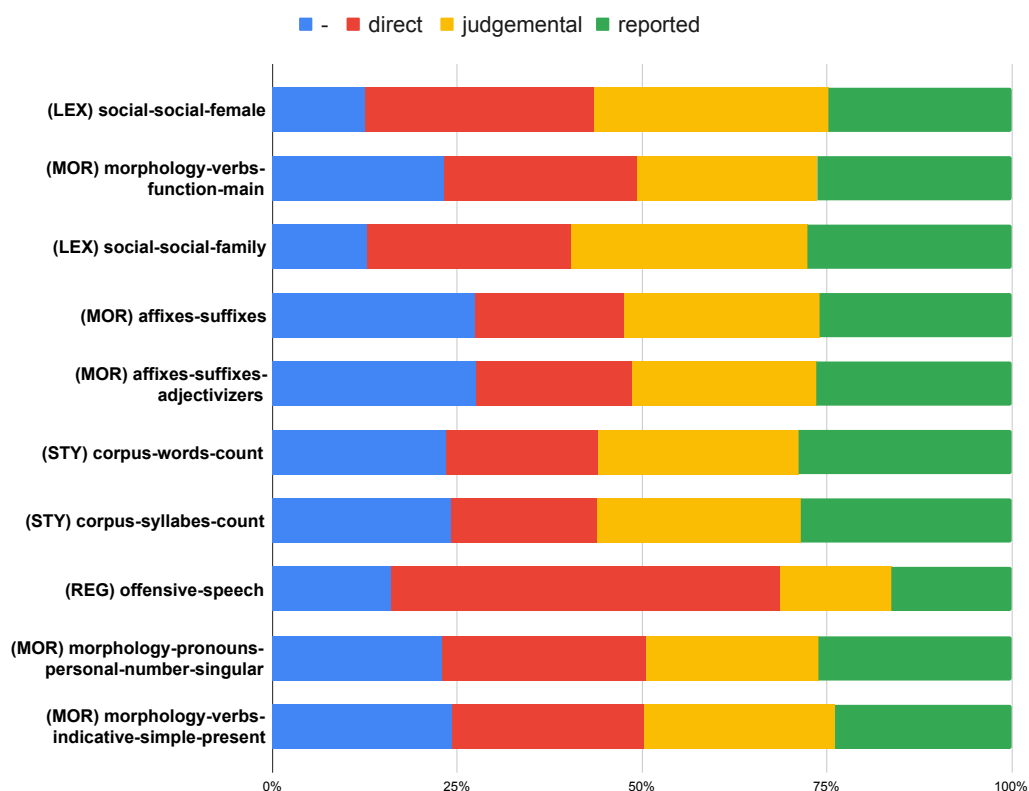


Figure 2: Information gain from linguistic features to labels for Task 2

3.1. Custom validation

For Task 1, we transform the binary classification task into a regression task to account for the disagreement between the annotators of the corpus. That is, we count how many annotators considered each text to be sexist.

We report the results for Task 1 with our own validation scheme using Explained Variance (EV), Root Mean Squared Logarithmic Error (RMSLE), Pearson R, R Square (R²), Mean Average Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). The results are reported in Table 4. The best results are achieved with Knowledge Integration (KI) for all metrics, and Ensemble Learning also achieves good results, suggesting that the combination of features is beneficial for this task.

For tasks 2 and 3, the results for custom validation split are reported in Table 5. In this sense, we train both tasks as they are, a multi-classification for task 2 and a multi-label for task 3. We report our results with the traditional scoring scheme, using the macro-average precision, recall, and f1-score. In this case, the best results are obtained with KI for Task 2 with the best recall and F1-score but with the best precision is obtained with an ensemble learning based on the mode. In the case of Task 3, the ensemble learning strategies achieve better results with

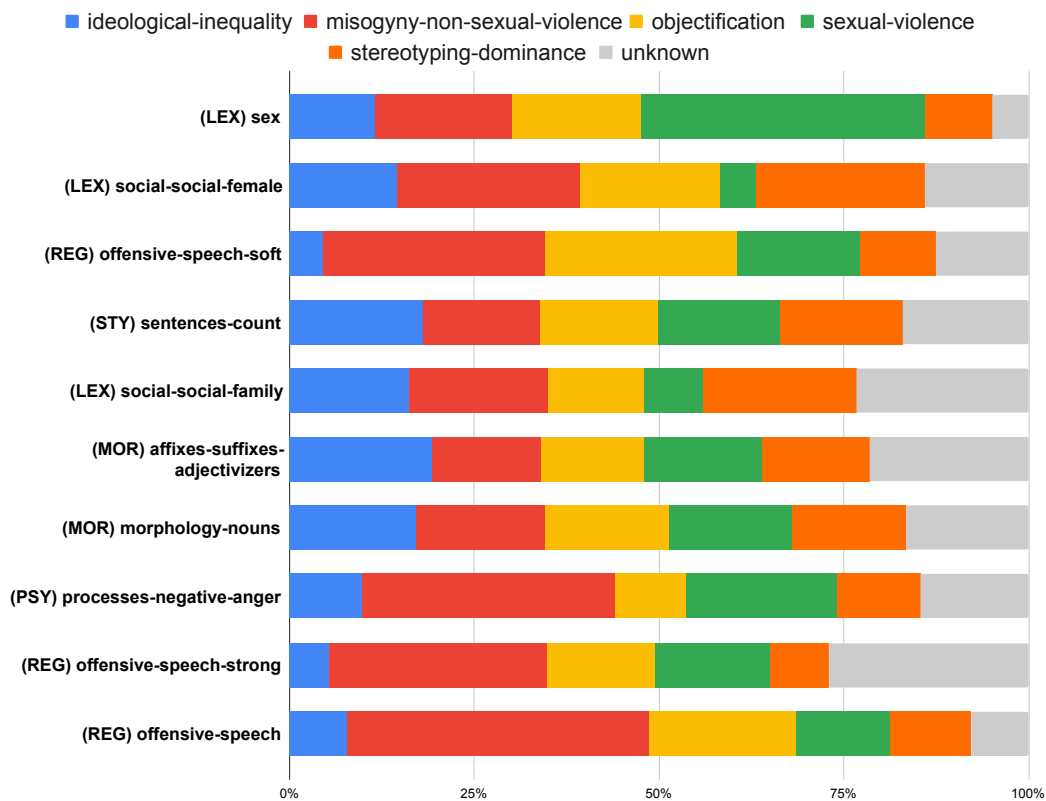


Figure 3: Information gain from linguistic features to labels for Task 3

an F1-score of 53.718% averaging probabilities and, the best recall of 82.633% with the highest probability, but the best precision is achieved by the multilingual DeBERTa model.

4. Results of the official Leaderboard

This edition of EXIST incorporates the paradigm of learning with disagreement, which is considered in both sides of the evaluation process [19]. There are three types of evaluation, known as hard vs hard, hard vs soft and soft vs soft, but all of them use the Information Contrast Measure (ICM) metric. ICM is a similarity function. It generalises the Pointwise Mutual Information (PMI) and computes the similarity to the ground truth categories. For the multi-label classification task (Task 3), the organisers have defined an extension of ICM called ICM-soft, which accepts both soft system outputs and soft ground truth assignments.

In the Hard vs. Hard scoring, the scoring compares fixed labels rather than probabilities. The organisers rely on a probabilistic threshold for each task. For task 1, the majority vote; for task 2, the class annotated by more than two annotators; and for task 3, the annotations made by more than one annotator. It should be noted that texts in which there is no majority

class are excluded from the evaluation. In addition, the ranking includes the F1 score for the positive class in task 1 and the macro-weighted F1 score for tasks 2 and 3. In the Hard vs. Soft evaluation, the evaluation compares the hard output with the probabilities assigned to each label. This uses the ICM-soft as the official scoring metric. The probabilities of the labels for each text are calculated according to the distribution of labels and annotators. Note that there are texts labelled as *unknown* and that annotations are not taken into account. Finally, the soft vs. soft evaluation compares the probabilities assigned by our systems with the probabilities assigned by the annotators. Again, the ICM-soft metric is used as a benchmark.

We have three runs for each task. We select these runs based on the best results using our

Table 2
Hyperparameter optimisation of LLMs

LLM	learning rate	epochs	batch size	warmup steps	weight decay
Task 1					
ALBETO	4.8e-05	3	16	500	0.045
BERTIN	1.8e-05	2	8	250	0.16
BETO	3.6e-05	5	16	500	0.074
DISTILBETO	2.4e-05	2	16	0	0.0038
MARIA	4.2e-05	4	8	0	0.065
MBERT	4.4e-05	3	16	500	0.13
MDEBERTA	3.9e-05	5	8	0	0.19
XLM	4e-05	4	8	1000	0.094
XLMTWITTER	4e-05	4	16	500	0.12
Task 2					
ALBETO	4.2e-05	5	16	0	0.13
BERTIN	1.6e-05	5	16	500	0.3
BETO	2.6e-05	3	8	0	0.26
DISTILBETO	1.4e-05	3	8	500	0.24
MARIA	4.9e-05	5	16	1000	0.14
MBERT	2.9e-05	4	8	1000	0.12
MDEBERTA	4.8e-05	5	8	0	0.28
XLM	1.4e-05	5	8	0	0.26
XLMTWITTER	5e-05	4	16	0	0.2
Task 3					
ALBETO	4.4e-05	5	8	250	0.18
BERTIN	4.3e-05	5	8	1000	0.2
BETO	1.8e-05	4	8	250	0.15
DISTILBETO	2.9e-05	5	8	1000	0.21
MARIA	4.6e-05	4	16	250	0.27
MBERT	4.5e-05	3	8	500	0.24
MDEBERTA	2.2e-05	5	8	1000	0.25
XLM	4.5e-05	3	8	500	0.025
XLMTWITTER	3.8e-05	3	8	250	0.011

custom validation split. For Task 1, our first run is based on knowledge integration, the second run is an ensemble learning strategy, and the third run is based on linguistic features. Our results for the first task are described in Table 6. We ranked 11th for the soft-soft scheme, out

Table 3

Results of the hyper-parameter optimisation stage using Keras of the LFs (LF), each LLM and the multi-input neural network using Knowledge Integration (KI).

feature set	shape	layers	neurons	dropout	lr	batch size	activation
Task 1							
LF	brick	1	128	0.2	0.01	32	sigmoid
ALBETO	brick	2	4	0.2	0.001	64	linear
BERTIN	brick	2	128	0.3	0.001	32	relu
BETO	brick	5	37	0.1	0.001	64	sigmoid
DISTILBETO	brick	2	4	0.1	0.01	64	linear
MARIA	long funnel	3	128	0.3	0.01	64	tanh
MBERT	brick	2	1	0.1	0.01	64	linear
MDEBERTA	funnel	8	128	False	0.001	32	sigmoid
XLM	brick	4	37	0.2	0.01	64	sigmoid
XLMTWITTER	brick	2	1	0.2	0.001	64	linear
KI	brick	4	4	0.1	0.001	64	sigmoid
Task 2							
LF	brick	1	128	0.1	0.01	512	linear
ALBETO	triangle	7	512	0.1	0.01	512	tanh
BERTIN	brick	6	128	0.1	0.01	128	elu
BETO	long funnel	5	256	0.3	0.01	512	elu
DISTILBETO	funnel	4	128	False	0.01	128	selu
MARIA	long funnel	5	128	False	0.01	256	selu
MBERT	brick	5	128	0.2	0.01	128	selu
MDEBERTA	long funnel	5	256	0.1	0.01	256	elu
XLM	triangle	5	512	0.1	0.01	256	elu
XLMTWITTER	rhombus	6	256	False	0.01	128	elu
KI	diamond	3	64	0.2	0.001	512	sigmoid
Task 3							
LF	brick	1	128	0.1	0.01	512	linear
ALBETO	triangle	7	512	0.1	0.01	512	tanh
BERTIN	brick	6	128	0.1	0.01	128	elu
BETO	long funnel	5	256	0.3	0.01	512	elu
DISTILBETO	funnel	4	128	False	0.01	128	selu
MARIA	long funnel	5	128	False	0.01	256	selu
MBERT	brick	5	128	0.2	0.01	128	selu
MDEBERTA	long funnel	5	256	0.1	0.01	256	elu
XLM	triangle	5	512	0.1	0.01	256	elu
XLMTWITTER	rhombus	6	256	False	0.01	128	elu
KI	diamond	3	64	0.2	0.001	512	sigmoid

of more than 50 results with the knowledge integration strategy. The rank of our runs varies according to the evaluation scheme. For example, we get better results with the ensemble learning strategy in Hard vs Hard and Hard vs Soft, but a worse rank. The third run, based

Table 4

Results with the custom validation split for subtask-1, reported as a regression task. The results are organised by feature set. The first block is the linguistic features, the second block are the LLMs, and the third and fourth blocks are the Knowledge Integration strategy and an ensemble learning based on average results respectively.

feature-set	EV	RMSLE	PEARSONR	R2	MAE	MSE	RMSE
LF	0.288	0.370	0.537	0.288	1.464	3.077	1.754
ALBETO	0.515	0.251	0.718	0.515	1.160	2.097	1.448
BETO	0.565	0.226	0.752	0.565	1.099	1.879	1.371
MBERT	0.487	0.260	0.699	0.487	1.187	2.217	1.489
MARIA	0.571	0.221	0.756	0.571	1.085	1.853	1.361
DILSTILBETO	0.524	0.250	0.724	0.524	1.161	2.059	1.435
MDEBERTA	0.558	0.230	0.747	0.558	1.112	1.912	1.383
BERTIN	0.539	0.234	0.734	0.539	1.127	1.995	1.413
XLM	0.482	0.268	0.694	0.482	1.217	2.240	1.497
XLMTWITTER	0.554	0.236	0.745	0.554	1.125	1.930	1.389
KI	0.599	0.204	0.774	0.599	1.039	1.734	1.317
EL (MEAN)	0.581	0.229	0.770	0.581	1.110	1.812	1.346

Table 5

Results with custom validation for tasks 2 and 3. The results are organised with the LFs (LF), all LLMs separately, the Knowledge Integration strategy (KI), and the three evaluated ensemble learning strategies (EL). All metrics are macro weighted

feature-set	Task 2			Task 3		
	precision	recall	f1-score	precision	recall	f1-score
LF	33.395	31.155	29.939	36.740	61.685	44.277
ALBETO	32.461	36.253	34.171	49.185	49.313	49.135
BERTIN	50.303	42.744	44.580	52.062	49.020	49.045
BETO	55.637	45.678	47.633	49.830	54.773	51.617
DILSTILBETO	49.672	44.760	46.242	50.299	50.527	49.382
MARIA	51.527	46.565	48.098	50.930	52.697	51.114
MBERT	41.059	39.481	39.737	44.477	57.825	49.913
MDEBERTA	52.533	45.767	47.808	54.002	48.575	50.576
XLM	49.597	43.910	45.366	43.494	66.196	50.201
XLMTWITTER	52.899	44.596	46.648	52.992	49.986	50.525
KI	55.932	53.245	54.356	49.999	56.565	52.552
EL (HIGHEST)	56.667	42.323	45.615	32.612	82.633	46.387
EL (MEAN)	57.878	42.252	44.495	52.385	56.616	53.718
EL (MODE)	59.281	40.998	43.015	51.355	54.345	52.542

on linguistic features, obtained the worst results. This result does not attract our attention, as the linguistic features are more limited than approaches based on fine-tuning large pre-trained models.

Table 6

Official results for Subtask 1, including soft and hard labels. Ranking is by runs

Team	Soft vs Soft		Hard vs Hard		Hard vs Soft	
	Rank	ICM-Soft	Rank	ICM-Hard	Rank	ICM-Hard
UMUTeam 1	11	0.6818	29	0.5053	23	0.1578
UMUTeam 2	23	0.4969	24	0.5083	21	0.1614
UMUTeam 3	36	-0.346	56	0.1882	58	-0.7329
baseline majority class	47	-2.3585	66	-0.4413	67	-2.3585
baseline minority class	52	-3.0717	69	-0.5742	70	-3.0717

The official results for Task 2 are given in Table 7. The first run is based on Knowledge Integration, but the second and third runs are based on individual LLMs, MarIA and multilingual DeBERTA. In this case, we get our best results with the Hard vs. Soft scheme, achieving second and third place with the individual models. These results attract our attention because MarIA is only focused on Spanish. However, in the other evaluation schemes, MarIA’s results are more limited. For example, in the Soft vs. Soft scheme we obtained our best result with the multilingual DeBERTA (run 3), but MarIA (run 2) obtained the most limited results.

Table 7

Official results for Subtask 2, including soft and hard labels. Ranking is by runs

Team	Soft vs Soft		Hard vs Hard		Hard vs Soft	
	Rank	ICM-Soft	Rank	ICM-Hard	Rank	ICM-Soft
UMUTeam 1	10	-2,5674	14	0.1409	5	-5.5369
UMUTeam 2	18	-4,0482	17	0.1409	2	-5.12
UMUTeam 3	9	-2,5405	21	-0.1349	3	-5.3093
baseline majority class	22	-5.4465	29	-0.9504	4	-5.446
baseline minority class	27	-32.9552	35	-3.1545	36	-32.9552

Finally, we report our results for the third task 3 in table 8. Note that we do not send probabilities for this task, so only the hard vs. hard and soft vs. hard results are reported. In this sense, the first run is based on ensemble learning averaging the results, the second run is based on knowledge integration, and the third run is based on ensemble learning based on mode. For both schemes, we achieve our best results with the third run based on mode based ensemble learning (rank 16 and rank 27).

5. Conclusions and further work

This working notes summarises our participation in the third shared task of EXIST, which includes the learning with disagreements paradigm and an additional multi-classification task to

determine the authors’ intentions. Our approach to solving all tasks is based on the combination of multilingual LLMs and LFs using KI and ensembles. We are satisfied with our participation, as we achieve competitive results in all tasks. As an improvement of our work, we need to perform our own validation taking into account the disagreements of the annotators, since we rely on hard metrics in order to select the best values.

Acknowledgments

This work is part of the research projects AIInFunds (PDC2021-121112-I00) and LT-SWM (TED2021-131167B-I00) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. This work is also part of the research project LaTe4PSP (PID2019-107652RB-I00/AEI/ 10.13039/501100011033) funded by MCIN/AEI/10.13039/501100011033.

References

- [1] 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: A. Arampatzis, E. Kanoulas, T. Tsirikika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Thessaloniki, Greece, 2023.
- [2] 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 (Extended Overview), in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), *Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum*, 2023.
- [3] J. A. García-Díaz, S. M. Jiménez-Zafra, M. A. García-Cumbreras, R. Valencia-García, Evaluating feature combination strategies for hate-speech detection in spanish using linguistic features and transformers, *Complex & Intelligent Systems* (2022) 1–22.
- [4] J. A. García-Díaz, M. Cánovas-García, R. Colomo-Palacios, R. Valencia-García, Detecting misogyny in spanish tweets. an approach based on linguistics features and word embeddings, *Future Generation Computer Systems* 114 (2021) 506–518. doi:j . future . 2020 . 08 . 032.

Table 8

Official results for Subtask 3, with hard labels only. Ranking is by runs

Team	Hard vs Hard		Soft vs Hard	
	Rank	ICM-Hard	Rank	ICM-Soft
UMUTeam 1	18	-0.5963	29	-35.0505
UMUTeam 2	26	-0.9727	31	-37.3056
UMUTeam 3	16	-0.5121	27	-34.5038
baseline majority class	27	-1.5984	2	-8.7089
baseline minority class	32	-3.1295	33	-46.108

- [5] J. A. García-Díaz, R. Colomo-Palacios, R. Valencia-García, Umuteam at exist 2021: Sexist language identification based on linguistic features and transformers in spanish and english, in: CEUR Workshop Proceedings, volume 2943, 2021, pp. 512–521.
- [6] J. A. García-Díaz, S. M. Jiménez-Zafra, R. Colomo-Palacios, R. Valencia-García, Umuteam at exist 2022: Knowledge integration and ensemble learning for multilingual sexism identification and categorization using linguistic features and transformers, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022), 2022.
- [7] C. Caparrós-Laiz, J. Antonio, G. Díaz, R. Valencia-García, Detecting hate speech on english and indo-aryan languages with bert and ensemble learning, in: Forum for Information Retrieval Evaluation (Working Notes)(FIRE), CEUR-WS. org, 2021, pp. 75–81.
- [8] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, Automatic classification of sexism in social networks: An empirical study on twitter data, *IEEE Access* 8 (2020) 219563–219576.
- [9] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, T. Donoso, Overview of exist 2021: Sexism identification in social networks, *Procesamiento del Lenguaje Natural* 67 (2021) 195–207.
- [10] J. A. García-Díaz, P. J. Vivancos-Vicente, A. Almela, R. Valencia-García, Umutextstats: A linguistic feature extraction tool for spanish, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 6035–6044.
- [11] J. Devlin, M. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186. URL: <https://doi.org/10.18653/v1/n19-1423>. doi:10.18653/v1/n19-1423.
- [12] P. He, J. Gao, W. Chen, DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing, *CoRR abs/2111.09543* (2021). URL: <https://arxiv.org/abs/2111.09543>. arXiv:2111.09543.
- [13] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: D. Jurafsky, J. Chai, N. Schluter, J. R. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, Association for Computational Linguistics, 2020, pp. 8440–8451. URL: <https://doi.org/10.18653/v1/2020.acl-main.747>. doi:10.18653/v1/2020.acl-main.747.
- [14] 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, pp. 1–10.
- [15] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. R. Penagos, A. Gonzalez-Agirre, M. Villegas, MarIA: Spanish language models, *Proces. del Leng. Natural* 68 (2022) 39–60. URL: <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6405>.
- [16] J. de la Rosa, E. G. Ponferrada, M. Romero, P. Villegas, P. González de Prado Salas, M. Grandury, BERTIN: efficient pre-training of a spanish language model using perplexity sampling, *Proces. del Leng. Natural* 68 (2022) 13–23. URL: <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6403>. doi:10.26342/2022-68-1.

- [17] J. Cañete, S. Donoso, F. Bravo-Marquez, A. Carvallo, V. Araujo, ALBETO and DistilBETO: Lightweight spanish language models, 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, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022, European Language Resources Association, 2022, pp. 4291–4298. URL: <https://aclanthology.org/2022.lrec-1.457>.
- [18] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, Association for Computational Linguistics, 2019, pp. 3980–3990. URL: <https://doi.org/10.18653/v1/D19-1410>. doi:10.18653/v1/D19-1410.
- [19] E. Amigó, A. Delgado, Evaluating extreme hierarchical multi-label classification, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 5809–5819. doi:10.18653/v1/2022.acl-long.399.