System Description for EXIST Shared Task at IberLEF 2021: Automatic Misogyny Identification Using Pretrained Transformers

Ignacio Talavera[1,2], David Fidalgo[1], and Daniel Vila-Suero[1]

[1] Recognai, Valencia, Spain. {david,daniel}@recogn.ai
http://www.recogn.ai
[2] Universidad Carlos III de Madrid, Madrid, Spain
100383487@alumnos.uc3m.es

Abstract. This shared task system description depicts two neural network architectures submitted to the EXIST task at IberLEF 2021, among them the twelfth classified in the second sub-task. We present in detail the approach and topologies used to obtain the two systems which we submitted. Both systems are based on pretrained language models and solve the two subtasks simultaneously, with the first system using different networks for English and Spanish and the second using a multilingual approach.

Keywords: Deep Learning · Misogynistic behaviours detection · Natural Language Processing · Sentiment Analysis.

1 Introduction

EXIST (sEXism Identification in Social neTworks)[16] is a shared task in Automatic Misogyny Identification in social networks at IberLEF 2021[14], a comparative evaluation campaign for Natural Language Processing Systems in Spanish and other Iberian languages. It aims to detect online proof of sexism in Spanish written language, which may help to determine the evolution of new equality policies in online environments, as well as to encourage better behaviours in society. AI and NLP researchers are working on Automatic Misogyny Identification (AMI) shared tasks like this one to distinguish misogynist contents from non-misogynous ones and to categorize their type [4, 7–9].

EXIST is divided into two subtasks:

– **Task 1: Sexism Identification.** It is a binary classification task, in which the system has to decide whether or not a given text extracted from Twitter or Gab is sexist.

IberLEF 2021, September 2021, Málaga, Spain.

– **Task 2: Sexism Categorization.** It is a multiclass classification task. The same texts analyzed in task 1 have to be classified into one of the five categories decided by the organization, which are **idelogical and inequality**, **stereotyping and dominance**, **objectification**, **sexual violence** and **misogyny and non-sexual violence** [16].

In this working notes we are going to explain our approach on this shared task and how we designed and trained the submitted models.

2 Our approach

We have submitted two systems capable of making predictions for the two subtasks. Both systems are based on pretrained Transformer models [17], and both were designed and trained using biome.text [1], a practical NLP open source library based on AllenNLP [10] and Pytorch [15].

These two systems were trained directly over the categories of the second task but were used to predict both tasks: if any of the categories of the second task surpassed a given threshold (independently calculated for each neural network), it is predicted as 'sexist' for the first task; otherwise, it is predicted as 'non-sexist'. The category of the second task is chosen as the output category from the neural network with the highest probability.

2.1 System 1

Our first system, denoted as Run 1 in the submitted results, has been designed using two Deep Neural Networks, one for English and one for Spanish Language. The Spanish Transformer-based language model was *BETO*, a BERT model trained on a big Spanish corpus [6], which is distributed via HuggingFace's [18] Model Hub under the name "*dccuchile/bert-base-spanish-wwm-cased*"; and the English Transformer-based language model was *Twitter-roBERTa-base Offensive Language Identification*, a roBERTa-base model trained on 58 million tweets and fine-tuned for offensive language identification [3]. It is also distributed via HuggingFace's Model Hub under the name "*cardiffnlp/twitter-roberta-base-offensive*".

Both neural networks were fine-tuned for the task, each one trained with the dataset corresponding to its language, and evaluated using the macro-averaged F-measure. The system was created combining those two networks in a basic decision tree: if the record of the test set to predict was in English, the English network is invoked to make the prediction; otherwise, the Spanish network was used.

2.2 System 2

Our second system, denoted as Run 2 in the submitted results, has been designed using one Deep Neural Network, following a multilingual approach. The

Transformer-based language model used was *twitter-XLM-roBERTa-base for Sentiment Analysis*, a XLM-roBERTa-base model trained on 198M tweets and fine-tuned for sentiment analysis [2]. It was fine-tuned over 8 languages (including English and Spanish). It can be found at HuggingFace's Model Hub under the name "*cardiffnlp/twitter-xlm-roberta-base-sentiment*".

This neural network was fine-tuned for the given task, using all records in the training dataset, and also evaluated using macro-averaged F-measure. At the end of the pipeline, this system was capable of predicting both English and Spanish input text.

3 Training

Both systems were trained using the same procedure, even though the hyperparameters obtained after optimizing each neural network and the thresholds used for predicting for each system were different.

Table 1. List of tuned hyperparameters during training. Search spaces define how hyperparameters were sampled initially, provided as Ray Tune search space functions.

Parameter	Search Space
Learning Rate	loguniform(5e-6. 1e-4)
Weight Decay	loguniform(1e-3, 1e-1)
Warmup Steps on Learning Rate Scheduler	randint(0,200)
Pooler Type	choice(["lstm", "gru"])
Hidden Size of Pooler's layers	choice([32, 64, 128, 256])
Number of Pooler's layers	choice([1, 2, 3])
Bidirectionality on Pooler's layers	choice([True, False])

For the parameter updates, we used the *AdamW* algorithm [13]. The parameters optimized can be seen in Table 1, along with their search spaces at the start of the hyperparameter process. These parameters were optimized by means of the Ray Tune Library [12], which is tightly integrated in *biome.text*.

Several Hyperparameter Optimization Processes (HPOs) were performed for each of the three neural networks, and each subsequent HPO fixed some parameters and reduced the search space for others, until we got the best-performing neural networks at the last HPO process. Spanish and Multilingual neural networks needed four HPO processes, and English neural network needed five HPO processes. The reference metric for all these processes was macro-averaged F-measure of Task 2. The training was done on a computer with 2 Tesla V100. These HPO processes included ASHA trial schedulers to terminate low-performing trials [11] and a tree-structured Parzen Estimator as search algorithm [5].

Once the best-performing models were obtained, a quick sweep across several random initialization seeds was performed, and then another sweep was made across different threshold values from 0.15 to 0.85, adding 0.05 in each step. The result of these last processes was the final model for the Spanish and English languages (which, together, compose System 1) and for the Multilingual approach (System 2).

In Table 2 we included the details of each of the three final models: the Spanish model, the English model and the Multilingual Model .

Table 2. Parameters of the best obtained models

Parameters	Spanish model	English model	Multilingual model
Learning Rate	$1.73 \cdot 10^{-5}$	$1.01 \cdot 10^{-5}$	$1.51 \cdot 10^{-5}$
Weight Decay	$4.97 \cdot 10^{-3}$	$7.77 \cdot 10^{-3}$	$7.44 \cdot 10^{-2}$
Batch Size	8	8	16
Warmup Steps (LR Scheduler)	12	91	14
Steps per epoch (LR Scheduler)	354	343	348
Pooler Type	gry	gru	gru
Hidden Size (Pooler)	128	128	64
Number of layers (Pooler)	1	1	1
Bidirectional (Pooler)	True	True	True
Threshold	0.5	0.55	0.5

4 Results

In Table 3 we present the evaluation metrics of both tasks for each of the submitted runs on the validation and the tests data sets, as well as the model size. Tables 4 and 5 show a comparison between the submitted runs, the best models of the shared task and the baselines models (provided by the organization), divided by tasks. System 1 obtained our highest score in both tasks. Our better model was System 1 (which made run 1), which was the twelfth classified for task 2 and the forty sixth for task 1. System 2 underpeformed System 1, being the thirty first classified on task 2 and the fifty sixth classified on task 1.

Both results obtained on Task 2 are close to the best ones of the competition, being 0.03 and and 0.09 F-measure points away from the winner, respectively. However, our results for Task 1 are significantly worse, which means that our initial premise (training a system to predict label and, if any label is predicted, to also predict 'sexist') was not effective.

Table 3. Competition results obtained and model size, divided by runs

Models	Task 1 Valid. (accuracy)	Task 2 Valid. (f-measure)	Task 1 Test (accuracy)	Task 2 Test (f-measure)	Model size (nr of params)
Spanish	0.751763	0.622708	0.751763	0.622708	$1.1 \cdot 10^8$
English	0.755814	0.563271	0.755814	0.563271	$1.3 \cdot 10^8$
Run 1 (Spanish + English)	0.753758	0.601608	0.753758	0.601608	$2.4 \cdot 10^8$
Run 2 (Multilingual)	0.762178	0.590333	0.762178	0.590333	$2.8 \cdot 10^8$

Table 4. Competition results of Task 1, compared to the two best models of the shared task and the baseline model

Ranking	Run	Accuracy	F-Measure
1	task1_AI-UPV_1	0,7804	0,7802
2	task1_SINAI_TL_1	0,78	0,7797
46	task1_recognai_1	0,7044	0,7041
52	Baseline_svm_tfidf	0,6845	0,6832
56	task1_recognai_2	0,6726	0,6717

Table 5. Competition results of Task 2, compared to the two best models of the shared task and the baseline model

Ranking	Run	Accuracy	F-Measure
1	task2_AI-UPV_1	0,6577	0,5787
2	task2_LHZ_1	0,6509	0,5706
12	task2_recognai_1	0,6243	0,55
31	task2_recognai_2	0,5996	0,5177
51	Baseline_svm_tfidf	0,5222	0,395

We also found that the multilingual approach simplified the training (we only had to train one pipeline instead of two) while obtaining good inference results. It did not reach the top performing models of the competition for the second task, and it performed even worse on task 1, but we find it a valid alternative to classic monolingual training.

5 Conclusions

To face this shared task, we designed two different systems with which we made the predictions that composed our two submitted runs. System 1 was designed with two Deep Neural Networks, one for English predictions (using

Twitter-roBERTa-base Offensive Language Identification as the pretrained language model) and one for the Spanish predictions (using *BETO* as the pretrained language model). In System 2 we followed a multilingual approach, using only one Deep Neural Network to make predictions in both English and Spanish (with *twitter-XLM-roBERTa-base for Sentiment Analysis* as the pretrained language model). Both systems followed a multilabel approach described in previous section, with which we were able to make prediction for Tasks 1 and 2 without making different pipelines.

We conclude that the exploitation of the transfer capabilities of a pretrained language model and its optimized fine tuning to the target domain provides a conceptually easy system architecture and seems to be the most straight forward method to achieve competitive performance, especially for tasks where training data is scarce. We also found that, for these types of competitions, creating a model for each subtask is the best-performing approach. Better results on task 1 could have been obtained if we had trained Deep Neural Networks on the binary classification task.

References

1. biome.text, https://www.recogn.ai/biome-text/v2.2.0/
2. Barbieri, F., Anke, L.E., Camacho-Collados, J.: XLM-T: A Multilingual Language Model Toolkit for Twitter. arXiv:2104.12250 [cs] (Apr 2021), http://arxiv.org/abs/2104.12250, arXiv: 2104.12250
3. Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., Neves, L.: TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 1644–1650. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.findings-emnlp.148, https://www.aclweb.org/anthology/2020.findings-emnlp.148
4. Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F.M., Rosso, P., Sanguinetti, M.: SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In: Proceedings of the 13th International Workshop on Semantic Evaluation. pp. 54–63. Association for Computational Linguistics, Minneapolis, Minnesota, USA (Jun 2019). https://doi.org/10.18653/v1/S19-2007, https://www.aclweb.org/anthology/S19-2007
5. Bergstra, J., Bardenet, R., Bengio, Y., Kégl, B.: Algorithms for hyper-parameter optimization. In: Proceedings of the 24th International Conference on Neural Information Processing Systems. pp. 2546–2554. NIPS'11, Curran Associates Inc., Red Hook, NY, USA (Dec 2011)
6. Canete, J., Chaperon, G., Fuentes, R., Pérez, J.: Spanish pre-trained bert model and evaluation data. PML4DC at ICLR **2020** (2020)
7. Carmona, M.A., Guzmán-Falcón, E., Montes, M., Escalante, H.J., Villaseñor-Pineda, L., Reyes-Meza, V., Rico-Sulayes, A.: Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in Mexican Spanish tweets (Aug 2018)

8. Fersini, E., Rosso, P., Anzovino, M.: Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In: Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), co-located with 34th Conf. of the Spanish Society for Natural Language Processing (SEPLN 2018). vol. 2150, pp. 214–228. Seville, Spain (Sep 2018)
9. Fersini, E., Nozza, D., Rosso, P.: Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI). In: Caselli, T., Novielli, N., Patti, V. (eds.) EVALITA Evaluation of NLP and Speech Tools for Italian : Proceedings of the Final Workshop 12-13 December 2018, Naples, pp. 59–66. Collana dell'Associazione Italiana di Linguistica Computazionale, Accademia University Press, Torino (Jun 2019), http://books.openedition.org/aaccademia/4497, code: EVALITA Evaluation of NLP and Speech Tools for Italian : Proceedings of the Final Workshop 12-13 December 2018, Naples
10. Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N., Peters, M., Schmitz, M., Zettlemoyer, L.: AllenNLP: A Deep Semantic Natural Language Processing Platform. arXiv:1803.07640 [cs] (May 2018), http://arxiv.org/abs/1803.07640, arXiv: 1803.07640
11. Li, L., Jamieson, K., Rostamizadeh, A., Gonina, E., Hardt, M., Recht, B., Talwalkar, A.: A System for Massively Parallel Hyperparameter Tuning. arXiv:1810.05934 [cs, stat] (Mar 2020), http://arxiv.org/abs/1810.05934, arXiv: 1810.05934
12. Liaw, R., Liang, E., Nishihara, R., Moritz, P., Gonzalez, J.E., Stoica, I.: Tune: A Research Platform for Distributed Model Selection and Training. arXiv:1807.05118 [cs, stat] (Jul 2018), http://arxiv.org/abs/1807.05118, arXiv: 1807.05118
13. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization. arXiv:1711.05101 [cs, math] (Jan 2019), http://arxiv.org/abs/1711.05101, arXiv: 1711.05101
14. Montes, M., Rosso, P., Gonzalo, J., Aragón, E., Agerri, R., Ángel Álvarez Carmona, M., Álvarez Mellado, E., de Albornoz, J.C., Chiruzzo, L., Freitas, L., Adorno, H.G., Gutiérrez, Y., Zafra, S.M.J., Lima, S., de Arco, F.M.P., Taulé, M. (eds.): Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) (2021)
15. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv:1912.01703 [cs, stat] (Dec 2019), http://arxiv.org/abs/1912.01703, arXiv: 1912.01703
16. Ródriguez-Sánchez, F., Carrillo-de Albornoz, J., Plaza, L., Gonzalo, J., Rosso, P., Comet, M., Donoso, T.: Overview of EXIST 2021: sEXism Identification in Social neTworks. Procesamiento del Lenguaje Natural **67**(0) (2021)
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need. arXiv:1706.03762 [cs] (Dec 2017), http://arxiv.org/abs/1706.03762, arXiv: 1706.03762
18. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv:1910.03771 [cs] (Jul 2020), http://arxiv.org/abs/1910.03771, arXiv: 1910.03771