Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 229-240 recibido 05-07-2022 revisado 22-07-2022 aceptado 25-07-2022

Overview of EXIST 2022: sEXism Identification in Social neTworks

Overview de EXIST 2022: Identificación de Sexismo en Redes Sociales

Francisco Rodríguez-Sánchez[1], **Jorge Carrillo-de-Albornoz**[1], **Laura Plaza**[1],
Adrián Mendieta-Aragón[1], **Guillermo Marco-Remón**[1],
Maryna Makeienko[1], **María Plaza**[1], **Julio Gonzalo**[1],
Damiano Spina[2], **Paolo Rosso**[3]
[1]Universidad Nacional de Educación a Distancia
[2]RMIT University, Australia
[3]Universitat Politècnica de València
frodriguez.sanchez@invi.uned.es,
{jcalbornoz, lplaza, gmarco,julio}@lsi.uned.es,
{mmakeienko,amendieta}@cee.uned, maria.plaza.morales95@gmail.com,
damiano.spina@rmit.edu.au, prosso@dsic.upv.es

Abstract: The paper describes the organization, goals, and results of the sEXism Identification in Social neTworks (EXIST)2022 challenge, a shared task proposed for the second year at IberLEF. EXIST 2022 consists of two challenges: sexism identification and sexism categorization of tweets and gabs, both in Spanish and English. We have received a total of 45 runs for the sexism identification task and 29 runs for the sexism categorization task, submitted by 19 different teams. In this paper, we present the dataset, the evaluation methodology, an overview of the proposed systems, and the results obtained. The final dataset consists of more than 12,000 annotated texts from two social networks (Twitter and Gab) labelled following two different procedures: external contributors and trained experts.
Keywords: Sexism Detection, Twitter, Gab, Spanish-English.

Resumen: El artículo describe la organización, objetivos y resultados de EXIST 2022 (sEXism Identification in Social neTworks), una competición que se celebra por segundo año consecutivo en el foro IberLEF. EXIST 2022 consta de dos tareas: detección de sexismo y categorización de sexismo de tweets y gabs, tanto en español como en inglés. Hemos recibido un total de 45 ejecuciones para la tarea de detección de sexismo y 29 ejecuciones para la tarea de categorización de sexismo, enviadas por 19 equipos diferentes. En el presente artículo, presentamos el conjunto de datos, la metodología de evaluación, una descripción general de los sistemas propuestos y los resultados obtenidos. El conjunto final de datos consta de más de 12.000 textos anotados de dos redes sociales (Twitter y Gab) etiquetados siguiendo dos procedimientos diferentes: colaboradores externos y expertos en el dominio.
Palabras clave: Detección de Sexismo, Twitter, Gab, Español-Inglés.

1 Introduction

The Oxford English Dictionary defines sexism as "prejudice, stereotyping or discrimination, typically against women, on the basis of sex". As stated in (Rodríguez-Sánchez, Carrillo-de Albornoz, and Plaza, 2020), sexism is frequently found in many forms in social networks, includes a wide range of behaviours (such as stereotyping, ideological issues, sexual violence, etc.) (Donoso-Vázquez and Rebollo-Catalán, 2018; Manne, 2017) and may be expressed in different forms (direct, indirect, descriptive, reported, etc.) (Mills, 2008; Chiril et al., 2020). Subtle forms of sexism are particularly dangerous as they can go unnoticed, and affect women in many facets of their lives (Swim et al., 2001; Berg, 2006).

However, research on sexism in online platforms has focused on detecting violent

ISSN 1135-5948. DOI 10.26342/2022-69-20

sexism and hate against women (Waseem, 2016; Waseem and Hovy, 2016; Frenda et al., 2019). The previous edition of EXIST (Rodríguez-Sánchez et al., 2021) was the first attempt to automatically detect and classify sexism in a broad sense, from explicit misogyny to other subtle expressions that involve implicit sexism behaviours. Therefore, the EXIST 2022 challenge is the second shared task on sexism detection in social networks whose aim is to identify and classify sexism in a broad sense. Like its first edition in 2021, EXIST 2022 has been proposed at IberLEF. During the first edition, we received a total of 70 runs for the sexism identification task and 61 for the sexism categorization challenge, submitted by 31 different teams from 11 countries, showing the great interest of the community around sexism detection in social networks.

The EXIST 2022 shared task has been focused on the same tasks as its first edition: sexism identification and categorization. Furthermore, we proposed a new test set labelled by six experts trained to perform the task. Thus, this new edition focuses on augmenting the quality of the labels and comparing the dataset labelled by crowdsourcing to expert annotators. Moreover, balance between the genders of the annotators was ensured in order to avoid gender bias in the labeling process. Annotators of different age groups were also considered.

In this second edition of EXIST, we have received a total of 45 runs for the sexism identification task and 29 runs for the sexism categorization task, submitted by 19 different teams. Results have improved with respect to the previous edition for task 1 (sexism identification) and have remained similar for task 2 (sexism categorization), which seems to indicate that classifying sexist expressions according to the facet of women they undermine is a difficult task that requires further research.

2 Tasks

2.1 Task Description

The EXIST 2022 shared task is defined as a multilingual classification task. In particular, the EXIST challenge is organized according to two main subtasks: (i) sexism identification (task 1), which aims to identify if a message or post contains sexist content; and (ii) sexism categorization (task 2), which aims to classify the type of sexism contained in a given sexist message or post. Participants were welcome to present systems that attempt both subtasks or one of them.

Task 1 is defined as a binary classification problem, where every system should determine whether a text or message is sexist or not. It includes any type of sexist expression or related phenomena, like descriptive or reported assertions where the sexist message is a report or a description of a sexist event. In particular, we consider two labels:

- **Sexist:** the tweet or gab expresses sexist behaviours or discourses.
- **Non-Sexist:** the tweet or gab does not express any sexist behaviour or discourse.

Once a message has been classified as sexist, task 2 aims to categorize the message according to the type of sexism it encloses. The categorization has been revised by two experts in gender issues, Trinidad Donoso and Miriam Comet from the University of Barcelona, and takes into account the different aspects of women that are undermined. This task is defined as a multi-class classification problem where each sexist tweet or gab must be categorized in one of the 5 following classes:

- **Ideological and inequality:** The text discredits the feminist movement, rejects inequality between men and women, or presents men as victims of gender-based oppression.
- **Stereotyping and dominance:** The text expresses false ideas about women that suggest they are more suitable to fulfill certain roles (mother, wife, family caregiver, faithful, tender, loving, submissive, etc.), or inappropriate for certain tasks (driving, hardwork, etc), or claims that men are somehow superior to women.
- **Objectification:** The text presents women as objects apart from their dignity and personal aspects, or assumes or describes certain physical qualities that women must have in order to fulfill traditional gender roles (compliance with beauty standards, hypersexualization of female attributes, women's bodies at the disposal of men, etc.).

Text	Task 1	Task 2
Where are all the white women at?	non-sexist	non-sexist
Feminism is a war on men, but it's also a war on women. It's a war on female nature, a war on femininity.	sexist	ideological-inequality
Woman driving, be careful!	sexist	stereotyping-dominance
No offense but I've never seen an attractive african american hooker. Not a single one	sexist	objectification
I wanna touch your tits..you can't imagine what I can do on your body.	sexist	sexual-violence
I hate misogyny more than I hate women	sexist	misogyny-non-sexual-violence

Table 1: Examples of messages for each task.

- **Sexual violence:** Sexual suggestions, requests for sexual favors or harassment of a sexual nature (rape or sexual assault) are made.
- **Misogyny and non-sexual violence:** The text expresses hatred and violence towards women.

Examples of each category for both tasks are reported in Table 1.

A substantial difference between EXIST 2022 and its first edition in 2021 is that, in 2022, the test set was labelled by 6 experts trained to perform the task, and therefore annotation quality is considerably higher. EXIST 2021, on the other hand, presented a bigger test set labelled by crowdsourcing annotators using the Amazon Mechanical Turk[1] (MTurk) platform. Moreover, EXIST 2022 takes into account women and men may differ in their perception of what is sexism, and therefore the annotation group is composed of three women and three men.

2.2 Evaluation Measures and Baselines

In order to evaluate the performance of the different approaches proposed by the participants, we will use the Evaluation Framework EvALL[2] (Amigó et al., 2017; Amigó, Spina, and Carrillo-de Albornoz, 2018; Amigó et al., 2020). Within this framework, we will evaluate the system outputs as classification tasks (binary and multi-class respectively) using standard evaluation metrics, including Accuracy, Precision, Recall, and macro-averaged F1-score.

In task 1, Sexism Identification, the results of participants will be ranked using Accuracy, as the distribution between sexist and non-sexist categories is balanced. Besides, other measures will be computed, such as Precision, Recall, and F1. All metrics will be also computed by language. In particular, Accuracy has been computed as follows:

$$Accuracy = \frac{\text{number of correctly predicted instances}}{\text{number of instances}}$$

In task 2, Sexism Categorization, we will use macro-averaged F1-score to rank the system outputs. Similarly, we will compute other measures such as Precision and Recall. The F1-score was computed as follows:

$$F_1 = \frac{F_1(\text{sexism categorization})}{6}$$

where F_1(sexism categorization) is calculated as the sum of all classes (including non-sexist):

F_1(sexism categorization) = F_1(non-sexist) + F_1(ideological-inequality) + F_1(misogyny-non-sexual-violence) + F_1(objectification) + F_1(sexual-violence) + F_1(stereotyping-dominance)

We propose two different baselines so that we can establish an expected performance of the submitted runs. First, we provided a benchmark (BASELINE_) based on Support Vector Machine (linear kernel) trained on tf-idf features built from the texts unigrams. Second, a model that labels each record based on the majority class (Majority Class).

3 Dataset

The EXIST 2022 shared task employs data from Twitter and Gab in English and Spanish. In particular, this edition uses the EXIST 2021 dataset for training and a new test set labeled by experts in the task for testing. Therefore, Twitter data was used for both training and testing while Gab was only included in the EXIST 2022 training set. This way, participants can analyse whether including data from a social network without "content control" in the training phase improves the performance of their systems. In order to build the testing data for both tasks, we employed the same terms used in EXIST 2021. In particular, the final set contains 116 seed terms for Spanish and 109 for English.

To create the new test set for this edition, we used the Twitter API to search for

[1] https://www.mturk.com/

[2] www.evall.uned.es

tweets written in English or Spanish containing some of the selected keywords. The setup of our crawler implies collecting 100 tweets for each term daily. Crawling was performed during the period from the 1st of January 2022 until the 31st of January 2022, gathering 170,210 tweets for Spanish and 206,549 for English. We have removed those with less than 60 tweets to ensure an appropriate balance between seeds. The final set of seeds used contains 91 seeds for Spanish and 94 seeds for English.

Regarding the sampling process, approximately 7 tweets were randomly selected for each seed term within the period from 1st to 31st of January 2022. We randomly resampled these tweets for each language to build the final sampled set composed of 600 tweets. The whole sampling process was defined taking into account different sources of bias. In particular, we considered three main sources: seed, temporal and user bias. We tried to mitigate seed bias by including a wide range of terms that are used in both sexist and non-sexist contexts (116 terms for Spanish and 109 for English). Temporal bias between training and testing data is mitigated since there is a temporal gap of almost one year between both sets. We also checked the temporal gap between tweets for each seed to ensure that data is spread all over the period. Finally, we checked messages generated by users to ensure an appropriate balance. We also took into account this principle to split the dataset into training and test sets and removed from the test set users who were also present in the training set to avoid user bias.

The sampled data set was labelled through a majority voting approach by six expert annotators trained to perform this task. Initially, we developed an annotation guide in English and Spanish in which we provided a clear explanation of each label along with a number of examples. We presented and explained the guidelines to ensure that all experts understood the task. Then, we did an annotation experiment proposing the 6 experts to annotate the 20% of the test set obtaining a 0.387 kappa for task 1 and 0.336 for task 2. These results indicated poor agreement and were used to modify the annotation guide and revise all the problems with the annotators. We repeated the experiment and obtained a 0.57 kappa for task 1 and 0.47 for task 2 showing a moderate agreement that aligns with the fact that the sexism detection task from a broad perspective is not simple. Sexism is even more subjective than misogyny or hate speech to women thus the labeling process is harder. The final labels were selected according to the majority vote between the 6 expert annotators in all cases. In the case of a tie, the tweet/gap was discarded. The final agreement for the whole dataset was 0.589 kappa for task 1 and 0.485 for task 2. Texts with disagreement for any of the classes were removed. The final EXIST 2022 test set consists of 1058 tweets, where all texts were randomly selected from the 1200 sampled set.

We have also tried to avoid gender bias in the annotation process by employing three female annotators and three men annotators. Gender bias may lead to algorithm bias.

The training data was provided as tab-separated, according to the following fields:

- test_case: contains the string "EXIST2021" or "EXIST2022" needed for the evaluation tool EvALL.
- id: denotes a unique identifier of the text.
- source: denotes the data source; it takes values "twitter" or "gab".
- language: denotes the language of the text; it takes values "en" or "es".
- text: contains the actual text.
- task1: defines whether the text is sexist or not; it takes values "sexist" and "non-sexist".
- task2: defines the type of sexism (if applicable); it takes values as:
  - "ideological-inequality": denotes the category "Ideological and inequality";
  - "misogyny-non-sexual-violence": denotes the category "Misogyny and non-sexual violence";
  - "objectification": denotes the category "Objectification";
  - "sexual-violence": denotes the category "Sexual violence";
  - "stereotyping-dominance": denotes the category "Stereotyping and dominance";

- "non-sexist": denotes that the tweet or gab does not express any sexist behaviours or discourses.

Concerning the test data, we removed "task1" and "task2" labels from the file that was provided to the participants.

The entire EXIST dataset contains 12,403 labeled texts, 11,345 for training corresponding to EXIST 2021 and a new test set consisting of 1,058 tweets. Table 2 summarizes the description of the dataset, as well as the number of texts per class for both training and test sets, and the distribution by language.

4 Overview of the Submitted Approaches

60 groups from 14 countries signed up for EXIST 2022, 19 of them submitted runs for task 1, and 15 for task 2. In this challenge, each team had the chance to submit a maximum of 6 runs, 3 runs for each task. We received a total of 45 runs for task 1 and 29 runs for task 2.

Regarding the classification approaches, all of the participants submitted their results using some sort of transformer-based system for both tasks with the exception of one team. In particular, 18 teams used some sort of transformer architecture, of which 8 teams used BERT (Devlin et al., 2019) (or multilingual BERT - mBERT), 5 used a Spanish version of BERT called BETO (Canete et al., 2020), 4 used RoBERTa (Liu et al., 2019), 3 used DeBERTa v3 (He, Gao, and Chen, 2021), 2 used a multilingual version of RoBERTa called XLM-R (Conneau et al., 2019) or other transformer versions. Traditional machine learning methods like decision trees or Logistic Regression (LR) have been adopted by only one team. This year, none of the teams experimented with other deep learning methods (i.e. Long short-term memory networks - LSTM) or libraries. Following, we list the participants and briefly describe the approaches used by each group.

2539404758 participated in both tasks and submitted one run for each task. They fine-tuned BERT for English texts and BETO for Spanish.

AI-UPV participated in both tasks and submitted 3 runs for each task. Their system was based on an ensemble of transformer models in a single-language and multilingual configuration. In particular, they used BERT, RoBERTa, ELECTRA (Clark et al., 2020) and GPT-2 (Radford et al., 2019) as transformer models.

AIT_FHSTP participated in both tasks and submitted 3 runs for each task. They experimented with two multilingual transformer models, such as mBERT and XLM-R, and a monolingual (English) T5 model (Raffel et al., 2019). To train the models, they used a two step approach. First, unsupervised pre-training with additional data and second, supervised fine-tuning with additional as well as augmented data. For these experiments, they employed the MeTwo dataset (Rodríguez-Sánchez, Carrillo-de Albornoz, and Plaza, 2020), HatEval 2019 dataset (Basile et al., 2019) and other hate speech related datasets.

avacaondata participated in both tasks and submitted 3 runs for each task. Their best approach to the task is based on an ensemble of different transformer models with BERTweet-large (Nguyen, Vu, and Nguyen, 2020), RoBERTa and DeBERTa v3 for English, and BETO, BERTIN (De la Rosa et al., 2022), MarIA-base (Gutiérrez-Fandiño et al., 2021) and RoBERTuito (Pérez et al., 2021) for Spanish. Models were trained in two phases. First, a validation set was used for hyperparameter optimization, second, models were trained using the whole training set.

besiguenza submitted one run for each task. Their best system was based on multilingual DeBERTa v3 and used back translation techniques to augment the EXIST dataset.

CIMATCOLMEX only participated in task 1 with three different runs. Their best approach consisted in an ensemble of 10 RoBERTuito and 10 BERT models each of them is trained individually using different seeds.

CompLingKnJ only participated in task 1 with two different runs. Their best run was based on transformers, where BETO was used for Spanish messages and BERT for English. They experimented with a system based on tf-idf features and traditional machine learning techniques.

ELiRF-VRAIN participated in both tasks and submitted three runs for each task. Their system was based in a ensemble of 5 different models for Spanish (XLM-R, RoBERTa and 3 BERT models) and other 5 models for En-

	Training				Testing		
	Twitter		Gab		Twitter		
	Spanish	English	Spanish	English	Spanish	English	Total
Sexist	2599	2494	265	300	254	215	6127
Non-sexist	2612	2658	225	192	271	305	6263
Ideological-inequality	695	619	73	100	97	64	1648
Misogyny-non-sexual-violence	600	436	58	63	32	25	1214
Objectification	368	377	50	29	18	21	863
Sexual-violence	304	494	71	48	44	43	1004
Stereotyping-dominance	632	568	13	60	60	55	1388

Table 2: Dataset distribution.

glish (XLM-R, RoBERTa, BERT, hateBERT (Caselli et al., 2020) and ALBERT (Lan et al., 2019)). Furthermore, they translated all English tweets to Spanish and vice versa and masked randomly selected tokens to augment the data available.

I2C participated with 3 runs for task 1 and one run for task 2. For their best system, they translated all Spanish tweets to English and created an ensemble of 3 models: RoBERTa, BETO and SiEBERT (Hartmann et al., 2020).

LPtower submitted 3 runs for each task. For their best run, they translated all tweets to 6 languages (French, Portuguese, Italian, German, Spanish and English) and created an ensemble of 6 models, each of them trained for a different language.

multiaztertest submitted two runs for task 1 and one run for task 2. In their best run, they fine-tuned RoBERTa for English texts and BETO for Spanish.

NIT Agartala NLP Team submitted one run for each task. Their system was based on Logistic Regression trained on tf-idf features built from the texts unigrams.

shm2022 submitted two runs for task 1 and one run for task 2. They trained the multilingual model LaBSE (Feng et al., 2020) to classify both English and Spanish tweets.

SINAI only participated in task 1 with three different runs. The best run was a system based on DistilBERT (Sanh et al., 2019). They experimented with other datasets for data augmentation.

SINAI-TL only participated in task 1 with three different runs. They followed a multitask learning approach using different auxiliary tasks. BETO for Spanish and BERT for English were used as base models. Their best run used the emotion detection task as the auxiliary one by training a shared model with the Universal Joy dataset (Lamprinidis et al., 2021) for Spanish and, for English, they used a BERT model without auxiliary.

ThangCIC submitted 3 runs for each task. Their best system was based on an majority vote ensemble of 2 different models: mBERT and DeBERTa.

UMUTeam submitted 3 runs for each task. Their system combined linguistic features and state-of-the-art transformers using ensemble techniques. Their best model is based on a weighted ensemble model using transformers.

UNED-UPM submitted two runs for each task. For both tasks, they used a Multilingual Universal Sentence Encoder (Yang et al., 2019) as textual representation and computed the nearest neighbors to find the definitive class.

xaiTUD only participated in task 1 with a run. Their system was based on a combination of byte-level model ByT5 (Xue et al., 2022) with tabular modeling via TabNet (Arık and Pfister, 2021).

5 *System Results*

Tasks 1 and 2 were evaluated independently. In the following subsections, we show the results for each task and language. Teams were ranked by accuracy for task 1 and macro-averaged F1-score (F1) for task 2. However, we also report standard evaluation metrics such as Precision and Recall.

Ranking	**Team_run**	**Accuracy**	**Precision**	**Recall**	**F1**
1	task1_avacaondata_1	0.7996	0.7982	0.7975	0.7978
2	task1_CIMATCOLMEX_1	0.7949	0.7935	0.7952	0.794
3	task1_I2C_1	0.7883	0.7889	0.7912	0.788
4	task1_SINAI-TL_1	0.7845	0.7846	0.7868	0.7841
5	task1_multiaztertest_1	0.7836	0.7831	0.7853	0.783
6	task1_ELiRF-VRAIN_2	0.7694	0.7684	0.7704	0.7686
7	task1_UMU_1	0.7647	0.7647	0.7668	0.7642
8	task1_2539404758	0.7637	0.7619	0.7628	0.7623
9	task1_AI-UPV_3	0.7637	0.7652	0.7671	0.7635
10	task1_ThangCIC_3	0.7609	0.7598	0.7616	0.76
11	task1_LPtower_1	0.758	0.7561	0.7558	0.7559
12	task1_shm2022_1	0.7533	0.754	0.756	0.753
13	task1_AIT_FHSTP_3	0.7505	0.7494	0.7512	0.7496
14	task1_CompLingKnJ_1	0.7457	0.7446	0.7463	0.7448
15	task1_SINAI_1	0.7316	0.7353	0.7362	0.7315
16	task1_besiguenza_1	0.7306	0.729	0.726	0.7269
17	task1_NIT Agartala NLP Team_1	0.7098	0.7075	0.7059	0.7065
18	*task1_BASELINE*	*0.6928*	*0.6919*	*0.685*	*0.6859*
19	task1_UNED-UPM_1	0.6824	0.7131	0.6968	0.6792
20	*Majority Class*	*0.5444*	*0.5444*	*0.5*	*0.3525*
21	task1_xaiTUD_1	0.4811	0.5034	0.5026	0.46

Table 3: Results task 1 (best run).

5.1 Task 1

19 teams participated in task 1 for both English and Spanish, presenting 45 runs in total. In Table 3, the best run for each team is shown, as well as the two baselines: task1_BASELINE and Majority Class. All runs ranking is available at the task website[3].

Regarding the best run ranking, 14 teams achieved an Accuracy above the task1_BASELINE, while only 5 teams were below the baseline. For the Majority Class baseline, 16 teams achieved a higher Accuracy, whereas only 3 teams were below. The best performing team is *avacondata*, which achieved an overall F1 of 0.7996. This team exploited an ensemble of transformers models for different hyper-parameter configurations. The baseline based on majority vote was one of the worst performing solutions.

Although the official ranking considered both languages, we also presented two separate rankings by language (English and Spanish) for each task. Table 4 shows the top-10 runs for English and Table 5 for Spanish. Regarding the English results, the winning team *avacaondata* achieved the best results with an accuracy of 0.8422. Regarding the Spanish results, *CIMATCOLMEX* ranked first with an accuracy of 0.7801. They used an ensemble of 10 RoBERTuito and 10 BERT models, each of them trained individually using different seeds. The winning team *avacaondata* ranked fifth with more than 2% of difference in terms of accuracy.

As expected, transformer-based models performed better than the other techniques, since the top-10 teams are all based on these techniques. Traditional machine learning approaches did not perform well even using extra features based on external resources. Similarly, the use of external datasets has been explored by some teams with relative success. Specific-domain transformers have been successfully employed by the top-performed teams. This may suggest that transformer-based models benefit from training with data from the same source (e.g. Twitter).

It is interesting to highlight the performance difference (around 6%) between English and Spanish tasks. As we expected, transformer models perform better in English since they have been trained on corpus mainly composed of English texts. However, since Spanish is well-represented in these datasets, multilingual transformers perform very well for this language.

[3] http://nlp.uned.es/exist2022/

Ranking	**Team_run**	**Accuracy**	**Precision**	**Recall**	**F1**
1	task1_avacaondata_1.tsv_en	0.8422	0.8388	0.8365	0.8376
2	task1_SINAI-TL_1.tsv_en	0.8194	0.8148	0.8206	0.8166
3	task1_CIMATCOLMEX_3.tsv_en	0.8137	0.8087	0.8132	0.8103
4	task1_I2C_1.tsv_en	0.8137	0.8107	0.8182	0.8117
5	task1_AI-UPV_3.tsv_en	0.8118	0.807	0.8122	0.8087
6	task1_multiaztertest_2.tsv_en	0.8023	0.7981	0.794	0.7958
7	task1_ELiRF-VRAIN_3.tsv_en	0.7947	0.7893	0.7893	0.7893
8	task1_LPtower_1.csv_en	0.7852	0.7795	0.7811	0.7802
9	task1_ThangCIC_3.tsv_en	0.7852	0.7795	0.7805	0.78
10	task1_AI-UPV_1.tsv_en	0.7795	0.7789	0.7862	0.7779
11	*task1_BASELINE.tsv_en*	*0.7167*	*0.7092*	*0.7053*	*0.7068*
12	*Majority Class*	*0.5798*	*0.5798*	*0.5*	*0.367*

Table 4: Top-10 results task 1 English.

Ranking	**Team_run**	**Accuracy**	**Precision**	**Recall**	**F1**
1	task1_CIMATCOLMEX_1.tsv_es	0.7801	0.7808	0.7805	0.7801
2	task1_multiaztertest_1.tsv_es	0.7744	0.7753	0.7749	0.7744
3	task1_I2C_3.tsv_es	0.7707	0.7706	0.7707	0.7706
4	task1_I2C_1.tsv_es	0.7632	0.7652	0.7639	0.763
5	task1_UMU_3_es	0.7613	0.7614	0.7614	0.7613
6	task1_avacaondata_1.tsv_es	0.7575	0.7574	0.7574	0.7574
7	task1_ELiRF-VRAIN_1.tsv_es	0.7556	0.7608	0.7569	0.755
8	task1_ThangCIC_1.tsv_es	0.7556	0.7567	0.7549	0.755
9	task1_UMU_1_es	0.7556	0.757	0.7563	0.7556
10	task1_2539404758.tsv_es	0.7538	0.7539	0.7534	0.7535
11	*task1_BASELINE.tsv_es*	*0.6692*	*0.6747*	*0.6673*	*0.6649*
12	*Majority Class*	*0.5094*	*0.5094*	*0.5*	*0.3375*

Table 5: Top-10 results task 1 Spanish.

Ranking	**Team_run**	**Accuracy**	**Precision**	**Recall**	**F1**
1	task2_avacaondata_1	0.7013	0.5907	0.5351	0.5106
2	task2_ELiRF-VRAIN_3	0.7042	0.587	0.5057	0.4991
3	task2_UMU_2	0.6767	0.5552	0.492	0.4741
4	task2_multiaztertest_1	0.6786	0.5451	0.4826	0.4706
5	task2_ThangCIC_8	0.6626	0.5414	0.5001	0.4706
6	task2_I2C_1	0.6465	0.5255	0.518	0.47
7	task2_AIT_FHSTP_3	0.6522	0.5301	0.4999	0.4675
8	task2_LPtower_1	0.6569	0.5477	0.4748	0.4635
9	task2_AI-UPV_3	0.6267	0.519	0.5005	0.4516
10	task2_besiguenza_1	0.6285	0.4941	0.4231	0.4198
11	task2_2539404758	0.6153	0.4511	0.3939	0.3809
12	task2_UNED-UPM_1	0.5274	0.4141	0.4279	0.3708
13	*task2_BASELINE*	*0.5784*	*0.4299*	*0.3395*	*0.342*
14	task2_NIT Agartala NLP Team_1	0.6229	0.5736	0.281	0.3194
15	*Majority Class*	*0.5539*	*0.5539*	*0.1429*	*0.1018*
16	task2_shm2022_1	0.138	0.38	0.1637	0.056

Table 6: Results task 2 (best run).

Ranking	Team_run	Accuracy	Precision	Recall	F1
1	task2_avacaondata_1.tsv_en	0.7471	0.6184	0.5532	0.5337
2	task2_AI-UPV_3.tsv_en	0.6996	0.5718	0.5631	0.5133
3	task2_ELiRF-VRAIN_3.tsv_en	0.73	0.5954	0.5218	0.5084
4	task2_ThangCIC_8.tsv_en	0.6939	0.549	0.5058	0.4792
5	task2_UMU_2_en	0.7091	0.5461	0.4899	0.4751
6	task2_AI-UPV_1.tsv_en	0.673	0.5353	0.5177	0.474
7	task2_multiaztertest_1.tsv_en	0.711	0.552	0.4789	0.4689
8	task2_I2C_1.tsv_en	0.6654	0.5112	0.5193	0.4658
9	task2_AIT_FHSTP_3.tsv_en	0.6635	0.5196	0.4767	0.4545
10	task2_LPtower_1.csv_en	0.6806	0.5289	0.461	0.4515
11	*task2_BASELINE.tsv_en*	*0.5722*	*0.3836*	*0.3471*	*0.3276*
12	*Majority Class*	*0.5932*	*0.5932*	*0.1429*	*0.1064*

Table 7: Top-10 results task 2 English.

Ranking	Team_run	Accuracy	Precision	Recall	F1
1	task2_ELiRF-VRAIN_3.tsv_es	0.6786	0.5891	0.4881	0.4867
2	task2_avacaondata_1.tsv_es	0.656	0.5718	0.5169	0.4864
3	task2_UMU_1_es	0.6541	0.5985	0.5026	0.4855
4	task2_ELiRF-VRAIN_1.tsv_es	0.6767	0.5818	0.4964	0.4841
5	task2_AIT_FHSTP_3.tsv_es	0.641	0.5416	0.5215	0.4775
6	task2_I2C_1.tsv_es	0.6278	0.5432	0.5152	0.4714
7	task2_LPtower_1.csv_es	0.6335	0.5666	0.4889	0.4709
8	task2_ThangCIC_4.tsv_es	0.6316	0.5494	0.5131	0.4699
9	task2_multiaztertest_1.tsv_es	0.6466	0.5457	0.4863	0.4679
10	task2_AI-UPV_2.tsv_es	0.5545	0.4754	0.4405	0.3974
11	*task2_BASELINE.tsv_es*	*0.5846*	*0.4827*	*0.3317*	*0.3488*
12	*Majority Class*	*0.515*	*0.515*	*0.1429*	*0.0971*

Table 8: Top-10 results task 2 Spanish.

5.2 Task 2

15 teams participated in task 2 for both English and Spanish, for a total of 29 runs. In Table 6, the best run for each team is shown, as well as the two baselines. Among all the runs, 11 teams achieved an F1 above the task2_BASELINE, while only 4 teams were below it. For the Majority Class baseline, 14 teams achieved a higher F1, whereas only 1 team is below the baseline.

It is interesting to highlight the strong difference between the best and the worst systems. The best performing team for task 2 is again *avacaondata*. The worst results have been obtained by teams that employ traditional machine learning methods. Furthermore, the difference between the first and second team is more significant. This could be due to the fact that, unlike the first task, the second team does not employ a domain-adapted transformer.

Tables 7 and 8 show results for the top-10 teams in English and Spanish, respectively. Again, the task winner *avacaondata* performed better in English than in Spanish, they ranked first and second respectively. Interestingly, *ELiRF-VRAIN* performed well in Spanish by using an ensemble of 5 different models for Spanish and other 5 models for English.

In this task, the difference in performance between English and Spanish is very similar to task 1. However, it is important to notice that most participants achieved relatively low results, demonstrating the difficulty of this task and the need for further research.

6 Conclusion

In this paper, we have presented the results of the second shared task on sexism detection in a broad sense, from explicit misogyny to other subtle expressions that involve implicit sexist behaviours. The task setup provided an opportunity to test classification systems in multilingual scenarios (English and Spanish) along with different social networks (Twitter and Gab) and labelling procedures (crowdsourcing and expert annotators). Perhaps the main contribution of this new edition is its high-quality dataset, which comprises more than 1,000 tweets labelled by experts trained to perform both tasks in the competition. We think that this dataset is a useful resource for researchers in online sexism detection.

Compared to the previous EXIST edition, the runs submitted show that the problem of sexism identification can be better addressed by using transformer-based models adapted to the Twitter domain. However, the sexism categorization still remains a challenging problem. Like in the previous edition, we found out that modern transformer-based models considerably overcome traditional machine learning approaches. Overall, the results confirm that sexism detection in social networks is challenging but there is still room for improvement.

Again, the high number of participating teams at EXIST 2022 confirms the growing interest of the community around sexism detection in social networks.

Acknowledgments

This work was partially supported by the Spanish Ministry of Science and Innovation under the project "FairTransNLP: Midiendo y Cuantificando el sesgo y la justicia en sistemas de PLN"(PID2021-124361OB-C31 and PID2021-124361OB-C32). This work was also partially funded by the Spanish Ministry of Economy and Competitiveness, as part of the research cooperation project "Space for Observation of AI in Spanish" (UNED and RED.ES, M.P., ref. C039/21-OT). The work of Paolo Rosso was in the framework of the research project PROMETEO/2019/121 (DeepPattern) by the Generalitat Valenciana.

References

Amigó, E., J. Carrillo-de Albornoz, M. Almagro-Cádiz, J. Gonzalo, J. Rodríguez-Vidal, and F. Verdejo. 2017. Evall: Open access evaluation for information access systems. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1301–1304.

Amigó, E., J. Gonzalo, S. Mizzaro, and J. Carrillo-de Albornoz. 2020. An effectiveness metric for ordinal classification: Formal properties and experimental results. *arXiv preprint arXiv:2006.01245*.

Amigó, E., D. Spina, and J. Carrillo-de Albornoz. 2018. An axiomatic analysis of diversity evaluation metrics: Introducing the rank-biased utility metric. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 625–634.

Arık, S. O. and T. Pfister. 2021. Tabnet: Attentive interpretable tabular learning. In *AAAI*, volume 35, pages 6679–6687.

Basile, V., C. Bosco, E. Fersini, N. Debora, V. Patti, F. M. R. Pardo, P. Rosso, M. Sanguinetti, et al. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *13th International Workshop on Semantic Evaluation*, pages 54–63. Association for Computational Linguistics.

Berg, S. H. 2006. Everyday sexism and posttraumatic stress disorder in women: A correlational study. *Violence Against Women*, 12(10):970–988.

Canete, J., G. Chaperon, R. Fuentes, and J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. *PML4DC at ICLR*, 2020.

Caselli, T., V. Basile, J. Mitrović, and M. Granitzer. 2020. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*.

Chiril, P., V. Moriceau, F. Benamara, A. Mari, G. Origgi, and M. Coulomb-Gully. 2020. He said "who's gonna take care of your children when you are at ACL?": Reported sexist acts are not sexist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4055–4066, Online, July. Association for Computational Linguistics.

Clark, K., M.-T. Luong, Q. V. Le, and C. D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116.*

De la Rosa, J., E. G. Ponferrada, M. Romero, P. Villegas, P. G. de Prado Salas, and M. Grandury. 2022. Bertin: Efficient pre-training of a spanish language model using perplexity sampling. *Procesamiento del Lenguaje Natural*, 68:13–23.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Donoso-Vázquez, T. and Rebollo-Catalán. 2018. Violencias de género en entornos virtuales. Ediciones Octaedro.

Feng, F., Y. Yang, D. Cer, N. Arivazhagan, and W. Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852.*

Frenda, S., B. Ghanem, M. Montes-y Gómez, and P. Rosso. 2019. Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *Journal of Intelligent & Fuzzy Systems*, 36(5):4743–4752.

Gutiérrez-Fandiño, A., J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, A. Gonzalez-Agirre, C. Armentano-Oller, C. Rodriguez-Penagos, and M. Villegas. 2021. Spanish language models. *arXiv preprint arXiv:2107.07253.*

Hartmann, J., M. Heitmann, C. Siebert, and C. Schamp. 2020. More than a feeling: Accuracy and application of sentiment analysis.

He, P., J. Gao, and W. Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543.*

Lamprinidis, S., F. Bianchi, D. Hardt, and D. Hovy. 2021. Universal joy: A data set and results for classifying emotions across languages. In *The 16th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Lan, Z., M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942.*

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692.*

Manne, K. 2017. *Down girl: The logic of misogyny*. Oxford University Press.

Mills, S. 2008. *Language and sexism*. Cambridge University Press.

Nguyen, D. Q., T. Vu, and A. T. Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200.*

Pérez, J. M., D. A. Furman, L. A. Alemany, and F. Luque. 2021. Robertuito: a pre-trained language model for social media text in spanish. *arXiv preprint arXiv:2111.09453.*

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683.*

Rodríguez-Sánchez, F., J. Carrillo-de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, and T. Donoso. 2021. Overview of exist 2021: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 67:195–207.

Rodríguez-Sánchez, F., J. Carrillo-de Albornoz, and L. Plaza. 2020. Automatic classification of sexism in social networks: An empirical study on twitter data. *IEEE Access*, 8:219563–219576.

Sanh, V., L. Debut, J. Chaumond, and T. Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Swim, J., L. Hyers, L. Cohen, and M. Ferguson. 2001. Everyday sexism: Evidence for its incidence, nature, and psychological impact from three daily diary studies. *Journal of Social Issues*, 57:31 – 53.

Waseem, Z. 2016. Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas, November. Association for Computational Linguistics.

Waseem, Z. and D. Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California, June. Association for Computational Linguistics.

Xue, L., A. Barua, N. Constant, R. Al-Rfou, S. Narang, M. Kale, A. Roberts, and C. Raffel. 2022. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Yang, Y., D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. H. Abrego, S. Yuan, C. Tar, Y.-H. Sung, et al. 2019. Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint arXiv:1907.04307*.