Detection and Classification of Sexism on Social Media Using Multiple Languages, Transformers, and Ensemble Models

Angel Felipe Magnossão de Paula[1], Roberto Fray da Silva[2]

[1]Universitat Politècnica de València, Spain. E-mail: adepau@doctor.upv.es
[2]Instituto de Estudos Avançados da Universidade de São Paulo, Brazil. E-mail: roberto.fray.silva@gmail.com

Abstract

Identifying and classifying sexist content in social media posts is a highly complex and relevant problem. Some characteristics such as sarcasm and multiple forms of sexism increase the difficulty of detecting and identifying this type of content. Nevertheless, it is essential to improve prediction quality to improve decision-making such as post removal, and user ban, among others. The main objective of this work is to propose a methodology and explore the use of different transformers architectures for two tasks in English and Spanish: sexism detection and sexism classification. Single-language and multilingual versions of the BERT, RoBERTa, and single-language versions of the Electra, and GPT2 architectures were evaluated on the EXIST 2022 shared task challenge at IberLEF 2022 dataset. It was observed that: (i) the use of the translation of the posts to English and then using single-language English and multilingual models present the best results; (ii) the best architectures were BERT and RoBERTa; (iii) using single-language Spanish models provided the worst results; (iv) sexism classification was more difficult than sexism detection; and (v) the use of ensembles were better than the GPT2 and Electra models, but worse than English single-language generally and multilingual models. An in-depth hyperparameters analysis was also conducted.

Keywords

Sexism detection, Sexism classification, Transformers, Deep learning

1. Introduction

Natural language processing (NLP) encompasses a group of problems and tasks that are highly relevant for machine learning and artificial intelligence [1, 2, 3]. Language-related problems are very complex and present different layers of challenges, from grammar and semantics to sarcasm detection, among others [4, 2, 3, 5, 6, 7, 8]. In a broad sense, NLP can be defined as a group of techniques and methods, traditionally using machine learning and artificial intelligence, to solve problems that have as inputs free-form text related to a specific or a group of languages[1, 2, 3, 5, 6, 7, 8].

Some of the main NLP tasks, are: (i) identification of polarity in sentences; (ii) sentiment analysis, providing a number or class that represents the sentiment in a given sentence; (iii)

IberLEF 2022, September 2022, A Coruña, Spain.
adepau@doctor.upv.es (A. F. M. d. Paula); roberto.fray.silva@gmail.com (R. F. d. Silva)
https://github.com/AngelFelipeMP (A. F. M. d. Paula); https://github.com/rfsilva1 (R. F. d. Silva)
0000-0001-8575-5012 (A. F. M. d. Paula); 0000-0002-9792-0553 (R. F. d. Silva)

specific content identification; (iv) specific content classification; among several others uses and tasks. It is crucial to observe that most works in the NLP literature focus on the English language, which has the most comprehensive number of tools, pre-trained models, and lexicons available [1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12]. As work is needed for other languages than English, this work focuses on both English and Spanish, using single-language models (also called monolingual models) and multiple language models (also called multilingual models).

In this work, our main focus is on using NLP techniques and state-of-the-art models for two main tasks: (i) sexism detection, which aims to identify if a specific sentence contains sexist content; and (ii) sexism classification, which aims to identify, for a sexist sentence, to which class it belongs (from a series of defined and widely used classes in the domain). Some important works related to sexism detection and classification: [4, 13, 14, 10, 11, 12, 6, 3, 15].

In the specific context of this work, we aim to use a purely automatic, machine learning approach to the problem of sexism detection and classification in social media posts. This is a relevant problem that has not yet been solved satisfactorily [4, 14, 16], and the proposed methodology is scalable and does not directly depend on human inputs.

It is important to observe several aspects related to the problem and the methodology proposed: (i) only pre-trained models are going to be used, as they are state-of-the-art and already contain essential information from the relevant languages; (ii) although the focus is on social media posts, the same models could be applied for other data points, such as news articles, blog posts, among others; (iii) the problems will be treated sequentially, with one group of models identifying the sexist posts, which are then fed to a second group of models, which will identify the sexism class of each post; (iv) the proposed architecture is based on different transformers architectures, as these are state-of-the-art models in several different NLP tasks [8, 17, 18, 2]; (v) this work focuses on applying the proposed methodology on the EXIST 2022 shared task challenge at IberLEF 2022 [16]; and (vi) the code used is freely available, to allow for better reproducibility of the results and to aid researchers and practitioners that want to adapt or use it.

The three main research questions of this work are: (i) What transformer architecture is better suited for the tasks of sexism detection and sexism classification?; (ii) What are the values of the main hyperparameters for the final models?; and (iii) Is the use of multilingual models better than the use of single-language models for sexism detection and sexism classification? Those questions will be answered in section 5, which contains the main results of this work.

This work is organized as follows: section 2 describes a series of important works in sexism detection and classification; section 3 describes the basic architecture of transformer models and their use for detecting and classifying sexism; section 4 describes the main steps of the methodology used; section 5 contains the main results and analysis of this work; section 6 contains a discussion of relevant topics and limitations; and section 7 concludes the work, providing also suggestions for future works.

2. Sexism detection and classification

Sexism is considered by many authors as a type of hate speech [2, 7, 19]. As with other forms of hate speech, it is considered an essential topic to be addressed to improve the quality of the

interactions on social media platforms [2, 7]. There are several accounts of sexist content on major platforms such as Twitter, motivating the development of models for better detecting and classifying social media posts.

Important examples of works that explore the impacts of sexism and hate speech on social media platforms are [6, 3, 5]. Additional information can also be obtained through the important literature reviews by [20, 7, 2]. The concepts discussed by those authors form the basis for the approach used in this work to detect and classify sexist content on social media posts.

One important observation is that sexism classification is a more complex task, due to the existence of multiple labels [14]. Additionally, the use of sarcasm and irony may increase considerably the complexity of classifying the posts into different classes. [14] also observe that the use of abbreviations, emojis, misspellings, and memes also increase the difficulty of detecting and classifying sexism in social media posts.

Several different models, techniques, and methodologies have been used for sexism detection and classification, as illustrated in the work by [2, 7, 17, 8, 20]. However, in comparison to the traditional models used (support vector machines, convolutional neural networks, long short-term neural networks, among others), the transformer models have been presenting increasingly better results [2, 7, 17, 8]. The first widely spread transformer model, BERT, has been widely used in the literature with satisfactory results [6, 21, 2]. In the last years, several variations of transformer models have been developed and successfully implemented in different NLP tasks [9, 10, 11, 12].

It is important to observe that the state-of-the-art methodologies used for the detection task could be separated into the following classes: (i) methods using only lexicons, which may be very precise in specific cases, but cannot learn; (ii) using deep learning models, that are more generic, but lack specific domain knowledge; and (iii) using lexicons and deep learning models, which incorporate aspects of both previous classes. One very relevant lexicon is Hurtlex [22], which is specific for hate speech and was used by several works, such as [6, 23].

In this work, we followed the approach by [14] and adopted the second method using pre-trained models, as an attempt to incorporate previous knowledge on the models without the need of using lexicons. This is important, as domain-specific lexicons are very costly to develop and must be constantly reviewed to incorporate new social media trends in terms of slang and concepts.

Most of the works in sexism detection and classification are conducted using the English language, both due to the quality of its resources (as most development is conducted focusing on the English language) and the availability of datasets for training and testing the models.

Nevertheless, there are several attempts of improving the quality of sexism detection and classification in other languages. These are conducted using several methodologies: (i) by using domain-specific lexicons; (ii) by using single-language models; (iii) by using multilingual models; (iv) by using translation from the specific language to English, then an English model; among others. In this work, we focus on approaches 2 to 4.

3. Transformer Models for detecting and classifying sexism

The transformer is a class of machine learning models proposed by [24] that uses as a basis the autoencoder architecture. Since its inception, several transformer architectures have been developed and applied in NLP tasks, both for single-language and multilanguage tasks [25]. Some authors consider the use of transformers in NLP a revolution in terms of quality of results and potential applications.

Its basic architecture is composed of deep encoder and decoder layers and a self-attention mechanism [25]. Several transformer models are used in this work: BERT [25], RoBERTa [26], Electra [27], and GPT2 [28]. Although they share the same basic architecture, there are several differences on their design and implementation. For more information about each of the models used, we refer the reader to the original papers: BERT [25], RoBERTa [26], Electra [27], and GPT2 [28].

The BERT model [25] is considered a language learning model that provides a structure that is general for several NLP tasks. This structure is then refined through fine-tuning procedures for specific tasks and domains. It is important to observe that, by using a vast corpus during the training phase, BERT and the other transformer models used in this work incorporate semantics features on the model, improving the quality of its predictions [25]. For an in-depth analysis of how the BERT model works, we refer the reader to [25].

It is vital to observe that the methodology proposed in this paper is based on the methodology used by the winning team of the EXIST 2021 shared task [14]. However, several changes are proposed to both improve the quality of the results, allow for easier implementation into other datasets, languages, and tasks, and to improve the replicability of the results obtained.

The main changes proposed are: (i) the use of multiple transformer architectures, with a modular approach; (ii) several changes in the code to facilitate adapting it to other languages, models, and tasks; (iii) the automation of the ensemble generation; and (iv) simplification of the training stage, by translating all the training data for single-language models (and maintaining the data as-is for the multilingual model). In the current approach, the tasks are conducted sequentially: first, a group of models will be responsible for detecting sexist content in the data points; then, another group of models will be used to classify the posts identified as sexist.

Some very important works that used transformer models in NLP are: [29, 30, 6]. The work of [30] conducted important work on the multi-class classification of sexist content, by proposing a model that uses both the outputs of a BERT model and linguistic word embeddings. The authors observed that their model provided better results than several state-of-the-art baselines, such as: convolutional neural networks with different architectures, bidirectional long short-term memory networks, and the use of only the BERT model. It is important to observe that the results were better for both metrics: F1-score and Accuracy.

As observed before, the work by [14] used transformer models for sexism detection and classification at the EXIST 2021 shared task. The authors compared the use of two models: BERT (the single-language version for English and the multilingual version for English and Spanish) and BETO (the single-language version for Spanish). They have also conducted an in-depth hyperparameters analysis. The authors concluded that their system obtained better results than the multilingual BERT model and that the use of ensembles provided better results than the use of single-language models.

Lastly, it is important to cite the work by [6], which used the BERT model and a recurrent neural network for misogyny detection on social media posts, a task similar to sexism detection. It is important to observe that the authors evaluated two datasets (AMI IberEval 2018 and AMI EVALITA 2018) in three languages (English, Italian, and Spanish). The authors concluded that the BERT model provided the best results, resulting in better predictions than the baseline model, a support vector machine model.

4. Methodology

The methodology used in this work was based on the work by [14], and was composed of six steps. It is important to observe that several improvements were made in the methodology, to make it: (i) more generalizable and applicable to other datasets and problems; and (ii) more adapted for use in different languages, by substituting the language-specific transformer models.

Therefore, we believe that the current methodology better suits the tasks of multilingual sexism detection and classification on social media messages. It is also important to emphasize that the code developed could be easily adapted for using other data sources, such as messages from other social media platforms, and news, among others.

In the methodology proposed in this work and illustrated in Figure 1, we follow two tasks sequentially: sexism detection and sexism classification. First, we identify, for a given data point (in the current implementation, a specific tweet), if it contains sexist content or not. After this classification, the tweets labeled sexist are then used on the sexism classification module, to identify one of the following categories: ideological and inequality; stereotyping and dominance; objectification; sexual violence; and misogyny and non-sexual violence. This classification follows the classes on the EXIST2022 shared task dataset [16]. For a further description of the labeling method used and each of the classes, we refer the reader to the EXIST 2022 shared task at IberLEF 2022 [16].

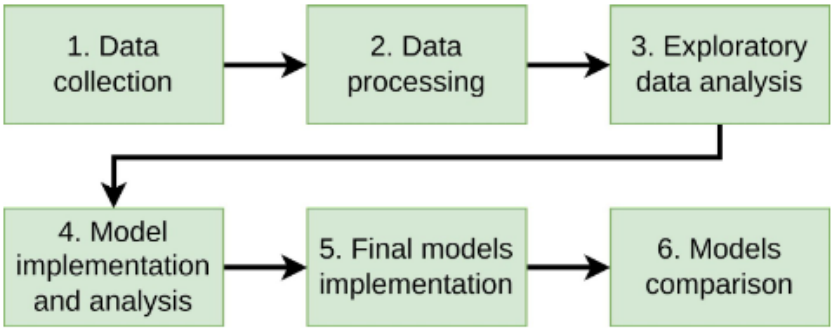

Figure 1: Steps of the methodology used in this work.

The steps of the methodology used in this work were:

1. **Data collection**: the dataset developed for the EXIST 2022 shared task at IberLEF 2022 [16] was used to train, validate, and test the models. This dataset contained labeled data

from two social media platforms: Twitter and Gab. Section 5.1 of this work contains a further description of this dataset;

2. **Data processing**: for both tasks, the following widely used processing techniques were applied: separation of the dataset between languages (English and Spanish) and tokenization. These methods are used in the state of the art NLP tasks in different contexts, as observed in the works by [2, 7, 17, 8]. It is important to observe that we used all data points in the dataset, as no clear outliers were detected. The training subset of the shared task was then divided into training (80%) and validation (20%) subsets for cross-validation. It is essential to observe that, as in the work by [14], one of the strategies used involved translating all the data points to each language (generating a dataset containing only English messages and a dataset containing only Spanish messages). The googletrans (https://github.com/ssut/py-googletrans) library was used for translating the data points;
3. **Exploratory data analysis**: an analysis of each data subset was conducted to characterize class distributions for both tasks 1 (sexism detection) and 2 (sexism classification);
4. **Model implementation and hyperparameters analysis**: several models were implemented, considering two main classes: (i) single-language models with translation between languages (increasing the size of the dataset for each language); and (ii) multi-language models. All implementations considered a 5-fold cross-validation process during the training stage. A hyperparameters analysis was conducted for all models, considering: different learning rates (0.00002, 0.00003, 0.00005, and 0.000005) and a different number of epochs (6, 7, 8, 9, and 10). The other hyperparameters used were the optimal hyperparameter values found by [14] for batch size (8) and max length (128). The metrics used for evaluating the tasks were the same as the EXIST 2022 shared task: Accuracy for task 1 (sexism detection) and F1-Macro for task 2 (sexism classification). The models implemented were:
    - Single-language models for English: BERT [25], RoBERTa [26], Electra [27], and GPT2 [28]. These were named, respectively: En-BERT, En-RoBERTA, En-Electra, and En-GPT2;
    - Single-language models for Spanish: BERT [25], RoBERTa [26], Electra [27], and GPT2 [28]. These were named, respectively: Sp-BERT, Sp-RoBERTA, Sp-Electra, and Sp-GPT2;
    - Multi-language models: BERT [25] and RoBERTa [26]. These were named: Mu-BERT, Mu-RoBERTA, XLM-Align, and Info-XLM;
    - Ensemble models: the highest sum model, which chose the specific class with the highest probability sum considering all models, and a majority vote model, which chose the specific class with the majority of vote between the models. If there is a tie, it randomly selects one of the classes among the tied classes
5. **Final models implementation**: considering the best hyperparameters for each model, identified in Step 4, the final models were built and trained with the whole training dataset (considering both training and validation subsets). Table 1 presents all the final models implemented;
6. **Models comparison**: the final comparison of all models was then conducted on the test subsets, considering the official metrics for the EXIST 2022 shared task at IberLEF 2022

Table 1
Models implemented in this work

Model	Language	Translation of training data
Sp-BERT	Spanish	X
Sp-RoBERTa	Spanish	X
Sp-Electra	Spanish	X
Sp-GPT2	Spanish	X
En-BERT	English	X
En-RoBERTa	English	X
En-Electra	English	X
En-GPT2	English	X
Mu-BERT	Multi	
Mu-RoBERTa	Multi	
XLM-Align	Multi	
Info-XLM	Multi	
Ens-Higher sum		
Ens-Majority vote		

[16]): Accuracy for task 1 and F1-Macro for task 2. Based on this analysis, the best model was chosen.

The implementation was done using Python on a personal computer with 2 GPUs GeForce RTX 2080 (CUDA Version 11.5), with the following technical specifications: Intel(R) Core(TM) i7-7800X CPU @ 3.50GHz CPU. The code implemented is available on an open Github repository: https://github.com/AngelFelipeMP/Transformers-Sexism-Classification.

5. Results

This section presents the main results of the current work and is divided into three subsections: 5.1, which contains a description of the EXIST 2022 shared task and its dataset; 5.2, which presents the results of the models' hyperparameters analysis; and 5.3, which presents a comparison of final trained models on the test subset of the EXIST 2022 shared task.

5.1. EXIST 2022 shared task description

In this work, the dataset used was provided by the EXIST 2022 shared task at IberLEF 2022 [16]. It is an expansion of the EXIST 2021 shared task at IberLEF 2021 dataset, and contains data from posts collected on two important social media platforms: Twitter and Gab, for two languages: English and Spanish. The training subset used contained 11.345 labeled posts, being 5.644 in English and 5.701 in Spanish. The testing subset contained 1.058 tweets from the month of January 2022.

It is important to note that the shared task is composed of two main tasks: (i) sexism detection, containing two classes (sexist and non-sexist); and (ii) sexism classification, containing five classes (ideological and inequality; stereotyping and dominance; objectification; sexual violence;

Table 2
Results for the hyperparameters analysis on the validation subset for task 1.

Model	Best hyperparameters found	Accuracy	F1-Macro
En-BERT	E: 10; Ml: 128; Bs: 8; Lr: 0.000005; D: 0.3	**0.764**	0.763
En-RoBERTa	E: 10; Ml: 128; Bs: 8; Lr: 0.000005; D: 0.3	**0.764**	0.762
En-Electra	E: 10; Ml: 128; Bs: 8; Lr: 0.00001; D: 0.3	0.747	0.745
En-GPT2	E: 10; Ml: 128; Bs: 8; Lr: 0.00003; D: 0.3	0.744	0.743
Sp-BERT	E: 10; Ml: 128; Bs: 8; Lr: 0.000005; D: 0.3	**0.696**	0.690
Sp-RoBERTa	E: 10; Ml: 128; Bs: 8; Lr: 0.00001; D: 0.3	0.693	0.693
Sp-Electra	E: 10; Ml: 128; Bs: 8; Lr: 0.000005; D: 0.3	0.693	0.693
Sp-GPT2	E: 10; Ml: 128; Bs: 8; Lr: 0.00001; D: 0.3	0.675	0.673
Mu-BERT	E: 10; Ml: 128; Bs: 8; Lr: 0.000005; D: 0.3	**0.752**	0.752
Mu-RoBERTa	E: 10; Ml: 128; Bs: 8; Lr: 0.000005; D: 0.3	0.747	0.747
XLM-Align	E: 10; Ml: 128; Bs: 8; Lr: 0.000005; D: 0.3	0.724	0.722
Info-XLM	E: 10; Ml: 128; Bs: 8; Lr: 0.00001; D: 0.3	0.707	0.702

Legend: E: number of epochs; Ml: maximum length; Bs: batch size; Lr: learning rate; D: dropout rate. In highlight: best model for each category (english, spanish, and multilingual) for Accuracy.

and misogyny and non-sexual violence) [16]. For an in-depth description of each class, we refer the reader to the work by [16].

It is also essential to note that the dataset was labeled considering different experts with wide experience on analyzing sexist content in social media. For a further description of the dataset, the tasks themselves, and the methodology used for labeling the posts and to create the training and test subsets, we refer the reader to the work by [16].

5.2. Hyperparameters analysis

Tables 2 and 3 illustrate the main results for the hyperparameters analysis on the validation subsets for each task, respectively. It is important to emphasize that the tasks had different official evaluation metrics: Accuracy for task 1 and F1-Macro for task 2. Therefore, for each task, the official metric will be evaluated. Nevertheless, for a better comparison between the complexity of the tasks, both metrics are presented for the two tasks.

Considering the Accuracy of the validation subset, it is possible to observe in Table 2 that: (i) the BERT model presented the best results for all three categories (English, Spanish, and multiple languages); (ii) that the multiple languages BERT presented better overall results than the Spanish BERT, considering the full validation subset; and (iii) the worst overall model, for all categories, was the GPT2. However, it is important two note that: (i) the RoBERTa model's results were considerably close to the BERT ones for all categories; and (ii) the results observed are similar when F1-Macro evaluated, with the two best models being BERT and RoBERTa.

Lastly, it is crucial to emphasize that most final models presented the values of the following hyperparameters: 10 epochs, a maximum length of 128, a batch size of 8, and a dropout rate of 0.3. The only hyperparameter that showed differences was the learning rate, which seems to be, at least for this specific validation subset, more related to the category than to the model itself. From these results, we can infer that the quality of the model's prediction is directly related to

Table 3
Results for the hyperparameters analysis on the validation subset for task 2.

Model	Best hyperparameters found	Accuracy	F1-Macro
En-BERT	E: 9; Ml: 128; Bs: 8; Lr: 0.00005; D: 0.3	0.667	**0.661**
En-RoBERTa	E: 6; Ml: 128; Bs: 8; Lr: 0.00001; D: 0.3	0.663	0.658
En-Electra	E: 10; Ml: 128; Bs: 8; Lr: 0.00003; D: 0.3	0.660	0.653
En-GPT2	E: 7; Ml: 128; Bs: 8; Lr: 0.00003; D: 0.3	0.652	0.645
Sp-BERT	E: 9; Ml: 128; Bs: 8; Lr: 0.00001; D: 0.3	0.619	**0.612**
Sp-RoBERTa	E: 10; Ml: 128; Bs: 8; Lr: 0.00003; D: 0.3	0.593	0.581
Sp-Electra	E: 9; Ml: 128; Bs: 8; Lr: 0.00001; D: 0.3	0.588	0.576
Sp-GPT2	E: 9; Ml: 128; Bs: 8; Lr: 0.00005; D: 0.3	0.569	0.556
Mu-BERT	E: 8; Ml: 128; Bs: 8; Lr: 0.000005; D: 0.3	0.654	**0.646**
Mu-RoBERTa	E: 8; Ml: 128; Bs: 8; Lr: 0.000005; D: 0.3	0.649	0.643
XLM-Align	E: 9; Ml: 128; Bs: 8; Lr: 0.00001; D: 0.3	0.642	0.636
Info-XLM	E: 9; Ml: 128; Bs: 8; Lr: 0.00001; D: 0.3	0.643	0.635

Legend: E: number of epochs; Ml: maximum length; Bs: batch size; Lr: learning rate; D: dropout rate. In highlight: best model for each category (english, spanish, and multilingual) for F1-Macro.

the specific language, as is widely observed in the literature. One of the possible explanations for this observation is that languages such as Spanish have fewer and lower quality models and word embeddings than English.

Analyzing Table 3, which is related to task 2, it is possible to observe considerable differences from what was observed for task 1. It is important to emphasize that the metric evaluated for this task was the F1-Macro and that the validation subset for this task only considered the posts that were considered sexist. Although the BERT model continues to show the best results in all categories, all models' results are considerably lower than in task 1. This may be explained by the increased complexity of a multi-class classification in relation to a binary label classification.

The best hyperparameters for each model and category varied a lot, which may be directly related to the complexity of the task in relation to task 1. Additionally, in the case of Spanish, the BERT model was considerably better than all other models. Lastly, comparing the F1-Macro between the BERT models for both tasks, it is possible to observe that it was 10.29% lower for English, 7.83% lower for Spanish, and 10.41% lower for multilingual for task 2 in relation to task 1.

Table 4 illustrates the number of winning models (models that presented the best metrics) for each task, considering the values of the hyperparameters explored on the validation subset. It is important to observe that: (i) as cited before, the values were more homogeneous between the models and categories for task 1, due to its lower complexity in relation to task 2; (ii) for task 1, the best learning rate was 0.000005 and the best number of epochs was 10; and (iii) for task 2, the best learning rate was 0.00001 and the best number of epochs was 9.

The results in this section can be important for both: (i) evaluating and comparing other models with the same languages explored in this work; (ii) evaluating and comparing the same models for different languages; and (iii) better guiding other researchers on hyperparameters values choice when starting to work with sexism detection and classification on social media

Table 4
Summary of the winning models for each hyperparameter value for both tasks.

Hyperparameter	Values	Task 1	Task 2
Learning rate	**0.000005**	**58.00%**	16.67%
	0.00001	33.30%	**41.67%**
	0.00003	8.30%	25.00%
	0.00005	0.00%	16.67%
Number epochs	**10**	**100.00%**	16.67%
	9	0.00%	**50.00%**
	8	0.00%	16.67%
	7	0.00%	8.30%
	6	0.00%	8.30%

Legend: In highlight: best hyperparameter value for each task.

Table 5
Results of the final models on the test subset for task 1 in English.

Model	Accuracy	F1-Macro
En-RoBERTa	**0.954**	0.954
Mu-BERT	**0.948**	0.948
Info-XLM	0.864	0.863
Highest Sum	0.861	0.858
Majority Vote	0.846	0.844
XLM-Align	0.807	0.803
En-BERT	0.767	0.764
Mu-RoBERTa	0.746	0.742
En-GPT2	0.731	0.723
En-Electra	0.710	0.687

Legend: in highlight: best model for each category (single-language and multilingual) for Accuracy.

posts.

5.3. Final models comparison

In this section, we will discuss the main results observed on the test subsets for both tasks. Tables 5 and 6 contain the main results for task 1, and tables 7 and 8 contain the main results for task 2. It is important to emphasize that the test subsets used were the ones provided by the EXIST 2022 Shared Task, as described in section 5.1.

Tables 5 and 6 illustrate the results of the final models on the test subset for task 1 for English and Spanish, respectively. In the case of English (Table 5), it important to observe that: (i) the best overall model was the RoBERTa, with an Accuracy of 0.954, closely followed by the multilingual BERT (Accuracy: 0.948) and the multilingual Info-XLM (Accuracy: 0.864); and (ii) in general, the multilanguage and ensemble models presented better results than single-language models.

Table 6
Results of the final models on the test subset for task 1 in Spanish.

Model	Accuracy	F1-Macro
Info-XLM	**0.831**	0.830
Highest Sum	0.819	0.819
Majority Vote	0.814	0.813
XLM-Align	0.797	0.795
Mu-BERT	0.780	0.774
Mu-RoBERTa	0.739	0.738
Sp-GPT2	**0.696**	0.684
Sp-RoBERTa	0.688	0.687
Sp-Electra	0.620	0.613
Sp-BERT	0.546	0.474

Legend: in highlight: best model for each category (single-language and multilingual) for Accuracy.

In the case of Spanish (Table 6), it is possible to observe that: (i) the best model was the Info-XLM (F1-Macro: 0.830), followed by the highest sum (F1-Macro: 0.819) and majority vote (F1-Macro: 0.813) ensembles; (ii) all multilingual models presented better results than single-language ones; and (iii) the worst model was the single-language BERT. Additionally, it is possible to observe that, in general, the models presented worse results for Spanish than for English.

From these results, a series of inferences can be made: (i) in general, for languages that contain fewer quality models and materials for NLP tasks, the use of multilingual models is recommended for sexism detection; (ii) the BERT and RoBERTa model presented better results, especially when considering their multilingual version; and (iii) translating the data points to English and using a single-language English model or a multilanguage model provided better results than using a single-language Spanish model. Nevertheless, more research is needed to further validate those observations, as they consider only one dataset, few models, and only two languages.

Tables 7 and 8 illustrate the results of the final models on the test subset for task 2 for English and Spanish, respectively. For English (Table 7), it is important to observe that: (i) the best overall model was the single-language BERT (F1-Macro: 0.997); (ii) the second-best model, the multilingual RoBERTa, had a considerably lower F1-Macro (0.900); and (iii) the difference between the F1-Macro metrics for the models was considerably larger than in task 1.

In the case of Spanish (Table 8), it is possible to observe that: (i) the multilingual RoBERTa was the best model (F1-Macro: 0.784), followed by the highest sum (F1-Macro: 0.732) and majority vote (F1-Macro: 0.716) ensembles; (ii) as observed for task 1, the multilingual models presented better results than the single-language ones; (iii) the overall results were considerably worse than for task 1; and (iv) the worst model was the single-language RoBERTa.

Even though the results were different from the ones observed in task 1, some similarities can be observed: (i) using translation to English and then implementing single-language English models or multilingual models improved considerably the quality of the results; (ii) the BERT and RoBERTa models for English seem to present the best results; and (iii) the single-language

Table 7
Results of the final models on the test subset for task 2 in English.

Model	Accuracy	F1-Macro
En-BERT	0.998	**0.997**
Mu-RoBERTa	0.938	**0.900**
Highest Sum	0.890	0.823
Majority Vote	0.871	0.791
Info-XLM	0.817	0.707
En-RoBERTa	0.817	0.703
En-Electra	0.819	0.700
XLM-Align	0.815	0.700
Mu-BERT	0.805	0.681
En-GPT2	0.587	0.300

Legend: in highlight: best model for each category (single-language and multilingual) for F1-Macro.

Table 8
Results of the final models on the test subset for task 2 in Spanish.

Model	Accuracy	F1-Macro
Mu-RoBERTa	0.869	**0.784**
Highest Sum	0.837	0.732
Majority Vote	0.827	0.716
XLM-Align	0.807	0.689
Info-XLM	0.809	0.685
Mu-BERT	0.778	0.633
Sp-Electra	0.779	**0.626**
Sp-BERT	0.750	0.558
Sp-GPT2	0.722	0.519
Sp-RoBERTa	0.682	0.477

Legend: in highlight: best model for each category (single-language and multilingual) for F1-Macro.

Spanish models presented the worst results. Therefore, we can infer that those observations could be further validated and explored by other researchers for different models and languages, as these could considerably improve the analysis of models and hyperparameters for both sexism detection and classification.

6. Discussion

In this section, we discuss how the proposed methodology compares to other models for the EXIST 2022 shared task for both tasks, how it compares with the work by the winners of the EXIST 2021 shared task challenge, and the main limitations observed in this research.

The best runs of the proposed methodology used in this work ranked 18 for task 1 (with an Accuracy of 0.764 and an F1-Macro of 0.764) and 17 for task 2 (with an Accuracy of 0.627 and an F1-Macro of 0.452). As expected, the results for the metrics on the test subset were considerably

lower than in the validation subset. Although the proposed methodology did not reach the top 5 best models for either of the categories, it is still possible to infer that the models presented good results.

As a comparison, the winning model for both tasks at EXIST 2021 shared task (whose whole dataset was used as the training subset for EXIST 2022) presented an Accuracy of 0.780 for task 1 and an F1-Macro of 0.579 [14]. Therefore, our proposed methodology achieved comparable results for task 1.

The two main differences between our proposal and the winning model of the EXIST 2021 shared task challenge [14] were: (i) we translated all data points for the single-language models, while that work considered also a non-translated dataset (which contained fewer points available for training the single-language models); and (ii) we considered other state-of-the-art language models other than BERT and BETO.

Additionally, we also considered implementation aspects, making our implementation much easier to adapt to other languages, tasks, and models. Therefore, although the results obtained were similar for task 1 and worse for task 2, we believe that the currently proposed methodology is more useful in real-world scenarios (for example, for using bots for crawling and detecting sexism in different social media platforms).

However, much work is still needed to tweak the models, explore more hyperparameters, and language models, and test for other languages and datasets. We also believe, especially in the case of the multilingual models, that the methodology can be used to conduct different tasks than the ones considered in the EXIST 2022 shared task.

As observed by [14] on the winning model for EXIST 2021, there is also the possibility to improve the quality of the proposed methodology by inserting sentiment lexicons, such as Vader [31] and Hurtlex [22]. Both are very important for sentiment analysis and are used in different tasks.

Additionally, feature engineering (for example, by using unsupervised learning models to generate new features to be used as inputs for the models) and transfer learning (for example, by training the models in other datasets with similar characteristics or tasks) could also be important ways to improve the quality of the models implemented in this work. The current methodology allows for the use of those techniques with fewer changes in the implementation, due to the modularity principles adopted when coding the steps of the methodology.

Finally, the main limitations of this work were: (i) not considering additional datasets in the analysis, as the focus was on developing a solution and participating in the EXIST 2022 shared task challenge; (ii) not considering the use of unsupervised learning methods to improve the quality of the inputs used by the models; and (iii) the lack of open code for other implementation of sexism detection and classification models that could be used for comparison with our results. However, we believe that we have successfully addressed those limitations within the scope of this work.

7. Conclusion and future work

In this work, we explored the use of different architectures of transformers on two very important problems related to social media posts: detecting sexist content and classifying it into relevant

labels. Both tasks are very complex due to two main reasons: (i) the volume of posts on social media platforms, demanding an automatic solution; and (ii) the wide variety of forms sexist content may have. Additionally, some languages have fewer high-quality resources for use in NLP problems, such as language models, domain-specific lexicons, and labeled datasets for model training.

Therefore, we proposed a methodology using four different architectures of transformers (BERT, RoBERTa, Electra, and GPT2) for both tasks in Spanish and English (including both single-language and multilingual models). The methodology was implemented on the EXIST 2022 shared task dataset, which consisted of sexism detection and sexism classification tasks. We have also conducted an in-depth hyperparameters analysis that can be used as a starting point by other researchers to explore both problems.

We compared the results of the different models in the validation and test subsets, and concluded that: (i) the use of the translation of the posts to English and then using single-language English and multilingual models present the best results; (ii) the best model for task 1 in English was single-language RoBERTa with an Accuracy of 0.954, and in Spanish was Info-XLM with an Accuracy of 0.831; (iii) the best model for task 2 in English was single-language BERT with an F1-Macro of 0.997, and in Spanish was multilingual RoBERTa with an F1-Macro of 0.784; (iv) the best architectures were BERT and RoBERTa; (v) using the single-language Spanish models provided the worst results; (vi) task 2 presented worse results for all models due to its higher complexity in relation to task 1; and (vii) the use of ensembles were better than the Info-XLM and XLM-Align models, but worse than English single-language and multilingual models.

Future works are related to: (i) implementing the proposed methodology with the best models in a real case scenario (for example, a crawler bot that also classifies the tweets and warns when sexist content is detected); (ii) testing and analyzing the behavior of the best models on different datasets and languages; (iii) implementing unsupervised models to generate features that may improve the quality of the models; (iv) evaluating the use of different labeled datasets and transfer learning techniques to improve the models' predictions; and (v) evaluating the use of general and domain-specific sentiment lexicons on the proposed methodology to improve the quality of the models' predictions.

References

[1] A. M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, N. Kourtellis, Large scale crowdsourcing and characterization of twitter abusive behavior, in: Twelfth International AAAI Conference on Web and Social Media, 2018.

[2] W. Yin, A. Zubiaga, Towards generalisable hate speech detection: a review on obstacles and solutions, arXiv preprint arXiv:2102.08886 (2021).

[3] 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.

[4] P. Fortuna, J. Soler-Company, L. Wanner, How well do hate speech, toxicity, abusive

and offensive language classification models generalize across datasets?, Information Processing & Management 58 (2021) 102524.
[5] P. Chiril, V. Moriceau, F. Benamara, A. Mari, G. Origgi, M. Coulomb-Gully, An annotated corpus for sexism detection in french tweets, in: Proceedings of The 12th Language Resources and Evaluation Conference, 2020, pp. 1397–1403.
[6] E. W. Pamungkas, V. Basile, V. Patti, Misogyny detection in twitter: a multilingual and cross-domain study, Information Processing & Management 57 (2020) 102360.
[7] N. Chetty, S. Alathur, Hate speech review in the context of online social networks, Aggression and violent behavior 40 (2018) 108–118.
[8] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and benchmark corpora for hate speech detection: a systematic review, Language Resources and Evaluation (2020) 1–47.
[9] A. F. M. de Paula, R. F. da Silva, I. B. Schlicht, Ai-upv at iberlef-2021 detoxis task: Toxicity detection in immigration-related web news comments using transformers and statistical models, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain, volume 2943, 2021, pp. 547–566. URL: http://ceur-ws.org/Vol-2943/detoxis_paper2.pdf.
[10] I. B. Schlicht, A. F. M. de Paula, P. Rosso, Upv at checkthat! 2021: Mitigating cultural differences for identifying multilingual check-worthy claims, in: Proceedings of The 12th Conference and Labs of the Evaluation Forum (CLEF 2021)., Bucharest, Romania, volume 2936, 2021, pp. 465–475. URL: http://ceur-ws.org/Vol-2936/#paper-36.
[11] I. B. Schlicht, A. F. M. de Paula, Unified and multilingual author profiling for detecting haters, in: Proceedings of The 12th Conference and Labs of the Evaluation Forum (CLEF 2021)., Bucharest, Romania, volume 2936, 2021, pp. 1837–1845. URL: http://ceur-ws.org/Vol-2936/#paper-157.
[12] I. B. Schlicht, A. F. M. de Paula, P. Rosso, Upv at trec health misinformation track 2021 ranking with sbert and quality estimators, Proceedings of The Thirtieth Text REtrieval Conference (TREC 2021)., Gaithersburg, United States (2021). URL: https://trec.nist.gov/pubs/trec30/papers/UPV-HM.pdf.
[13] M. Sajjad, F. Zulifqar, M. U. G. Khan, M. Azeem, Hate speech detection using fusion approach, in: 2019 International Conference on Applied and Engineering Mathematics (ICAEM), IEEE, 2019, pp. 251–255.
[14] A. F. M. de Paula, R. F. da Silva, I. B. Schlicht, Sexism prediction in spanish and english tweets using monolingual and multilingual bert and ensemble models, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain, volume 2943, 2021, pp. 356–373. URL: http://ceur-ws.org/Vol-2943/exist_paper2.pdf.
[15] H. Abburi, P. Parikh, N. Chhaya, V. Varma, Fine-grained multi-label sexism classification using a semi-supervised multi-level neural approach, Data Science and Engineering 6 (2021) 359–379.
[16] 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).
[17] F. A. Acheampong, H. Nunoo-Mensah, W. Chen, Transformer models for text-based emotion detection: a review of bert-based approaches, Artificial Intelligence Review (2021) 1–41.
[18] T. Wolf, J. Chaumond, L. Debut, V. Sanh, C. Delangue, A. Moi, P. Cistac, M. Funtowicz, J. Davison, S. Shleifer, et al., Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 38–45.
[19] B. Gambäck, U. K. Sikdar, Using convolutional neural networks to classify hate-speech, in: Proceedings of the first workshop on abusive language online, 2017, pp. 85–90.
[20] F. E. Ayo, O. Folorunso, F. T. Ibharalu, I. A. Osinuga, Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions, Computer Science Review 38 (2020) 100311.
[21] O. Istaiteh, R. Al-Omoush, S. Tedmori, Racist and sexist hate speech detection: Literature review, in: 2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA), IEEE, 2020, pp. 95–99.
[22] E. Bassignana, V. Basile, V. Patti, Hurtlex: A multilingual lexicon of words to hurt, in: 5th Italian Conference on Computational Linguistics, CLiC-it 2018, volume 2253, CEUR-WS, 2018, pp. 1–6.
[23] A. Koufakou, E. W. Pamungkas, V. Basile, V. Patti, Hurtbert: Incorporating lexical features with bert for the detection of abusive language, in: Proceedings of the Fourth Workshop on Online Abuse and Harms, 2020, pp. 34–43.
[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
[25] 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.
[26] 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, arXiv preprint arXiv:1907.11692 (2019).
[27] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, Electra: Pre-training text encoders as discriminators rather than generators, arXiv preprint arXiv:2003.10555 (2020).
[28] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.
[29] M. Bugueño, M. Mendoza, Learning to detect online harassment on twitter with the transformer, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2019, pp. 298–306.
[30] P. Parikh, H. Abburi, N. Chhaya, M. Gupta, V. Varma, Categorizing sexism and misogyny through neural approaches, ACM Transactions on the Web (TWEB) 15 (2021) 1–31.
[31] C. Hutto, E. Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 8, 2014.