Enhancing Sexism Identification and Categorization in Low-data Situations

Vicent Ahuir[1], José Ángel González[1] and Lluís-Felip Hurtado[1]

[1] *Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València, Camino de Vera, s/n, 46022 Valencia, Spain*

Abstract

With the consolidation of social media networks as a backbone communication channel of our society, the freedom of speech, anonymity, and global communication have been greatly extended. However, adverse movements like racism, xenophobia, or sexism have also been spread under the umbrella of those benefits. For this reason, it is essential to address this problem. In this work, we present the participation of the ELiRF-VRAIN team in the sEXism Identification in Social neTworks (EXIST) shared task at IberLEF 2022. Our work focuses on addressing the problem of sexism identification and categorization by using Transformer models, data augmentation, and ensembles to increment the performance in the sexism classification, when we do not have at our disposal a large amount of data to work with. Our submissions achieved 76.94% Accuracy on sexism identification, and 49.91 Macro F1 on sexism categorization.

Keywords

Sexism identification, Sexism categorization, Transformer ensembles, BERT-like models

1. Introduction

Social media networks have created new communication channels that have brought new modes of interaction between people. These networks have facilitated the accessibility to information and have provided tools for people to spread information and thoughts easily and quickly; thus, they can express themselves in different ways (text, emoticons, GIFs, images, videos) [1, 2]. The positive effects of social media are primarily apparent: global communication, anonymity, freedom of speech, and better accessibility to information, for instance. However, all those benefits also facilitate the movement of racism, xenophobia, or sexist expressions [3], which is a severe negative aspect. Due to the volume of information generated each day on social media, using technology is essential to address this problem and mitigate the appearance of these kinds of expressions. This work focuses on automatic sexism identification and categorization in text using deep learning.

Sexism is prejudice or discrimination based on sex [4], a set of actions or attitudes that discriminate against people based entirely on their gender [5]. This mindset or behavior

IberLEF 2022, September 2022, A Coruña, Spain.

viahes@dsic.upv.es (V. Ahuir); jogomba2@dsic.upv.es (J. González); lhurtado@dsic.upv.es (L. Hurtado)

00000.01-5636-651X (V. Ahuir); 00000.03-3812-5792 (J. González); 00000.02-1877-0455 (L. Hurtado)

is linked to beliefs around the roles that women and men should play in society. Sexism can affect everyone, but women are predominantly affected [6]. Sexism towards women manifests in different ways, but a sexist attitude would show certain signals: inequality between men and women, a degree of dominance of the male gender over the female one, separation of roles between genres, objectification of women, or words that would suggest any violence towards women [7].

Sexism identification in text is an open problem in Natural Language Processing (NLP). The problem can be seen from two perspectives: (i) detecting whether there is any presence of sexism, and (ii) identifying how the sexist attitude is expressed. Both of them are essential depending on the level of detail needed.

For Spanish and other Iberian languages, the Iberian Languages Evaluation Forum (IberLEF) proposed the first edition of a shared task about detecting sexism in social media, the sEXism Identification in Social neTworks (EXIST)[7]. The EXIST shared task consists of the following tasks: sexism identification (a binary classification problem) and sexism categorization (a multiclass classification problem). This article presents our contributions to the second edition of the EXIST shared task [8].

Our approach for the EXIST shared task was based on Transformers[9], the deep learning architecture that has pushed the state-of-the-art in most of the NLP problems [10, 11, 12], including classification [13, 14]. Since there is typically a small amount of labeled data for downstream tasks, we fine-tuned models that were pre-trained in a self-supervised way with large amounts of unlabeled data; which is a commonly used strategy. We used data augmentation to expand the number of training samples which could increase performance, and also, as a regularization method that would avoid overfitting [15, 16, 17]. We use monolingual models instead of multilingual ones, which would help in low-data situations. Also, we integrate all the previous ideas in an ensemble of models with the aim of increasing the performance of our classification system [18].

We participate in EXIST 2022 by sending three different runs for each task. The first run is a system that combines two monolingual classification models. The second run is a system based on an ensemble of five classification monolingual models per language. The third run is a system that has the same architecture as the second one; however, the models are trained with all the available data, including the one that was used for validating or testing during the development. With these approaches, we achieve the 6^{th} place on the sexism identification task, and the 2^{nd} place on the sexism categorization task.

The main contributions of this work are: (i) explore the benefits and downsides, depending on the task, of a classification system based on a ensemble of models with voting system, (ii) describe the process of fine-tuning models on a low-data situation, and how to increase robustness and performance with data augmentation.

2. The tasks

The EXIST shared task focuses on Spanish and English languages, and consists of two tasks. The first task is about sexism identification, where it is needed to identify whether

a social media text shows sexism or not. The second task is about identifying five different manifestations of sexism: *ideological and inequality*, *stereotyping and dominance*, *objectification*, *sexual violence*, and *misogyny and non-sexual violence*. Also, in the second task, non-sexist texts should be identified separately. Thus, the first task is a binary classification problem, and the second task is a multiclass-classification one.

The evaluation metric differs depending on the task. In the first task *Accuracy* is used as the evaluation metric; thus, the more samples correctly classified, the better. In contrast, in the second task, the *Macro F1*. Here, the classifiers should balance the performance among all the classes, regardless of the class frequency in the datasets.

3. The datasets

All our models were fine-tuned using the training dataset provided in the EXIST shared task. We also used the testing dataset from the first edition of EXIST for validation and testing purposes. One self-imposed requirement in our work was to avoid including other external data. This restriction aimed to find ways to enrich the training process when data is limited, as having a small amount of data could produce a handicap when we want to use deep learning architectures like Transformers [9].

All the EXIST samples in the datasets are in Spanish or English, and were collected from two social media platforms: *Twitter* and *Gab Social*. Each sample in the datasets is labeled with (i) the presence of sexism, and (ii) the categorization of sexism, if present. The samples that do not show sexism were labeled with *non-sexist* for both tasks. The samples that contain sexism were labeled as *sexist* for the first task and as one of the five different classes for the second task: *ideological-inequality*, *stereotyping-dominance*, *objectification*, *sexual-violence*, or *misogyny-non-sexual-violence*.

Table 1 shows the class distribution of the samples across the two tasks. In the first task, the percentage of samples is similar between the two classes. In the second task, the distribution is strongly imbalanced due to the *non-sexist* class. Moreover, the imbalance also appears between the *sexist* classes; being the *ideological-inequality* the most represented, and *objectification* and *sexual-violence* the classes with fewer samples.

Regarding the distribution in the training dataset of the samples by language, the 50.8% of the samples are in Spanish and the 49.2% are in English. All the samples of this dataset were captured from Twitter social media.

Table 1
Class distribution of samples in the EXIST training dataset (6977 samples).

Task 1			Task 2		
Class	Count	%	Class	Count	%
non-sexist	3600	51.6	non-sexist	3600	51.6
sexist	3377	48.4	ideological-inequality	866	12.41
			stereotyping-dominance	809	11.6
			objectification	500	7.17
			sexual-violence	517	7.41
			misogyny-non-sexual-violence	685	9.81

The class distribution of the testing dataset from the previous EXIST edition is shown in Table 2. In this dataset we observe that the presence of *sexist* samples is higher than the *non-sexist*, showing a slight imbalance between both classes for the first task. For the second task, appears the imbalance in the training dataset for the same reasons previously mentioned. Regarding the distribution by language, 49.5% of the samples are in Spanish and 50.5% in English. This dataset was created by capturing the 77.5% of the messages from Twitter and the 22.5% from Gab Social.

Table 2
Class distribution of samples in the testing dataset from EXIST 2021 (4368 samples).

Task 1			Task 2		
Class	Count	%	Class	Count	%
non-sexist	2087	47.78	non-sexist	2087	47.78
sexist	2281	52.22	ideological-inequality	621	14.22
			stereotyping-dominance	464	10.62
			objectification	324	7.42
			sexual-violence	400	9.16
			misogyny-non-sexual-violence	472	10.80

4. System architecture

We experimented with two main alternatives to approach the EXIST shared task. The first alternative was a single-model approach, that consisted in fine-tuning different pre-trained BERT-like models for Spanish and English, and then using the best model in terms of the evaluation metrics for each language independently. The second alternative was to use all the fine-tuned models and create an ensemble for each language by creating a voting system.

Figure 1 shows the architecture of the classification system that uses only one model for

Spanish and another for English. The system receives a set of samples to classify. Each sample has a property that identifies the language of the messages; thus, no language detection is needed for our purpose. The system splits the samples into two sets, one for the Spanish samples and another one for the English samples. After that, the best model for each language classifies the samples, outputting the class labels. Finally, the class labels are saved, preserving the original order of the dataset.

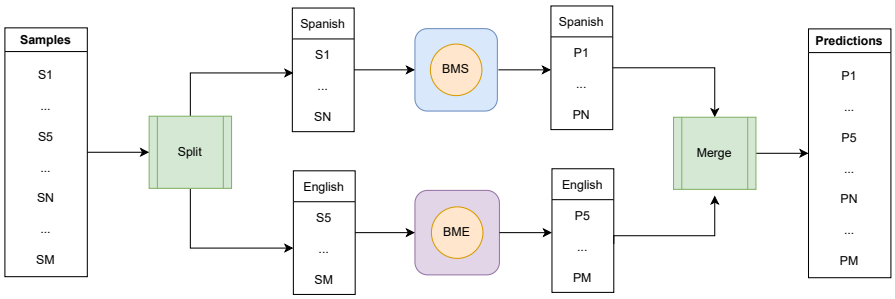

Figure 1: Architecture of the single-model approach. BMS stands for "Best Model for Spanish" and BME refers to the "Best Model for English".

Figure 2 shows the classification system based on ensembles of models. The general structure of the system is very similar to the one shown in Figure 1. The main difference is that the system uses an ensemble of five models to classify the samples of each language independently. Each model in the ensemble for Spanish classifies all the Spanish samples during the classification process, and the same applies to English. Later, a voting method selects the predicted class for each sample. The voting method for each task is different. In the first task, the method selects the class that more models chose; in a tied vote, the class that maxed first is selected. In the second task, a weighted voting method is used. Every model voted for a single class, and its vote was weighted with the likelihood of its prediction. The voting method chooses the class that had received more likelihood, selecting the class with higher accumulated likelihood.

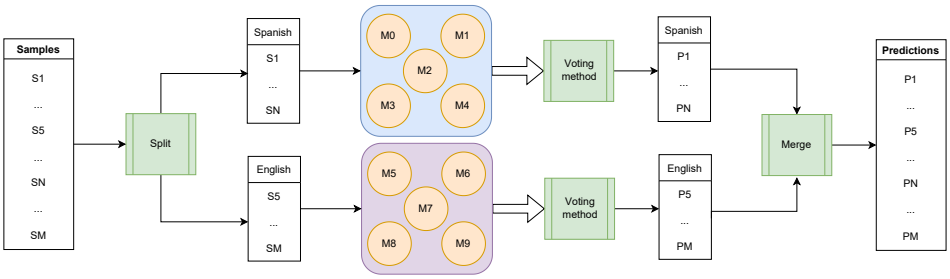

Figure 2: Architecture of the ensemble approach. In blue and purple background, the group of models for Spanish and English respectively.

The architectures shown in Figures 1 and 2 were used indistinctly for the two tasks in EXIST; the difference reside in the models utilized within the system, and the voting method. We fine-tuned twenty different models, from five different pre-trained

Transformers per language and task.

All our classifications models were obtained by fine-tuning a set of pre-trained Transformer models publicly available in the HuggingFace hub [19]. For Spanish we chose the following four pre-trained models: `roberta-large-bne` [20], `bert-base-spanish-wwm-cased` [21], `bert-base-es-cased` [22], and `bert-base-5lang-cased` [22]. For English we choose the following ones: `roberta-large` [11], `bert-large-cased` [10], `hateBERT` [23], and `albert-base-v2` [24]. Additionally, we fine-tuned the `twitter-xlm-roberta-base` model[13] for Spanish and English separately, resulting in an additional model for each language.

5. Dataset partitioning and data augmentation

For the fine-tuning process, three datasets are required: the training dataset for adjusting the pre-trained model, the validation dataset for selecting the best checkpoint during the training process, and the testing dataset, which gives us additional information about how the models generalize to unseen distributions. As it was shown in Section 3, we only had two datasets. Thus we split the EXIST testing dataset from 2021 into two datasets; one for validation and one for testing. In these datasets, we preserved the original distribution of the classes regarding the second task.

We also split the three resulting datasets by the language of the samples since we trained monolingual models. After splitting the training dataset by language, less than 4000 samples were available for training any model. We aim to increase this training dataset by using data augmentation. The first step to enlarge the training datasets was to translate the samples from the opposite language automatically. We used translation Transformer models from the Tatoeba Translation Challenge[25]. The second step was to back-translate the messages from the same language. We translated the Spanish samples to English and then back to Spanish; if the resulting text differed from the original one, it was added as a new sample. The same process was applied to the English samples, translating them to Spanish and then back to English. All the new samples created during the previous data augmentation steps were joined with the original samples and created extended datasets.

Table 3 shows how many samples each data augmentation action added to the final datasets. All samples from the opposite language were translated and added as new samples. It is not the case of back-translation where some of the back-translated texts did not differ from the original texts; thus, they were not added. Also, it could be noticed that the data augmentation techniques were also applied to validation and testing. This is because, during the final proposed systems, we add some of those datasets for training. However, only the original samples were utilized when those datasets were used for validating or testing.

In addition to text translation, we augmented the datasets by masking randomly selected tokens following a uniform distribution. Between the 10% and 30% of the tokens of each message are masked. This strategy helped the models to reduce overfitting and and effectively increased their performance.

Table 3
Amount of samples obtained by each action (original split, translation, and back-translation), and the final size of each dataset.

	Spanish				English			
Dataset	Ori.	Trans.	Back-T.	Total	Ori.	Trans.	Back-T.	Total
Training	3541	3436	3476	10 453	3436	3541	3334	10 311
Validation	1080	1104	1048	3232	1104	1080	1082	3266
Testing	1080	1104	1045	3229	1104	1080	1088	3272

6. The fine-tuning process and model selection

We took the following workflow to obtain the ten monolingual classification models for each task. First, we fine-tuned the Transformer models only with the training datasets, and selected the best single-model checkpoint, in terms of the evaluation metrics, after each validation step. Second, we evaluated the models with the evaluation metric for the specific task; in order to find the best models for the single-model approach. Finally, we did the fine-tuning process again, but adding to the training data the testing dataset; which would add the chance to increase the performance of the final models.

As explained in Section 5, in addition to the two translation-based data augmentation strategies, we added a third one based on masking randomly selected tokens. Using this strategy, we enlarged the training dataset five times its original size. During the fine-tuning process, we empirically observed that enlarging more than five times the training dataset did not provide higher benefits.

Table 4 shows the performance of the models on sexism identification in the validation and testing datasets. The tables 5a and 5b present the results for the models obtained without using the random token masking and the performance for the ones that included this data augmentation strategy during the fine-tuning process. In the first task, the models that has the best performance were `bert-base-spanish-wwm-cased` for Spanish, and `roberta-large` for English. We used the models obtained from the final fine-tuning of those pre-trained models for the single-model approach for the first task.

Table 5 shows similar information as Table 4, but regarding sexism categorization. In this second task the best performance models were `bert-base-spanish-wwm-cased` for Spanish, and `roberta-large` for English. Thus, for the single-model approach of the second task, we used the models obtained from final fine-tuning of those pre-trained models.

On both, Tables 4 and 5, it can be seen that in most cases, the use of random token masking during the fine-tuning process has resulted in an increment in performance. For this reason, we included this action in the final fine-tuning process.

Table 6 shows the twenty models used by the systems that we sent to the EXIST-2022 workshop. Each line represents the two models obtained with the pre-trained Transformer model, one for the first task and another for the second. As was expected, all the obtained models outperformed those trained without the testing dataset samples.

Finally, we also obtained another set of models for each task. These were fine-tuned

Table 4: *Accuracy* of models in the first task, sexism identification. Models fine-tuned only with the training datasets.

(a) Spanish

	Without masking		With masking	
Model	Val.	Test.	Val.	Test.
bert-base-es-cased	72.30	72.51	74.23	73.75
bert-base-5lang-cased	74.42	73.49	75.97	74.2
bert-base-spanish-wwm-cased	**79.15**	**79.45**	**80.98**	**80.34**
roberta-large-bne	71.62	72.78	76.74	78.11
twitter-xlm-roberta-base$_{es}$	79.05	78.11	77.61	76.69

(b) English

	Without masking		With masking	
Model	Val.	Test.	Val.	Test.
albert-base-v2	46.37	46.37	54.29	52.33
bert-large-cased	77.03	77.03	74.48	76.25
hateBERT	75.56	75.56	74.86	75.82
roberta-large	**78.32**	**78.32**	**77.43**	**77.46**
twitter-xlm-roberta-base$_{en}$	74.53	74.53	76.00	72.54

Table 5: *Macro F1* of models in the second task, sexism identification. Models fine-tuned only with the training datasets.

(a) Spanish

	Without masking		With masking	
Model	Val.	Test.	Val.	Test.
bert-base-es-cased	52.12	53.68	53.68	53.72
bert-base-5lang-cased	52.22	53.68	54.21	54.37
bert-base-spanish-wwm-cased	**59.21**	**59.67**	59.67	**59.12**
roberta-large-bne	58.30	59.72	**60.92**	59.08
twitter-xlm-roberta-base$_{es}$	53.92	56.31	56.31	56.01

(b) English

	Without masking		With masking	
Model	Val.	Test.	Val.	Test.
albert-base-v2	47.33	49.08	48.70	50.65
bert-large-cased	55.02	55.81	56.13	53.77
hateBERT	50.00	50.81	60.83	54.38
roberta-large	**57.15**	**57.68**	**62.10**	**57.70**
twitter-xlm-roberta-base$_{en}$	50.50	48.93	61.20	50.39

with all the samples available and without any validation process. For this reason, we did not have a set of samples to find out the best checkpoint during the training process; the final checkpoint was assumed to be the best checkpoint. Since we did not have samples to test these models, we cannot present any details about their performance. These models were used in our third run in the workshop.

Table 6
Performance of the final models, fine-tuned with the training and testing data, and evaluated with the validation dataset. First task measured with *Accuracy*, and the second with *Macro F1*.

	Model	T1: *Accu.*	T2: *Macro F1*
Spanish	bert-base-es-cased	80.07	68.46
	bert-base-5lang-cased	77.02	68.95
	bert-base-spanish-wwm-cased	87.49	68.17
	roberta-large-bne	80.82	68.91
	twitter-xlm-roberta-base$_{es}$	82.21	67.10
English	albert-base-v2	64.67	58.74
	bert-large-cased	76.63	64.60
	hateBERT	75.27	59.64
	roberta-large	78.62	67.67
	twitter-xlm-roberta-base$_{en}$	76.72	59.02

7. Results

For each EXIST task, we sent the following three systems. The first system, consists in the single-model approach with –`bert-base-spanish-wwm-cased` models for Spanish, and `roberta-large` models for English. The second system, based on the voting ensemble of all models listed in Table 6. The third system is also an ensemble, but it uses the models fine-tuned with all the available data.

Table 7 shows the results and rankings of our three runs in the first task. Our best run (system #2), ranks 12$^{\text{th}}$ in the competition, and 6$^{\text{th}}$ if we consider single-team submissions. The system achieved 3.02 less *Accuracy* than the best system in the competition. The run just before our best one obtained 0.38 better *Accuracy*. We observe that our best ensemble system increased the *Accuracy* by 0.38 in comparison to our single-model system. Also, we notice that the ensemble systems performed similarly.

Table 7
Task 1: Sexism Identification. Results of ELiRF-VRAIN in EXIST 2022. *Accuracy* is the reference metric for the first task.

System	Ranking	Team Rank.	*Accuracy*	*F1-score*	1$^{\text{st}}$: *Acc.*	1$^{\text{st}}$: *F1*	11$^{\text{th}}$: *Acc.*	11$^{\text{th}}$: *F1*
#1	14	-	76.56	76.55	-	-	-	-
#2	12	6	76.94	76.86	79.96	79.78	77.32	77.08
#3	13	-	76.84	76.79	-	-	-	-

Table 8 shows the results of our three runs in the second task. In this task, with the third system, we achieved the 3$^{\text{rd}}$ best run result and the 2$^{\text{nd}}$ place as a team. Our system obtained 1.15 less *Macro F1* than the best system in the competition, and we obtained 2.44 better *Macro F1* than the first next run from another team (the 3$^{\text{rd}}$ place). Also, it could be noticed that our system obtained 0.29 better *Accuracy* than the first place; however, *Accuracy* was not the reference metric for the second task. Regarding our runs, the system #3 obtained 0.28 better *Macro F1* than the single-model system (#1). Moreover, we observe that the other ensemble system achieved 2.05 less *Macro F1*

than our best run, which is a sensible difference in performance between both ensemble systems. System #2 also underperformed in comparison to the single-model system.

Table 8
Task 2: Sexism Categorization. Results of ELiRF-VRAIN in EXIST 2022. *Macro F1* (M-F1) is the reference metric for the second task.

System	Ranking	Team Rank.	*Accuracy*	*M-F1*	1st: *Acc.*	1st: *M-F1*	6th: *Acc.*	6th: *M-F1*
#1	4	-	70.13	49.63	-	-	-	-
#2	5	-	68.62	47.86	-	-	-	-
#3	3	2	70.42	49.91	70.13	51.06	66.07	47.47

8. Discussion

After analyzing the results, they showed that the data augmentation in the fine-tuning process had more impact on the final results than the ensemble approach. Although our best systems were both ensemble systems, the results of the single-model systems were not far from them. We expected to achieve better results with the ensemble system, an improvement that could justify the cost of training ten different models. However, in both tasks, the ensemble systems could not distance themselves from the performance of the single-model systems.

With those results, the first conclusion could be that the ensemble approach does not add any benefit to this problem. However, this can not be ensured, since the results of the ensembles are highly dependent on modeling aspects such as the performance of the models, and how their outputs are combined. So, other ensemble designs could be more beneficial. Further work is required to investigate the reasons of these results.

9. Conclusions

In this work, we have presented the participation of the ELiRF-VRAIN team in EXIST 2022 shared task. We have described the methodology to construct the six classification systems submitted to EXIST 2022, three for sexism identification and three for sexism categorization. We achieved the 6$^{\text{th}}$ place in the sexism identification task and the 2$^{\text{nd}}$ in the sexism categorization task.

We have detailed our approach to both EXIST 2022 tasks. We performed data augmentation actions that helped to increase the performance without additional external data. We enlarged the datasets by performing translation-based data augmentation and randomly masking tokens, which effectively increased the performance of our classification models.

We have also described the construction of our ensemble-based systems. In some cases, these systems increased the performance of the classification systems in comparison to the single-model systems. The improvements of the performance was below the initial expectations. Further work is required to study the reasons of these results.

In future works, we would like to improve our ensemble approach and study data augmentation strategies tailored to the sexism identification task.

Acknowledgments

This work is part of the AMIC-PoC project (PDC2021-120846-C44), funded by MCIN/AEI/10.13039/501100011033 and by the European Union "NextGenerationEU/PRTR". It is also partially supported by the Vicerrectorado de Investigación de la Universitat Politècnica de València (PAID-01-21 and PAID-11-21).

References

[1] W. Akram, A study on positive and negative effects of social media on society, International Journal of Computer Sciences and Engineering 5 (2018). doi:10.26438/ijcse/v5i10.351354.

[2] S. Siddiqui, T. Singh, Social media its impact with positive and negative aspects, International Journal of Computer Applications Technology and Research 5 (2016) 71–75. doi:10.7753/IJCATR0502.1006.

[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. doi:10.1109/ACCESS.2020.3042604.

[4] Sexism definition, 2022. URL: https://www.merriam-webster.com/dictionary/sexism.

[5] Glossary & thesaurus: Sexism, 2022. URL: https://eige.europa.eu/thesaurus/terms/1367.

[6] Sexism at work: Part 1. understand, 2022. URL: https://eige.europa.eu/publications/sexism-at-work-handbook/part-1-understand/what-sexism.

[7] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, T. Donoso, Overview of EXIST 2021: sEXism Identification in Social neTworks, Procesamiento del Lenguaje Natural 67 (2021) 195–207.

[8] F. Rodríguez-Sánchez, J. C. 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, volume 69, september 2022.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), Association for Computational Linguistics, 2019, pp. 4171–4186. URL: https://www.aclweb.org/anthology/N19-1423. doi:10.18653/v1/N19-1423.

[11] 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 (2019). arXiv:1907.11692.

[12] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.

[13] F. Barbieri, L. E. Anke, J. Camacho-Collados, XLM-T: A multilingual language model toolkit for twitter, CoRR abs/2104.12250 (2021). URL: https://arxiv.org/abs/2104.12250. arXiv:2104.12250.

[14] F. Barbieri, J. Camacho-Collados, L. Espinosa Anke, L. Neves, TweetEval: Unified benchmark and comparative evaluation for tweet classification, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 1644–1650. URL: https://aclanthology.org/2020.findings-emnlp.148. doi:10.18653/v1/2020.findings-emnlp.148.

[15] L. Taylor, G. Nitschke, Improving deep learning with generic data augmentation, in: 2018 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, 2018, pp. 1542–1547.

[16] C. Shorten, T. M. Khoshgoftaar, B. Furht, Text data augmentation for deep learning, Journal of big Data 8 (2021) 1–34.

[17] S. Chen, E. Dobriban, J. Lee, Invariance reduces variance: Understanding data augmentation in deep learning and beyond, ArXiv abs/1907.10905 (2019).

[18] T. Gonalves, P. Quaresma, Multilingual text classification through combination of monolingual classifiers, CEUR Workshop Proceedings 605 (2010).

[19] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[20] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022) 39–60. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6405.

[21] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[22] A. Abdaoui, C. Pradel, G. Sigel, Load what you need: Smaller versions of mutlilingual bert, in: SustaiNLP / EMNLP, 2020.

[23] T. Caselli, V. Basile, J. Mitrović, M. Granitzer, HateBERT: Retraining BERT for

abusive language detection in English, in: Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021), Association for Computational Linguistics, Online, 2021, pp. 17–25. doi:10.18653/v1/2021.woah-1.3.

[24] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, ALBERT: A lite BERT for self-supervised learning of language representations, CoRR abs/1909.11942 (2019). URL: http://arxiv.org/abs/1909.11942. arXiv:1909.11942.

[25] J. Tiedemann, The tatoeba translation challenge – realistic data sets for low resource and multilingual MT, in: Proc. of the 5th Conference on Machine Translation, ACL, 2020, pp. 1174–1182. URL: https://aclanthology.org/2020.wmt-1.139.