UMUTeam at EXIST 2022: Knowledge Integration and Ensemble Learning for Multilingual Sexism Identification and Categorization using Linguistic Features and Transformers

José Antonio García-Díaz[1], Salud María Jiménez-Zafra[2], Ricardo Colomo-Palacios[3] and Rafael Valencia-García[1]

[1]Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100, Spain
[2]Computer Science Department, SINAI, CEATIC, Universidad de Jaén, 23071, Spain
[3] Faculty of Computer Sciences, Østfold University College, Halden, Norway

Abstract

This paper presents the contribution of the UMUTeam to the second edition of the EXIST 2022 shared task at IberLEF 2022. This task deals with the identification and categorization of sexism language in English and Spanish. Specifically, two tasks were proposed. Task 1 consisting of a binary classification of sexism and Task 2 being a multi-classification task for the categorization of sexism traits. Our proposal for these tasks is based on the use of linguistic features and transformers combined using knowledge integration and ensemble learning strategies. Our team ranked 7th in Task 1 and 3rd in Task 2, achieving an accuracy of 76.47% and 67.67%, respectively.

Keywords

Sexism Identification, Feature Engineering, Negation processing, Transformers, Knowledge Integration, Ensemble learning, Natural Language Processing

1. Introduction

This work describes the participation of the UMUTeam at EXIST 2022 [1], the second edition of the sEXism Identification in Social neTworks task, organized at IberLEF workshop. This shared task focuses on the identification and categorization of sexist behaviors in social networks.

Sexism is a discriminatory attitude of those who undervalue or distinguish people based on their sex. The presence of sexist comments on social networks is very frequent and they range from explicit forms of misogyny to subtle or "friendly" expressions that can go unnoticed, making them difficult to identify, even for humans. Most of the existing works have focused on one form of sexism, misogyny, developing systems for its detection [2, 3, 4, 5, 6], providing datasets, such as the Spanish MisoCorpus-2020 [4], or providing datasets and organizing shared

IberLEF 2022, September 2022, A Coruña, Spain
joseantonio.garcia8@um.es (J. A. García-Díaz); sjzafra@ujaen.es (S. M. Jiménez-Zafra); ricardo.colomo-palacios@hiof.no (R. Colomo-Palacios); valencia@um.es (R. Valencia-García)
0000-0002-3651-2660 (J. A. García-Díaz); 0000-0003-3274-8825 (S. M. Jiménez-Zafra); 0000-0002-1555-9726 (R. Colomo-Palacios); 0000-0003-2457-1791 (R. Valencia-García)

tasks for its detection, such as the Automatic Misogyny Identification task (AMI) [7, 8], or the HatEval task [9] for the detection of hate speech against immigrants and women. However, sexism is not limited to hatred and violence towards women (misogyny), but also includes stereotyping and dominance, ideological issues, objectification, or sexual violence [10].

The aim of EXIST 2022 is to promote the development of tools in English and Spanish to detect sexism and categorize it according to the facet of the women that is undermined. Specifically, the organizers proposed two tasks:

- *Task 1: Sexism identification.* It is a binary classification task and consists of given a tweet written in English or Spanish, classify it as *SEXIST* or *NOT SEXIST*.
- *Task 2: Sexism categorization.* It is a multi-class classification task. For each tweet classified as *SEXIST* in the first task, the aim is to categorize the type of sexist in the following traits: (1) ideological and inequality, (2) stereotyping and dominance, (3) objectification, (4) sexual violence, and (5) misogyny and non-sexual violence.

We have participated in both tasks, testing different approaches based on the use of linguistic features and transformers combined using knowledge integration and ensemble learning strategies.

The rest of the paper is organized as follow. First, in Section 2, we give some insights regarding the dataset made available to the participants. Following, in Section 3, the methodology of our proposal is described. In Section 4, we show the results achieved by our team and compare them with those obtained by the rest of participants. Finally, the conclusions and future research directions are shown in Section 5.

2. Dataset

The dataset provided in the 2022 edition of the competition consists of a set of texts from Twitter and Gab written in English and Spanish that include expressions used to underestimate the role of women in our society. As training set, the complete EXIST 2021 dataset was supplied. It consists of 5,644 English and 5,701 Spanish tweets and posts. More details about the EXIST 2021 dataset can be found in the task overview [11]. We split this dataset into two subsets, train and dev, to perform our experiments and conduct parameter tuning. The distribution of these subsets by class for *Task 1: Sexism identification* and *Task 2: Sexism categorization* are presented in Table 1 and Table 2, respectively. We can observe that for Task 1, the distribution by class is similar, while for Task 2, the majority class is the non-sexist label, but the rest of the traits are similar in English and Spanish.

As test set, 1,058 tweets crawled from January 1st, 2022 to January 31st, 2022 were released in order to test the systems of the participants. These tweets were written in English and Spanish and annotated by 6 experts in sexism content, considering the balance between gender, 3 women and 3 men, to avoid gender bias in the labelling process. In this edition, the organizers decided not to make the test set public after the end of the competition, so it is not possible to provide statistics on the distribution of the test data, beyond the total per language, nor to analyze it.

Table 1
Corpus statistics for Task 1: Sexism identification. Sexist (S), Non-Sexist (NS) and Total (T)

Split	English			Spanish			Total		
	S	NS	T	S	NS	T	S	NS	T
Train	1676	1710	2794	1718	1702	2864	3394	3412	6806
Dev	1118	1140	2850	1146	1135	2837	2264	2275	4539
Test	-	-	526	-	-	532	-	-	1058
Total	2794	2850	6170	2864	2837	6233	5658	5687	12403

Table 2
Corpus statistics for Task 2: Sexism categorization

Data	Class	English	Spanish	Total
Training	ideological-inequality	432	474	906
	misogyny-non-sexual-violence	287	391	678
	non-sexist	1710	1702	3412
	objectification	228	245	473
	sexual-violence	337	219	556
	stereotyping-dominance	392	389	781
Validation	ideological-inequality	287	294	581
	misogyny-non-sexual-violence	212	267	479
	non-sexist	1140	1135	2275
	objectification	178	173	351
	sexual-violence	205	156	361
	stereotyping-dominance	236	256	492
Test	-	526	532	1058
Total	-	6170	6233	12403

3. Methodology

In a nutshell, our pipeline is the following. First, we split the dataset into training and validation as depicted in Section 2. Second, we clean the dataset. Third, we extract the following features: generic linguistic features from UMUTextStats and fine-grained negation features (LF), and sentence embeddings from FastText (SE), BERT (BF), and RoBERTa (RF). Forth, we train several neural network models using the features separately. Fifth, we evaluate two strategies for combining the strengths of each feature set: knowledge integration and ensemble learning. Finally, we obtain our final run using the best strategies that achieved better results with our custom validation split. Next, each step is described in detail.

3.1. Data-cleaning

During this step, we pre-process the documents removing punctuation marks, hyperlinks, and emojis. Besides, misspellings are fixed using the PSPELL library (http://aspell.net/) and

Table 3
Hyper-parameters of each feature set trained separately and combined using knowledge integration. The hyperparameters are the shape of the neural network, their number of hidden layer and neurons, the dropout rate, the learning rate and the activation function

	shape	layers	neurons	dropout	lr	activation
			English			
LF	funnel	6	256	.2	0.001	tanh
SE	brick	2	16	.2	0.010	sigmoid
BF	brick	2	4	.2	0.001	sigmoid
RF	brick	1	4	.3	0.001	relu
KI	long funnel	6	48	-	0.010	elu
			Spanish			
LF	brick	1	64	.2	0.001	sigmoid
SE	brick	1	48	.3	0.010	sigmoid
BF	brick	4	256	.2	0.010	elu
RF	brick	1	256	-	0.010	linear
KI	brick	5	512	.3	0.001	sigmoid

acronyms and abbreviations are expanded. Finally, all texts are encoded into their lowercase forms. The normalized version of the documents is used to extract the features based on sentence embeddings and certain linguistic features. However, the uncleaned version of the texts is used to obtain certain linguistic features regarding correction and style and stylometry.

3.2. Feature extraction

For the linguistic features we combine UMUTextStats [12, 13] with fine-grain negation [14, 15]. UMUTextStats extracts 389 features organised in (1) phonetics, (2) morphosyntax, (3) correction and style, (4) semantics, (5) pragmatics and figurative language, (6) stylometry, (7) lexis, (8) psycho linguistic processes, (9), and (10) social media jargon. The fine-grain negation features include simple cues (e.g., “no”/ *not*), continuous cues (e.g. “en mi vida”/ *in my life*) and discontinuous cues (e.g. “ni...ni”/ *nor...nor*).

For the non-contextual sentence embeddings (SE) we rely on FastText for English [16] and Spanish [17]. For the contextual sentence embeddings we rely on two models based on transformers: BERT (BF) and RoBERTa (RF) for English [18, 19] and Spanish [20, 21].

The sentence embeddings from BERT and RoBERTa are the value of the `[CLS]` token (similarly as described in [22]). Before this, we apply a fine-tuning approach and a hyper-parameter optimization stage using RayTune [23]. Specifically, we evaluate for each language and task 10 models with Tree of Parzen Estimators (TPE) [24]. TPE selects the next hyper-parameter combination using Bayesian reasoning and the expected improvement. During the hyper-parameter optimization stage we evaluate the (1) weight decay, (2) the batch size, (3) the warm-up speed, (4) the number of epochs, and (5) the learning rate.

Table 4
Results for the first task with our custom validation split combining the features using Knowledge Integration and four strategies for Ensemble Learning

	English			Spanish		
	precision	recall	f1-score	precision	recall	f1-score
Knowledge Integration						
non-sexist	84.995	75.526	79.981	85.266	79.031	82.030
sexist	77.590	86.404	81.760	80.635	86.475	83.453
macro avg	81.293	80.965	**80.871**	82.950	82.753	82.741
weighted avg	81.329	80.912	80.862	82.939	82.771	82.745
Ensemble learning: mode						
non-sexist	80.492	80.351	80.421	78.155	85.110	81.485
sexist	80.000	80.143	80.071	83.828	76.440	79.963
macro avg	80.246	80.247	80.246	80.992	80.775	80.724
weighted avg	80.248	80.248	80.248	81.005	80.754	80.720
Ensemble learning: weighted mode						
non-sexist	83.734	76.316	79.853	87.407	83.172	85.237
sexist	77.851	84.884	81.215	84.097	88.133	86.067
macro avg	80.793	80.600	80.534	85.752	85.652	**85.652**
weighted avg	80.821	80.558	80.528	85.744	85.664	85.654
Ensemble learning: averaging probabilities						
non-sexist	83.539	77.018	80.146	83.077	80.881	81.964
sexist	78.293	84.526	81.290	81.548	83.682	82.601
macro avg	80.916	80.772	80.718	82.312	82.282	82.283
weighted avg	80.942	80.735	80.713	82.309	82.288	82.284
Ensemble learning: highest probability						
non-sexist	94.145	35.263	51.308	93.263	54.890	69.107
sexist	59.694	97.764	74.127	68.258	96.073	79.812
accuracy	66.209	66.209	66.209	75.581	75.581	75.581
macro avg	76.920	66.514	62.718	80.761	75.482	74.459
weighted avg	77.088	66.209	62.606	80.700	75.581	74.485

3.3. Hyper-parameter optimization

In the next step of our pipeline, a neural network per feature set is obtained. For the network architecture, we evaluate only Multi-Layer Perceptrons (MLP) as all the feature sets are of fixed size. However, we distinguish among shallow and deep neural networks, according to the number of hidden layers. The shallow neural networks have only one or two hidden layers maximum and these layers have the same number of neurons in all layers. Deep neural networks are between 3 and 8 hidden layers, and the number of neurons of each layer are organized in shapes (brick, triangle, diamond, rhombus, and short and long funnel). Besides, we evaluate different activation functions, learning rates and dropout mechanisms. Table 3 reports the best

Table 5
Results for the second task with our custom validation split using Knowledge Integration and four Ensemble Learning strategies

	English			Spanish		
	precision	recall	f1-score	precision	recall	f1-score
	Knowledge Integration					
ideological-inequality	65.580	64.875	65.225	72.157	61.333	66.306
misogyny-non-sexual-violence	44.340	45.631	44.976	59.350	56.154	57.708
non-sexist	78.751	79.649	79.198	76.989	86.960	81.671
objectification	49.133	46.961	48.023	55.147	45.181	49.669
sexual-violence	61.275	60.386	60.827	70.690	74.096	72.353
stereotyping-dominance	55.000	53.878	54.433	67.553	50.000	57.466
macro avg	59.013	58.563	**58.780**	66.981	62.287	64.196
weighted avg	67.431	67.538	67.479	71.244	71.986	71.217
	Ensemble learning: mode					
ideological-inequality	68.127	61.290	64.528	68.013	67.333	67.672
misogyny-non-sexual-violence	44.172	34.951	39.024	56.940	61.538	59.150
non-sexist	69.274	87.018	77.138	74.923	85.551	79.885
objectification	57.273	34.807	43.299	56.589	43.976	49.492
sexual-violence	58.904	41.546	48.725	76.667	55.422	64.336
stereotyping-dominance	54.487	34.694	42.394	65.823	40.945	50.485
macro avg	58.706	49.051	52.518	66.492	59.127	61.836
weighted avg	63.325	65.058	63.016	69.744	70.232	69.298
	Ensemble learning: weighted mode					
ideological-inequality	68.699	60.573	64.381	71.269	63.667	67.254
misogyny-non-sexual-violence	44.172	34.951	39.024	58.029	61.154	59.551
non-sexist	74.409	82.895	78.423	77.409	84.229	80.675
objectification	55.000	42.541	47.975	52.667	47.590	50.000
sexual-violence	58.373	58.937	58.654	72.973	65.060	68.790
stereotyping-dominance	52.609	49.388	50.947	60.194	48.819	53.913
macro avg	58.877	54.881	56.567	65.423	61.753	63.364
weighted avg	65.554	66.696	65.859	70.352	70.890	70.425
	Ensemble learning: averaging probabilities					
ideological-inequality	66.023	61.290	63.569	70.980	60.333	65.225
misogyny-non-sexual-violence	44.086	39.806	41.837	57.087	55.769	56.420
non-sexist	77.200	80.789	78.954	74.558	85.463	79.639
objectification	54.605	45.856	49.850	51.370	45.181	48.077
sexual-violence	57.727	61.353	59.485	72.917	63.253	67.742
stereotyping-dominance	53.226	53.878	53.550	62.983	44.882	52.414
macro avg	58.811	57.162	57.874	64.982	59.147	61.586
weighted avg	66.601	67.139	66.793	69.000	69.706	68.902
	Ensemble learning: highest probability					
ideological-inequality	67.300	63.441	65.314	76.892	65.646	70.826
misogyny-non-sexual-violence	43.284	42.233	42.752	64.542	60.674	62.548
non-sexist	77.797	80.526	79.138	76.420	86.520	81.157
objectification	53.165	46.409	49.558	61.850	61.850	61.850
sexual-violence	57.746	59.420	58.571	68.966	64.103	66.445
stereotyping-dominance	52.263	51.837	52.049	76.136	52.344	62.037
macro avg	58.592	57.311	57.897	70.801	65.189	**67.477**
weighted avg	66.768	67.139	66.914	73.444	73.564	73.031

hyper-parameter combination for each feature set and the knowledge integration strategy for Spanish and English.

3.4. Model integration

We evaluate two strategies for combining the strengths of each feature set: knowledge integration and ensemble learning. On the one hand, knowledge integration consists of training a new neural network with multiple inputs. Then, each input is fed to its own hidden layers and then combined in new hidden layers until the final prediction. On the other hand, ensemble learning consists of generating the final predictions based on the predictions of the neural networks trained with each feature set separately. For this, we evaluate four strategies: (1) mode, (2) weighted mode, (3) averaging probabilities, and (4) highest probability.

We report the results with the custom validation split in Table 4 for the first task and in Table 5 for the second task. We can observe that the knowledge integration strategy achieves the best result for English, and the ensemble learning with the weighted mode for Spanish in the first task. However, the results are similar with all the strategies, both in terms of precision and recall. The knowledge integration strategy achieves better results in English and Spanish in the second task. Besides, the weighted mode strategy achieves less performance than averaging the predictions. Besides, we can observe that the highest probability strategy achieves the best result in the sexism categorization task.

4. Results

This section presents the results of our participation in *Task 1: Sexism identification* and *Task 2: Sexism categorization*. The organizers used the Evaluation Framework EvALL [25] to evaluate the performance of the approaches proposed by the participants. They selected Accuracy for ranking the systems in Task 1, while the macro-averaged F1-score was used for Task 2. Each participant could submit 3 runs. Table 6 shows the approach used by our team in each of the runs. These strategies are selected based on the results achieved with our custom validation split.

Table 6
Approaches tested in each of the runs of the UMUTeam

Run	Approach
UMU_1	Knowledge Integration
UMU_2	Ensemble learning: weighted mode
UMU_3	Ensemble learning: averaging the predictions

4.1. Task 1: Sexism identification

The performance of the three runs submitted for Task 1 is shown in Table 7. For the binary classification task (sexism, non-sexism), the approach that provided the best result was the one based on combining the fine-tuned embeddings from BETO and from RoBERTa with linguistic features from UMUTextStats and fine-grain negation by means of knowledge integration.

Table 7
Results of the three runs of the UMUTeam for Task 1: sexism identification

Rank	Team	Accuracy	F1-score
16	UMU_1	0.7647	0.7642
19	UMU_3	0.7637	0.7628
20	UMU_2	0.7618	0.7605

Regarding the position reached in the competition, taking into account the best run of each team, the UMUTeam (our team) obtained the 7th position in the first task, as it is shown in Table 8. Our results are only 3 hundredths from the first position, showing the success of our proposal. Furthermore, this is an indicator of the need to explore new mechanisms to detect sexism to achieve higher accuracy.

Table 8
Comparison of the UMUTeam with the best three runs and the baselines for Task 1: sexism identification

Rank	Team	Accuracy	F1-score
1	avacaondata_1	0.7996	0.7978
2	CIMATCOLMEX_1	0.7949	0.7940
3	I2C_1	0.7883	0.7880
7	**UMU_1**	**0.7647**	**0.7642**
20	BASELINE	0.6928	0.6859
22	Majority Class	0.5444	0.3525

If we analyze the results by language, we can see that, although our team obtains a similar accuracy for Spanish (Table 9) and English (Table 10), it performs better in Spanish compared to the rest of the teams, reaching the 4th position and being only 2 hundredths away from the best position. However, in English, our team obtains the 12th position and is 7 hundredths away from the first one. This may be due to the team's experience in Spanish text classification and the use of specific tools for this language, such as the UMUTextsStats [26, 4] and a Spanish negation detector [15, 14], both developed by the team members.

Table 9
Top-5 results for Task 1 - Spanish

Rank	Team	Accuracy	F1-score
1	CIMATCOLMEX_1	0.7801	0.7801
2	multiaztertest_1	0.7744	0.7744
3	I2C_3	0.7707	0.7706
4	**UMU_3**	**0.7613**	**0.7613**
5	avacaondata_1	0.7575	0.7574

Table 10
Top-5 results for Task 1 - English

Rank	Team	Accuracy	F1-score
1	avacaondata_1	0.8422	0.8376
2	SINAI-TL_1	0.8194	0.8166
3	I2C_1	0.8137	0.8117
4	CIMATCOLMEX_3	0.8137	0.8103
5	AI-UPV_3	0.8118	0.8087
12	**UMU_1**	**0.7738**	**0.7711**

4.2. Task 2: Sexism categorization

The performance of the three runs submitted for Task 2 is shown in Table 11. For the multi-class classification task the best result was achieved with ensemble learning and weighted mode. This finding draws our attention, as the ensemble learning strategy is the one that reported the most limited results with the custom validation split. In general, the three runs provide similar results in both tasks.

Table 11
Results of the three runs of the UMUTeam for Task 2: sexism categorization

Rank	Team	Accuracy	F1-score
7	UMU_2	0.6767	0.4741
8	UMU_1	0.6730	0.4724
12	UMU_3	0.6720	0.4680

Regarding the position reached in the competition, the UMUTeam obtained the 3rd position in the second task, as it is shown in Table 12.

Table 12
Comparison of the UMUTeam with the best three runs and the baselines for Task 2: sexism categorization

Rank	Team	Accuracy	F1-score
1	avacaondata_1	0.7013	0.5106
2	ELiRF-VRAIN_3	0.7042	0.4991
3	**UMU_1**	**0.6767**	**0.4741**
16	BASELINE	0.5784	0.3420
18	Majority Class	0.5539	0.1018

If we analyze the results by language, our system obtained similar F1-score for Spanish (Table 13) and English (Table 14), but also performs better for Spanish considering that it is one thousandth of a thousandth of the best F1 value obtained in the task.

Table 13
Top-5 results for Task 2 - Spanish

Rank	Team	Accuracy	F1-score
1	ELiRF-VRAIN_3	0.6786	0.4867
2	avacaondata_1	0.656	0.4864
3	ThangCIC_2	0.656	0.4514
4	**UMU_1**	**0.6541**	**0.4855**
5	multiaztertest_1	0.6466	0.4679

Table 14
Top-5 results for Task 2 - English

Rank	Team	Accuracy	F1-score
1	avacaondata_1	0.7471	0.5337
2	ELiRF-VRAIN_2	0.7319	0.5049
3	multiaztertest_1	0.7110	0.4689
4	**UMU_2**	**0.7091**	**0.4751**
5	AI-UPV_3	0.6996	0.5133

5. Conclusions

These working notes summarizes the participation of the UMUTeam at EXIST 2022 shared task concerning misogyny identification and categorization in Spanish and English languages. Our team ranked 7th in Task 1 (misogyny identification) and 3rd in Task 2 (misogyny categorization), achieving an accuracy of 76.47% and 67.67%, respectively. For solving these challenges, we built several deep-learning classifiers separately for each challenge and language. These systems rely on multiple feature sets, including linguistic features, fine-grained negation features, and sentence embeddings from BERT, RoBERTa and FastText. Our best classifiers combined the strengths of each feature set using different strategies, such as knowledge integration and ensemble learning. We achieved our best result with knowledge integration for Task 1 whereas our best result for Task 2 was obtained with an ensemble learning based on the weighted mode.

It is worth mentioning that, as can be seen in the results obtained by language, both in Task 1 and Task 2, the levels of accuracy reached by the participants are higher for English than for Spanish, which shows the need to continue working on the development of tools and methods for this language. In addition, the results in English have not reached their maximum development, indicating that there is still room for improvement in the detection of messages that dismiss women and, specially, in categorizing the facet of women that is undermined.

As further work we will focus on the interpretability of the deep-learning models and features. One of the drawbacks we faced during this competition is that we did not know the reason why some deep learning classifiers and feature sets performed better in some cases than in others. Therefore, we will evaluate the training of new deep-learning classifiers using the linguistic and negation features but adding the classifiers which correctly classified those instances as one multi-label classification task. Thus, we expect to determine which traits can explain the differences between the sentence embeddings.

Acknowledgments

This work was supported by Project LaTe4PSP (PID2019-107652RB-I00) funded by MCIN/AEI/10.13039/501100011033, Project AllInFunds (PDC2021-121112-I00) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, Project LIVING-LANG (RTI2018-094653-B-C21) funded by MCIN/AEI/10.13039/501100011033 and by ERDF A way of making Europe, and Big Hug project (P20_00956, PAIDI 2020) and WeLee project (1380939, FEDER Andalucía 2014-2020) funded by the Andalusian Regional Government. In addition, José Antonio García-Díaz has been supported by Banco Santander and University of Murcia through the industrial doctorate programme, and Salud María Jiménez-Zafra has been partially supported by a grant from Fondo Social Europeo and Administración de la Junta de Andalucía (DOC_01073).

References

[1] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, A. Mendieta-Aragón, G. Marco-Remón, M. Makeienko, M. Plaza, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2022: sexism identification in social networks, Procesamiento del Lenguaje Natural 69 (2022).

[2] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on twitter, in: Proceedings of the NAACL student research workshop, 2016, pp. 88–93.

[3] S. Frenda, B. Ghanem, M. Montes-y Gómez, P. Rosso, Online hate speech against women: Automatic identification of misogyny and sexism on twitter, Journal of Intelligent & Fuzzy Systems 36 (2019) 4743–4752.

[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. URL: http://www.sciencedirect.com/science/article/pii/S0167739X20301928. doi:10.1016/j.future.2020.08.032.

[5] F.-M. Plaza-Del-Arco, M. D. Molina-González, L. A. Ureña-López, M. T. Martín-Valdivia, Detecting misogyny and xenophobia in spanish tweets using language technologies, ACM Transactions on Internet Technology (TOIT) 20 (2020) 1–19.

[6] 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.

[7] E. Fersini, P. Rosso, M. Anzovino, Overview of the task on automatic misogyny identification at ibereval 2018., IberEval@ SEPLN 2150 (2018) 214–228.

[8] E. Fersini, D. Nozza, P. Rosso, Overview of the evalita 2018 task on automatic misogyny identification (ami), EVALITA Evaluation of NLP and Speech Tools for Italian 12 (2018) 59.

[9] V. Basile, C. Bosco, E. Fersini, N. Debora, V. Patti, F. M. R. Pardo, P. Rosso, M. Sanguinetti, et al., Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and

women in twitter, in: 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 54–63.
[10] 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.
[11] 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.
[12] J. A. García-Díaz, R. Valencia-García, Compilation and evaluation of the spanish saticorpus 2021 for satire identification using linguistic features and transformers, Complex & Intelligent Systems (2022) 1–14.
[13] J. A. García-Díaz, R. Colomo-Palacios, R. Valencia-García, Psychographic traits identification based on political ideology: An author analysis study on spanish politicians' tweets posted in 2020, Future Generation Computer Systems 130 (2022) 59–74.
[14] S. M. Jiménez-Zafra, R. Morante, E. Blanco, M. T. M. Valdivia, L. A. U. Lopez, Detecting negation cues and scopes in spanish, in: Proceedings of The 12th Language Resources and Evaluation Conference, 2020, pp. 6902–6911.
[15] S. M. Jiménez-Zafra, M. Taulé, M. T. Martín-Valdivia, L. A. Urena-López, M. A. Martí, Sfu review sp-neg: a spanish corpus annotated with negation for sentiment analysis. a typology of negation patterns, Language Resources and Evaluation 52 (2018) 533–569.
[16] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, A. Joulin, Advances in pre-training distributed word representations, CoRR abs/1712.09405 (2017). URL: http://arxiv.org/abs/1712.09405. arXiv:1712.09405.
[17] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, T. Mikolov, Learning word vectors for 157 languages, CoRR abs/1802.06893 (2018). URL: http://arxiv.org/abs/1802.06893. arXiv:1802.06893.
[18] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.
[19] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.
[20] J. Cañete, G. Chaperon, R. Fuentes, J. Pérez, Spanish pre-trained bert model and evaluation data, PML4DC at ICLR 2020 (2020).
[21] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022) 39–60. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6405.
[22] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: https://arxiv.org/abs/1908.10084.
[23] R. Liaw, E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez, I. Stoica, Tune: A research platform for distributed model selection and training, arXiv preprint arXiv:1807.05118 (2018).
[24] J. Bergstra, D. Yamins, D. Cox, Making a science of model search: Hyperparameter opti-

mization in hundreds of dimensions for vision architectures, in: International conference on machine learning, PMLR, 2013, pp. 115–123.
[25] E. Amigó, J. Carrillo-de Albornoz, M. Almagro-Cádiz, J. Gonzalo, J. Rodríguez-Vidal, F. Verdejo, Evall: Open access evaluation for information access systems, in: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2017, pp. 1301–1304.
[26] J. A. García-Díaz, M. Cánovas-García, R. Valencia-García, Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in latin america, Future Generation Computer Systems 112 (2020) 614–657. doi:10.1016/j.future.2020.06.019.