Procesamiento del Lenguaje Natural, Revista nº 67, septiembre de 2021, pp. 195-207 recibido 02-07-2021 revisado 09-07-2021 aceptado 13-07-2021

Overview of EXIST 2021: sEXism Identification in Social neTworks

Overview de EXIST 2021: Identificación de Sexismo en Redes Sociales

Francisco Rodríguez-Sánchez[1], **Jorge Carrillo-de-Albornoz**[1], **Laura Plaza**[1], **Julio Gonzalo**[1], **Paolo Rosso**[2], **Miriam Comet**[3], **Trinidad Donoso**[3]
[1]UNED NLP & IR Group, Universidad Nacional de Educación a Distancia
[2]PRHLT Research Center, Universitat Politècnica de València
[3]Universitat de Barcelona
frodriguez.sanchez@invi.uned.es, {jcalbornoz, lplaza, julio}@lsi.uned.es,
prosso@dsic.upv.es, {miriamcomet, trinydonoso}@ub.edu

Abstract: The paper describes the organization, goals, and results of the sEXism Identification in Social neTworks (EXIST) challenge, a shared task proposed for the first time at IberLEF 2021. EXIST 2021 proposes two challenges: sexism identification and sexism categorization of tweets and gabs, both in Spanish and English. We have 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. We present the dataset, the evaluation methodology, an overview of the proposed systems, and the results obtained. The final dataset consists of more than 11,000 annotated texts from two social networks (Twitter and Gab) and its development has been supervised and monitored by experts in gender issues.
Keywords: Sexism Detection, Twitter, Gab, Spanish, English.

Resumen: El presente artículo describe la organización, objetivos y resultados de la competición sEXism Identification in Social neTworks (EXIST), una tarea propuesta por primera vez en IberLEF 2021. EXIST 2021 propone dos tareas: la identificación y la categorización de sexismo en inglés y español. Se han recibido un total de 70 runs para la tarea de identificación de sexismo y 61 para la categorización de sexismo, enviadas por 31 equipos de 11 países. En este trabajo, se presentan el dataset, la metodología de evaluación, un análisis de los sistemas propuestos por los participantes y los resultados obtenidos. El dataset final está compuesto por más de 11,000 textos anotados procedentes de dos redes sociales (Twitter y Gab) y su elaboración ha sido supervisada por expertas en temas de género.
Palabras clave: Detección de Sexismo, Twitter, Gab, Español, Inglés.

1 Introduction

The phenomenal development of web technologies has facilitated the interaction among people from many different backgrounds. With more than 4 billion people around the world now using social media each month[1], social networks are undoubtedly one of the most important ways of communicating. Although the advantages and positive effects of this global communication are obvious, the invisibility, anonymity and accessibility have made the expression of xenophobic, racist and sexist discourses easy and unpunished. The anonymity online makes users report greater hostile sexism (Fox, Cruz, and Lee, 2015) and emboldens them to engage in behaviours they are unlikely to perform face-to-face. Furthermore, the rapid spread of online information in social networks has made these behaviours extremely dangerous. In this context, inequality and discrimination against women that remain embedded in society are increasingly being replicated and spread on online platforms.

However, the detection of sexist content is still a difficult task for social media platforms.

[1] https://datareportal.com/reports/digital-2020-october-global-statshot

ISSN 1135-5948. DOI 10.26342/2021-67-17

For instance, Amnesty International published a report[2] where they describe Twitter as a "toxic place" for women. According to this report, Twitter is promoting violence and hate against people based on their gender. The report also suggests that Twitter is failing to protect women against harassment and it could harm their freedom of speech. Recently, members of the U.S. Congress asked Facebook to do more to protect women in their platform[3]. According to some lawmakers, social networks have become "the number one place" in which psychological violence is perpetrated against female parliamentarians. The seriousness of the problem, combined with the rapid dissemination of information online, the possibility of anonymity and lastingness, especially on social networks, has made these harassment behaviours extremely dangerous so that solutions are required to perform a faster and even better user generated-content moderation or to serve as a tool that helps human moderators to reduce the volume of sexist content still present in online platforms.

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). Sexism may sound "friendly": the statement "Women must be loved and respected, always treat them like a fragile glass" may seem positive, but is actually considering that women are weaker than men. Sexism may sound "funny", as it is the case of sexist jokes or humour ("You have to love women... just that... You will never understand them."). Sexism may sound "offensive" and "hateful", as in "Humiliate, expose and degrade yourself as the fucking bitch you are if you want a real man to give you attention".

However, subtle forms of sexism can be as pernicious as other forms of sexism and affect women in many facets of their lives. According to (Swim et al., 2001), non-hateful sexism can affect women's psychological wellbeing by decreasing their comfort, increasing their feelings of anger and depression, and decreasing their stated self-esteem. Similarly, (Berg, 2006) found a relationship between the experience of non-violent sexism and post-traumatic stress disorder.

Current research on sexism in online platforms is focused on detecting misogyny or hatred towards women (Waseem, 2016; Waseem and Hovy, 2016; Frenda et al., 2019). Consequently, previous works have dealt with hostile and explicit sexism, overlooking subtle or implicit expressions of sexism. An exception is the approach proposed by (Rodríguez-Sánchez, Carrillo-de Albornoz, and Plaza, 2020), where authors released the first Spanish corpus of sexist expressions in Twitter, the MeTwo dataset. They also compared Machine Learning (ML) methods to detect sexism and discussed the generalization of their approach with respect to misogyny detection systems. In line with previous hate speech research, the AMI shared task focused on the automatic identification of misogyny (hate or prejudice against women) in Twitter (Fersini, Rosso, and Anzovino, 2018). Teams were proposed to identify misogynist tweets both in Spanish and English.

Given this important social problem, the sEXism Identification in Social neTworks (EXIST) shared task has been proposed at IberLEF 2021 (Montes et al., 2021). The EXIST challenge is the first shared task on sexism detection in social networks whose aim is to identify and classify sexism in a broad sense, from explicit misogyny to other subtle expressions that involve implicit sexist behaviours. To this aim, we proposed a new categorization of sexism and built a dataset using posts from Twitter and the uncensored social network Gab.com (Gab) in English and Spanish. To collect these posts, we defined as seed terms a set of a number of popular terms, both in English and Spanish, commonly used to underestimate the role of women in our society. All these terms, as well as the sexism categorization proposed in this work, have been supervised by two experts in gender issues. The EXIST dataset incorporates any type of sexist expression or related phenomena, including descriptive or reported assertions where the sexist message

[2] https://bit.ly/2TMPAJD

[3] https://www.reuters.com/article/us-facebook-women-politics-idUSKCN2522KK

is a report or a description of a sexist behaviour. To the best of our knowledge, the EXIST dataset is the first multilingual corpus designed to identify sexism in a broad sense, from hostile to subtle and benevolent sexism.

2 Tasks

2.1 Task Description

The EXIST 2021 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.).
- **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.

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[4] (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:

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

[4] www.evall.uned.es

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.

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 $F_1(\text{sexism categorization})$ is calculated as the sum of all classes (including non-sexist):

$F_1(\text{sexism categorization}) = F_1(\text{non-sexist}) + F_1(\text{ideological-inequality}) + F_1(\text{misogyny-non-sexual-violence}) + F_1(\text{objectification}) + F_1(\text{sexual-violence}) + F_1(\text{stereotyping-dominance})$

We propose two different baselines so that we can establish an expected performance of the submitted runs. First, we provided a benchmark (Baseline_svm_tfidf) 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 2021 shared task uses data from Twitter and Gab in English and Spanish. Twitter data was used for both training and testing while Gab was only included in the EXIST test set so that it can be analysed the differences between social networks with and without “content control”. In order to provide training and testing data for both tasks, we have collected a number of popular expressions and terms, both in English and Spanish, commonly used to underestimate the role of women in our society. The terms have been extracted from different sources: (i) previous works in the area; (ii) Twitter accounts (journalist, teenagers, etc.) and hashtags used to collects phrases and expressions that women (Twitter users) have received on a day-to-day basis or experiences; (iii) expressions extracted from the Everyday Sexism Project[5]. We have also included other common hashtags and expressions that are not so frequently used in sexist contexts in order to ensure a correct distribution between sexist and non-sexist expressions. These terms were analysed and filtered by Trinidad Donoso and Miriam Comet, which examined examples of tweets extracted using these terms as seeds. The final set contains 116 seed terms for Spanish and 109 for English.

We used the Twitter API to search for tweets written in English or Spanish containing some of the selected keywords selected keywords. The setup of our crawler implies collecting 100 tweets for each term daily. Crawling was performed during the period from the 1st December 2020 until the 28th February 2021, gathering 545,717 tweets for Spanish and 662,895 for English. To ensure an appropriate balance between seeds, we have removed those with less than 60 tweets. The final set of seeds used contains 91 seeds for Spanish and 93 seeds for English. To extract posts from Gab (gabs), we downloaded the most recent Gab dump from pushshift[6] (Baumgartner et al., 2020) and searched for gabs containing the selected keywords. We gathered 1853 gabs for Spanish between the 12th September 2016 and the 12th August 2019, and 1,356,266 between 12th August 2016 and 12th August 2019 for English. In this case, we did not remove any information since we did not have many gabs for Spanish. We only could find 38 seeds for Spanish and 81 for English, introducing a considerable seed bias for this subset of the dataset.

The sampling process was different for each data source. Regarding Twitter, approximately 50 tweets (50 tweets for Spanish and 48 for English) were randomly selected for each seed term within the period from 1st to 31st of December 2020 for the training set, and 22 tweets per seed within the period from 1st to 28th February of 2021 for the test set. We randomly resampled these tweets for

[5] https://everydaysexism.com/

[6] https://files.pushshift.io/gab/

each language to build the final sampled set composed of 4500 tweets per language for the training set and 2000 tweets per language for the test set. The Gab sampling process was more complex since we did not have an uniform distribution of gabs by seed. We included all available seeds and removed gabs containing those seeds that were more numerous. Previously, we removed gabs from users with more information to mitigate user bias. The final sampled set was composed of 500 gabs for each language.

The whole sampling process was defined taking into account different sources of bias. In particular, we considered three main sources of bias: seed, temporal and user bias. We tried to mitigate seed bias by including a wide range of terms which are used in both sexist and non-sexist context (116 terms for Spanish and 109 for English). To control temporal bias, we set a temporal gap of one month (January) between the training and test data and checked the temporal gap between tweets for each seed (around 0.5 days for training and 1 day for testing) to ensure that data is spread over all the period. Finally, we checked messages generated by users to ensure an appropriate balance. In particular, around 1 message was generated per user except for gabs in Spanish where each user posted 2 gabs. 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 sets were labelled through a majority voting approach by external contributors on the Amazon Mechanical Turk[7] (MTurk) platform involving different steps. Initially, we developed along with the experts in gender issues an annotation guide in English and Spanish in which we provided a clear explanation of each label along with a number of examples. In order to evaluate the quality of the annotation guide, three experts (proposed by the gender issue experts) labeled 50 Spanish tweets obtaining a 0.58 kappa for task 1 and 0.45 for task 2. These results indicated 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 results from this experiment were used to modify the annotation guide.

	Task 1		Task 2	
	Kappa	% Agreement	Kappa	% Agreement
Spanish	0.74	0.87	0.57	0.71
English	0.62	0.83	0.49	0.72

Table 2: EXIST 2021 agreement analysis.

Then, we did an annotation experiment using MTurk. To this aim, a gold standard was created and labeled by two experts (one man and one woman with 2 years of experience in sexism classification), whose cases of disagreement were solved by a third experienced contributor. It was composed of 100 Spanish tweets and 100 English tweets. Each tweet from the gold standard was annotated by 5 crowdsourcing annotators, following the modified guidelines. Some filters were applied to select annotators: location in USA or UK for English, location in Spain, USA or Chile for Spanish, work approval rate bigger than 98% and more than 50 tasks approved. In order to determine interannotator agreement, we compared the majority vote from crowdsourcing annotators to the label selected by the experts. Table 2 shows results for this experiment. As we can see, results indicate a substantial agreement thus crowdsourcing annotators performed the task correctly.

The final labels were selected according to the majority vote between 5 crowdsourcing annotators in all cases (same filters as before were used to select annotators). Texts with 3 votes in one class for task 1 (binary problem) and with disagreement for task 2 (2 categories with 2 votes) were manually reviewed by two experts (one man and one woman) with more than two years of experience analysing sexist content in social networks. Around 10% of all posts were changed by the experts for English and 14% for Spanish. We implemented further quality control mechanisms to avoid random judgements throughout the labeling process (deviation from label distribution by annotator, time to complete the task, etc.). The final EXIST dataset consists of 6977 tweets for training and 3386 tweets for testing, where both sets are randomly selected from the 9000 and 4000 sampled sets, training, and test respectively, to ensure class balancing according to Task 1. Gab information was labeled follow-

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

ing the same process, obtaining 492 gabs in English and 490 in Spanish from the 500 labeled sets. We discarded posts in both data sources due to a number of reasons: posts written in another language, messages containing only hashtags or URLs, etc. Emojis were also removed since Mturk does not support them.

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

- test_case: contains the string "EXIST2021" 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. Once the evaluation phase was over, we shared the labels for the test set in case participants wanted to perform further tests.

The entire EXIST dataset contains 11,345 labeled texts, 5644 for English and 5701 for Spanish. Table 3 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

76 groups from 11 countries (Spain, China, Germany, India, Italy, Mexico, Austria, Switzerland, England, Greece, and Pakistan) signed up for EXIST 2021, 31 of them submitted runs for task 1, and 27 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 70 runs for task 1 and 61 runs for task 2.

Regarding the classification approaches, the majority of participants exploited transformer-based systems for both tasks. In particular, 23 teams used some sort of transformer architecture, of which 14 teams used BERT (Devlin et al., 2019) (or multilingual BERT - mBERT), 10 used a Spanish version of BERT called BETO (Canete et al., 2020), 6 used RoBERTa (Liu et al., 2019) and 5 used a multilingual version of RoBERTa called XLM-R (Conneau et al., 2019). Traditional machine learning methods like Support Vector Machines (SVM), Random Forest (RF), or Logistic Regression (LR) have been adopted by a subset of participants. Similarly, a few teams experimented with other deep learning methods (i.e. Long short-term memory networks - LSTM) and with the fastText library (Joulin et al., 2017). Following, we list the participants and briefly describe the approaches used by each group.

AI-UPV participated in both tasks and submitted one run for each task. They used an ensemble of different transformer models with BERT for English, BETO for Spanish and mBERT for multilingual models. They also implemented individual models with translation for both English and Spanish texts.

AIT_FHSTP participated in both tasks and submitted 3 runs for each task. Their best approach to the task is based on a fine-tuned XLM-R on the provided EXIST dataset, and additionally on the MeTwo dataset (Rodríguez-Sánchez, Carrillo-de Albornoz, and Plaza, 2020) and HatEval 2019 dataset (Basile et al., 2019).

Alclatos submitted 3 runs for each task.

	Training		Test				
	Spanish	English	Spanish		English		
	Twitter	Twitter	Twitter	Gab	Twitter	Gab	Total
Sexist	1741	1636	858	265	858	300	5658
Non-sexist	1800	1800	812	225	858	192	5687
Ideological-inequality	480	386	215	73	233	100	1487
Misogyny-non-sexual-violence	401	284	199	58	152	63	1157
Objectification	244	256	124	50	121	29	824
Sexual-violence	173	344	131	71	150	48	917
Stereotyping-dominance	443	366	189	13	202	60	1273

Table 3: Dataset distribution.

Their best system was based on transformers, where BETO was used for Spanish messages and RoBERTa for English.

Almuoes3 submitted one run for each task. They employed RF, LR and SVM trained on tf-idf features built from the texts unigrams. For task 1, they used an ensemble of the 3 models whereas LR was used for task 2.

Andrea_Lisa submitted one run for each task. They proposed a multilingual classification system based on mBERT for both tasks.

BilaUnwanPk1 submitted 3 runs for each task. They used the fastText library and tuned models for different n-grams configurations.

CIC submitted 3 runs for each task. They used back translation techniques to augment the dataset and applied some preprocessing steps like URLs, mails, numbers, and punctuation removal. Their best-performing algorithms were BERT, SVM, and RF.

Codec submitted one run for each task. Their system was an ensemble of 3 models for Spanish (BETO) and English (BERT) with different hyperparameter configurations for task 1. For task 2, they fine-tuned one model for each language using only the sexist texts.

Free submitted one run for each task. They trained one model for each language, RoBERTa for English and mBERT for Spanish.

GuillemGSubies submitted 3 runs for each task. Their best system used back translation from English to Spanish and vice versa. They fine-tuned BERT for English texts and BETO for Spanish.

IREL_hatespeech_group submitted 3 runs for each task. They trained a RoBERTa model using unlabeled data (Parikh et al., 2019) and experimented with feature engineering using Empath tool (Fast, Chen, and Bernstein, 2016), Hurtlex lexicon (Bassignana, Basile, and Patti, 2018), and Perspective API[8] to create tweet representations. A biLSTM and Attention layer was applied to each representation followed by a linear layer to obtain the final predictions.

LaSTUS submitted one run for each task. For both tasks, they used a multilingual BERT (mBERT) transformer model.

LHZ submitted one run for each task. They used a different transformer model for each language: DEBERTA (He et al., 2020) for English and XLM-R for Spanish. They applied an LSTM network to each representation to obtain the final predictions.

MB-Courage submitted 3 runs for each task. Their best model was based on a Graph Convolutional Network (GCN) model where nodes contain word features from BERT encoding as well as morpho-syntactic annotations. For task 1, edges indicate the word neighborhood whereas for task 2 they indicate a syntactic dependency link between words.

MessGroupELL only participated in task 1 with three different runs. Their best approach consisted in an ensemble of three classifiers for each language: XGBoost, SVM, RoBERTa for English and mBERT for Spanish. They also used the MeTwo dataset with balanced classes to augment Spanish data.

MiniTrue participated in task 1 and submitted one run. They developed a voting mechanism for the sexist label prediction taking as input the output of three different models. The first two models used BERT for English texts and BETO for Spanish, and the last model used mBERT.

Multiaztertest only participated in task 1 with three different runs. Their best run used a different transformer model for each language: BERT for English texts and BETO

[8] https://www.perspectiveapi.com/

for Spanish.

Nerin participated in both tasks and submitted 3 runs for each one. They used traditional machine learning methods like SVM, LR and AdaBoost and processed each language independently. Their best approach for task 1 was a SVM model for each language whereas LR was their best solution for task 2.

nlp_uned_team submitted 3 runs for each task. They developed a multilingual system based on pre-trained transformers and compared single-task to multi-task learning approaches. Their best approach for task 1 was a multilingual single-task model based on XLM-T (Barbieri, Anke, and Camacho-Collados, 2021) whereas a multi-task model based on XLM-R had the best results for task 2.

ORDS_CLAN submitted 3 runs for each task. They applied some preprocessing (such as the removal of URLs, mentions, or stopwords) before training one classifier per language using the fastText library.

QMUL-SDS submitted 3 runs for each task. They applied simple preprocessing and added lexical features using the Hurtlex lexicon. They used XLM-R as base model and the outputs from the last 4 hidden layers were fed into a BiLSTM layer.

Recognai submitted 2 runs for each task. Their best run used a different transformer model for each language: RoB-Tw (Barbieri et al., 2020) for English texts and BETO for Spanish.

SINAI-TL submitted 3 runs for each task. They followed a multi-task learning approach using different auxiliary tasks. BETO for Spanish and BERT for English were used as base models. Their best run used polarity classification as the auxiliary task by training a shared model with the InterTASS dataset (Martínez-Cámara et al., 2017).

Soumya submitted 3 runs for task 1 and 2 runs for task 2. They employed two machine learning techniques (RF, SVM) and one deep learning model (LSTM). Previously, they employed some extra features such as the number of slang words used in the English tweets or hashtags count. For task 1, RF was their best result and SVM for task 2.

S_exist submitted 3 runs for each task. They experimented with transformers (RoBERTa) and traditional machine learning methods such as SVM. They also tried data augmentation techniques translating all Spanish tweets to English. For both tasks, the transformer-based model achieved the best results.

Uja only participated in task 1 with one run. They used a different transformer model for each language: BERT for English texts and BETO for Spanish.

UMUTeam submitted 3 runs for each task. Their system combined linguistic features and state-of-the-art transformers using ensemble techniques. They developed their tool to create linguistic features and a different transformer model for each language: BERT for English texts and BETO for Spanish.

UNEDBiasTeam submitted 3 runs for each task. They transferred features commonly used in the task of bias and propaganda detection and studied the applicability of these features with the detection of sexism. They combined these features with machine learning methods (LR and Bi-LSTM).

Zimtstern submitted 3 runs for each task. Their system was based on mBERT and experimented with different hyperparameter configurations.

ZK submitted one run for each task. They fine-tuned 3 different models (mBERT, RoBERTa and XLM-R) and conducted soft voting on the predicted results of the three models.

ZZW submitted one run for each task. They proposed a multilingual classification system based on XLM-R.

5 System Results

Tasks 1 and 2 were evaluated independently. In the following subsections, we will show 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.

5.1 Task 1

31 teams participated in task 1 for both, English and Spanish, presenting 70 runs in total. In Table 4, the best run for each team is shown, as well as the two baselines: Baseline_svm_tfidf and Majority Class. All runs ranking is available at the task website[9].

Regarding the best run ranking, 26 teams achieved an Accuracy above the Baseline_svm_tfidf, while only 5 teams are below

[9] http://nlp.uned.es/exist2021/

Ranking	Team_run	Accuracy	Precision	Recall	F1
1	task1_AI-UPV_1	0.7804	0.7801	0.7806	0.7802
2	task1_SINAI_TL_1	0.78	0.7796	0.78	0.7797
3	task1_AIT_FHSTP_2	0.7754	0.7751	0.7756	0.7752
4	task1_multiaztertest_1	0.774	0.7741	0.7727	0.7731
5	task1_nlp_uned_team_1	0.772	0.7737	0.7696	0.7702
6	task1_free_1	0.7708	0.7712	0.7717	0.7708
7	task1_GuillemGSubies_2	0.7683	0.7693	0.7695	0.7683
8	task1_LHZ_1	0.7665	0.766	0.7661	0.7661
9	task1_zk_1	0.7647	0.7645	0.765	0.7645
10	task1_Alclatos_1	0.7637	0.7635	0.764	0.7636
11	task1_QMUL-SDS_2	0.761	0.7613	0.7618	0.7609
12	task1_s_exist_1	0.7598	0.7615	0.7614	0.7598
13	task1_MiniTrue_1	0.7553	0.7551	0.7555	0.7551
14	task1_IREL_hatespeech_group_3	0.7532	0.7536	0.754	0.7532
15	task1_zzw_1	0.7527	0.7525	0.753	0.7526
16	task1_UMUTEAM_3	0.7514	0.7537	0.7532	0.7514
17	task1_Zimtstern_3	0.7356	0.7354	0.7359	0.7354
18	task1_LaSTUS_1	0.7317	0.7321	0.7325	0.7316
19	task1_CIC_1	0.7278	0.7273	0.7269	0.727
20	task1_MessGroupELL_3	0.7237	0.7254	0.7253	0.7237
21	task1_Andrea_Lisa_1	0.7186	0.7181	0.7183	0.7182
22	task1_MB-Courage_1	0.7145	0.7154	0.7156	0.7145
23	task1_Soumya_2	0.7115	0.7147	0.7137	0.7114
24	task1_Nerin_1	0.7072	0.7129	0.7103	0.7068
25	task1_UNEDBiasTeam_2	0.7056	0.7068	0.7069	0.7056
26	task1_recognai_1	0.7044	0.7093	0.7073	0.7041
27	*Baseline_svm_tfidf*	*0.6845*	*0.6943*	*0.6888*	*0.6832*
28	task1_ BilaUnwanPk1_3	0.6763	0.6808	0.679	0.6759
29	*Majority Class*	*0.5222*	*0.5222*	*0.5*	*0.3431*
30	task1_uja_1	0.519	0.5134	0.5122	0.5035
31	task1_ORDS_CLAN_1	0.4924	0.5417	0.5114	0.3934
32	task1_almuoes3.0_1	0.4876	0.5173	0.5058	0.3979
33	task1_codec_1	0.4096	0.7725	0.3922	0.3892

Table 4: Results task 1 (best run).

the baseline. For the Majority Class baseline, 27 teams achieved a higher Accuracy, whereas only 4 teams are below the benchmark model. The best performing team is *AI-UPV*, which achieved an overall of 0.7804. In *AI-UPV* the participants exploited an ensemble of transformers models for different configurations: multilingual, language-specific, and language-specific with data augmentation (via translation). The baseline based on majority vote was one of the worst-performing solutions (29 of 33).

Although the official ranking considered both languages, we also presented two rankings by language (English and Spanish) for each task. Table 5 shows the top-10 runs for English and Table 6 for Spanish. Regarding the English results, *SINAI-TL* achieved the best results with an accuracy of 0.7772. They followed a multi-task learning approach with two base models for each language. The winning team *AI-UPV* ranked third with around 1% difference in terms of accuracy. Regarding the Spanish results, *AI-UPV* ranked first.

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 lexicons has been explored by two teams without success. Data augmentation techniques have been successfully employed by the top-performed teams. This may suggest that transformer-based models benefit from training with more data from related tasks, even if the EXIST dataset is one of the largest corpus in this area.

It is interesting to highlight the perfor-

Ranking	Team_run	Accuracy	Precision	Recall	F1
1	task1_SINAI_TL_3.tsv_en	0.7772	0.7805	0.7739	0.7747
2	task1_multiaztertest_1.tsv_en	0.7717	0.7753	0.7683	0.7691
3	task1_AI-UPV_1.tsv_en	0.7668	0.7666	0.7654	0.7657
4	task1_free_1.tsv_en	0.7668	0.7662	0.7661	0.7662
5	task1_zk_1.tsv_en	0.7663	0.7664	0.7646	0.7651
6	task1_LHZ_1.tsv_en	0.7659	0.7687	0.7626	0.7633
7	task1_nlp_uned_team_1.tsv_en	0.7609	0.7666	0.7566	0.7571
8	task1_GuillemGSubies_1_en	0.7604	0.7598	0.7599	0.7599
9	task1_AIT_FHSTP_2.tsv_en	0.7595	0.7594	0.7579	0.7583
10	task1_IREL_hatespeech_group_3.tsv_en	0.7577	0.7575	0.7563	0.7566

Table 5: Top-10 results task 1 English.

Ranking	Team_run	Accuracy	Precision	Recall	F1
1	task1_AI-UPV_1.tsv_es	0.7944	0.796	0.7958	0.7944
2	task1_AIT_FHSTP_2.tsv_es	0.7917	0.7938	0.7933	0.7916
3	task1_SINAI_TL_1.tsv_es	0.7907	0.7955	0.7931	0.7906
4	task1_nlp_uned_team_1.tsv_es	0.7833	0.7832	0.7826	0.7828
5	task1_Alclatos_1.tsv_es	0.7792	0.7808	0.7806	0.7792
6	task1_GuillemGSubies_2_es	0.7764	0.7821	0.779	0.7761
7	task1_multiaztertest_1.tsv_es	0.7764	0.7764	0.7769	0.7763
8	task1_free_1.tsv_es	0.775	0.7785	0.7771	0.7749
9	task1_s_exist_1.tsv_es	0.7704	0.777	0.7733	0.77
10	task1_LHZ_1.tsv_es	0.7671	0.7705	0.7692	0.767

Table 6: Top-10 results task 1 Spanish.

mance difference (around 2%) between Spanish and English tasks. We expected that transformers models would 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.

5.2 Task 2

27 teams participated in task 2 for both, English and Spanish, presenting 61 runs in total. In Table 7, the best run for each team is shown, as well as the two baselines. Among all the runs, 24 teams achieved an F1 above the Baseline_svm_tfidf, while only 3 teams are below the benchmark model. For the Majority Class baseline, 27 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, underlying an F1 ranging from 0.5787 to 0.1069. The best performing team for task 2 is again *AI-UPV*. The worst results have been obtained by teams that used traditional machine learning techniques such as SVM and RF to solve the task.

Tables 8 and 9 show results for the top-10 teams in English and Spanish respectively. Again, the task winner *AI-UPV* performed better in Spanish than in English, they ranked first and third respectively. Interestingly, *LHZ* performed really well in English by using DeBERTa, an enhanced version of BERT and RoBERTa models.

In this task, the difference in performance between English and Spanish increases. However, it is important to notice that most participants achieved relatively low results, showing the difficulty of this task.

6 *Conclusions*

In this paper, we have presented the results of the first 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). The runs submitted show that the problem of sexism identification can be reasonably well addressed by using transformer-based models, while the sexism categorization still remains a challenging problem. We found out that modern transformer-based models overcome

Ranking	Team_run	Accuracy	Precision	Recall	F1
1	task2_AI-UPV_1	0.6577	0.5815	0.5774	0.5787
2	task2_LHZ_1	0.6509	0.5772	0.5649	0.5706
3	task2_SINAI_TL_1	0.6527	0.5848	0.5527	0.5667
4	task2_QMUL-SDS_1	0.6426	0.5626	0.5573	0.5594
5	task2_AIT_FHSTP_2	0.6445	0.5689	0.5531	0.5589
6	task2_Alclatos_1	0.6369	0.5668	0.5535	0.5578
7	task2_IREL_hatespeech_group_3	0.6403	0.5717	0.5429	0.5556
8	task2_zk_1	0.649	0.5821	0.532	0.5521
9	task2_nlp_uned_team_3	0.6232	0.5543	0.5515	0.5509
10	task2_recognai_1	0.6243	0.5782	0.5303	0.55
11	task2_UMUTEAM_2	0.617	0.5449	0.5332	0.5362
12	task2_codec_1	0.6239	0.5377	0.5366	0.5354
13	task2_s_exist_1	0.5682	0.5126	0.5857	0.5342
14	task2_GuillemGSubies_2	0.6293	0.5412	0.5201	0.5295
15	task2_LaSTUS_1	0.612	0.5375	0.5128	0.5227
16	task2_Zimtstern_1	0.6108	0.5344	0.5122	0.5208
17	task2_Andrea_Lisa_1	0.6129	0.534	0.5114	0.5204
18	task2_zzw_1	0.6296	0.5494	0.5068	0.5192
19	task2_CIC_2	0.5527	0.4837	0.5064	0.4908
20	task2_Nerin_3	0.6046	0.5744	0.4388	0.4817
21	task2_UNEDBiasTeam_3	0.5797	0.5154	0.4484	0.4704
22	task2_MB-Courage_2	0.5946	0.5307	0.428	0.459
23	task2_Soumya_1	0.5923	0.6023	0.3999	0.4504
24	task2_free_1	0.5847	0.4792	0.4232	0.4194
25	*Baseline_svm_tfidf*	*0.5222*	*0.4315*	*0.3772*	*0.395*
26	task2_BilaUnwanPk1_1	0.5062	0.4097	0.3709	0.3788
27	task2_ORDS_CLAN_1	0.4833	0.5724	0.1747	0.1244
28	*Majority Class*	*0.4778*	*0.4778*	*0.1667*	*0.1078*
29	task2_almuoes3.0_1	0.1291	0.1043	0.1792	0.1069

Table 7: Results task 2 (best run).

Ranking	Team_run	Accuracy	Precision	Recall	F1
1	task2_LHZ_1.tsv_en	0.6336	0.5512	0.5742	0.5604
2	task2_IREL_hatespeech_group_3.tsv_en	0.6277	0.5486	0.5588	0.5531
3	task2_AI-UPV_1.tsv_en	0.6291	0.5468	0.5647	0.5507
4	task2_zk_1.tsv_en	0.6368	0.5662	0.5359	0.5432
5	task2_AIT_FHSTP_2.tsv_en	0.6187	0.539	0.5497	0.5419
6	task2_SINAI_TL_3.tsv_en	0.6255	0.5428	0.5405	0.5375
7	task2_nlp_uned_team_1.tsv_en	0.6178	0.545	0.5374	0.5371
8	task2_QMUL-SDS_1.tsv_en	0.6187	0.5306	0.5505	0.5351
9	task2_recognai_1.tsv_en	0.6123	0.5666	0.5022	0.5252
10	task2_Alclatos_1.tsv_en	0.6033	0.5241	0.5303	0.5245

Table 8: Top-10 results task 2 English.

Ranking	Team_run	Accuracy	Precision	Recall	F1
1	task2_AI-UPV_1.tsv_es	0.687	0.6286	0.5913	0.6073
2	task2_SINAI_TL_2.tsv_es	0.6815	0.6425	0.5783	0.6014
3	task2_Alclatos_1.tsv_es	0.6713	0.6211	0.5759	0.5922
4	task2_QMUL-SDS_1.tsv_es	0.6671	0.6142	0.5652	0.5843
5	task2_LHZ_1.tsv_es	0.6685	0.6202	0.5565	0.5805
6	task2_AIT_FHSTP_2.tsv_es	0.6708	0.613	0.5578	0.5761
7	task2_nlp_uned_team_3.tsv_es	0.6491	0.6089	0.5687	0.576
8	task2_codec_1.tsv_es	0.6648	0.618	0.5596	0.5759
9	task2_recognai_1.tsv_es	0.6366	0.6073	0.56	0.575
10	task2_GuillemGSubies_2_es	0.6634	0.6039	0.541	0.5646

Table 9: Top-10 results task 2 Spanish.

considerably traditional machine learning approaches. Overall, the results confirm that sexism detection in social networks is challenging, with a large room for improvement. The high number of participating teams at EXIST 2021 confirms the growing interest of the community around sexism detection in social networks. We think that the provided dataset will foster research on this topic.

Acknowledgments

This work was supported by the Spanish Ministry of Science and Innovation under the MISMIS research project on Misinformation and Miscommunication in Social Media (PGC2018-096212-B-C32).

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.

Barbieri, F., L. E. Anke, and J. Camacho-Collados. 2021. Xlm-t: A multilingual language model toolkit for twitter. *arXiv preprint arXiv:2104.12250.*

Barbieri, F., J. Camacho-Collados, L. Espinosa Anke, and L. Neves. 2020. Tweet-Eval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online, November. Association for Computational Linguistics.

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.

Bassignana, E., V. Basile, and V. Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *5th Italian Conference on Computational Linguistics, CLiC-it 2018*, volume 2253, pages 1–6. CEUR-WS.

Baumgartner, J., S. Zannettou, B. Keegan, M. Squire, and J. Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

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.

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.

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.*

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.

Fast, E., B. Chen, and M. S. Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 4647–4657.

Fersini, E., P. Rosso, and M. Anzovino. 2018. Overview of the task on automatic misogyny identification at ibereval 2018. *IberEval@ SEPLN*, 2150:214–228.

Fox, J., C. Cruz, and J. Y. Lee. 2015. Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media. *Computers in Human Behavior*, 52:436–442.

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.

He, P., X. Liu, J. Gao, and W. Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Joulin, A., E. Grave, P. Bojanowski, and T. Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics, April.

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.

Martínez-Cámara, E., M. Díaz-Galiano, M. García-Cumbreras, M. García-Vega, and J. Villena-Román. 2017. Overview of tass 2017. *IberEval@ SEPLN*, 1896:13–21.

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

Montes, M., P. Rosso, J. Gonzalo, E. Aragón, R. Agerri, M. Ángel Álvarez Carmona, E. Álvarez Mellado, J. C. de Albornoz, L. Chiruzzo, L. Freitas, H. G. Adorno, Y. Gutiérrez, S. Lima, A. Montejo-Ráez, F. M. P. de Arco, and M. Taulé. 2021. Proceedings of the iberian languages evaluation forum (iberlef 2021). In *CEUR Workshop Proceedings*.

Parikh, P., H. Abburi, P. Badjatiya, R. Krishnan, N. Chhaya, M. Gupta, and V. Varma. 2019. Multi-label categorization of accounts of sexism using a neural framework. *arXiv preprint arXiv:1910.04602*.

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.

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.