Multiaztertest@Exist-Iberlef2022: Sexism Identification in Social Networks*

Kepa Bengoetxea[1,*,†], Amaia Aguirregoitia[1,†]

[1]University of the Basque Country (UPV/EHU), Spain

Abstract

The automatic identification of sexism and the categorization of the message in different types of sexism in social networks is the focus of the systems taking part in the EXIST-IberLEF 2022 shared task. Detecting online sexism and encouraging better behaviours in society is a fundamental duty since the Internet is not safe from many of the misconducts against women. This paper explores a proposal by the Multiaztertest team that combines Bert and Roberta to fight discrimination against women. The best approach developed by this team in Spanish to the sexism identification task is based on pre-trained xlm-roberta-large and it has been fine-tuned only on the provided sexism identification labels EXIST dataset obtaining an accuracy of 0.7744 in the test data. Moreover, in English, the system uses pre-trained roberta-large which has been fine-tuned only on the provided sexism categorization labels EXIST dataset obtaining an accuracy of 0.8023 in the test data. Regarding the sexism categorization task, the best approach is based on pre-trained BETO in Spanish and pre-trained roberta-large in English and both have been fine-tuned on the provided sexism categorization labels EXIST dataset obtaining a macro-average F-measure of 0.4867 and 0.5337 respectively in the test data. Overall, the results confirm that these challenging tasks have been reasonably well addressed by the proposed tools.

Keywords

Sexism detection, Exist-IberLEF, Language Models

1. Introduction

Sexism is defined by the Oxford English Dictionary as “prejudice, stereotyping or discrimination, typically against women, on the basis of sex”. Nowadays women rely on social media platforms like Twitter to communicate, access information and gain visibility. However, they are not free from abusive conducts and the Internet has been reported as a “toxic place” for women and a place that promotes "sexism". Previous research has addressed the topic of sexism in the online world and has confirmed its presence on Twitter [1, 2, 3].

Given the relevance of this social problem, the sEXism Identification in Social neTworks (EXIST) shared task has been proposed at IberLEF 2021 [4] and IberLEF 2022, which aims at identifying and classifying tweets and gabs that contain sexist expressions or behaviours, both in Spanish and English.

IberLEF 2022, September 2022, A Coruña, Spain
*Corresponding author.
†These authors contributed equally.
kepa.bengoetxea@ehu.eus (K. Bengoetxea); amaia.aguirregoitia@ehu.eus (A. Aguirregoitia)
0000-0002-0289-6897 (K. Bengoetxea); 0000-0003-2465-5768 (A. Aguirregoitia)

The sexism identification task is defined as a binary classification problem, where every system should determine whether a text or message is sexist or not. Also, sexism categorization of tweets and gabs is defined as a multi-class classification problem where each sexist tweet or gab must be categorized as ideological and inequality, stereotyping and dominance, objectification, sexual violence, and misogyny and non-sexual violence.

In the first EXIST-IberLEF shared task [5] more than 70 models for the sexism identification task and 61 for the sexism categorization challenge were presented, which might assist human moderators to reduce the volume of sexist content in social networks. The majority of participants exploited transformer-based systems for both tasks. The best performing team was AI-UPV [6], which achieved an accuracy of 0.7804 in the sexism identification task and a F1-macro of 0.5787 in the sexism categorization task using an ensemble of different transformer models with BERT for English, BETO for Spanish and mBERT for multilingual models. They also implemented monolingual models with translation for both English and Spanish texts. The monolingual without translation model presented better results for English and Spanish than the monolingual with translation in the sexism identification, but the monolingual with the translation model presented better results for English and Spanish than the monolingual without translation in the sexism categorization task. Our team [7] got the fourth position in the sexism identification task obtaining an accuracy of 0.774 combining a monolingual BERT transformer model for English and BETO for Spanish.

In this paper, we present our experience testing BERT [8] (or multilingual BERT - mBERT), a Spanish version of BERT called BETO [9], RoBERTa [10] and a multilingual version of RoBERTa called XLM-R [11] in the sexism identification task (task 1) and the sexism categorization (task 2) in the context of the EXIST-IberLEF 2022 Shared Task [12].

This paper is structured as follows: in Section 2, we describe our approaches and the experimental setup; in Section 3, we present the results and in Section 4 we summarise the conclusions of the study and outline the future work.

2. Methodology

In this section, we describe our approaches and the experimental setup using the dataset provided by the organizers for both tasks.

The system has been implemented using Python on a Google Collaboratory Pro (https://colab.research.google.com/) TPU with the following technical specifications: Intel(R) Xeon(R) CPU @ 2.30GHz CPU, 26GB of RAM, and TPU v2. We have used the PyTorch framework to create our model.

We have shared the best language models and the code to use them in the GitHub repository (https://github.com/kepaxabier/exist2022).

2.1. Dataset

The dataset we have used has been provided by the organisers [12]. The EXIST dataset consists of 11345 tweets for training (of which 5644 are in English and 5701 are in Spanish) and 1058 tweets for testing. Organisers discarded posts in both data sources due to several reasons: posts written in another language, messages containing only hashtags or URLs, etc. Emojis were also removed since Mturk does not support them.

2.2. Evaluation Measures

The results are calculated using accuracy for task 1 considering that the distribution between sexist and non-sexist labels is balanced. Additionally, for task 2, the macro average F-measure is calculated for unbalanced labels and the same importance of classes.

2.3. Language Models

Many strategies for training general-purpose language representation models using the immense amounts of unannotated text are available on the web (this is known as pre-training). When working with challenges like sexism detection, these general-purpose pre-trained models can subsequently be fine-tuned on smaller task-specific datasets. When compared to training on smaller task-specific datasets from scratch, this strategy produces significant accuracy improvements.

In this research, we have experimented with the following pre-trained transformer models provided by Hugging Face [13]: (i) the following BERT [8] models for English: bert-base-uncased and bert-large-uncased; (ii) the following RoBERTa [10] models for English: roberta-base and roberta-large; (iii) the following BETO [9] models for Spanish: bert-base-spanish-wwm-uncased and bert-base-spanish-wwm-cased; (iv) the following multilingual BERT [8] models for Spanish: bert-base-multilingual-uncased and bert-base-multilingual-cased; and finally (v) the following multilingual RoBERTa [11] models for Spanish: xlm-roberta-base and xlm-roberta-large.

2.3.1. BERT, mBERT and BETO

BERT is a transformers model pre-trained on the BooksCorpus [14] and English Wikipedia for English in a self-supervised fashion. It was pre-trained on the raw texts only, with no humans labelling them in any way, with two objectives:

- Masked language modeling (MLM): taking a sentence, the model randomly masks 15% of the words in the input then run the entire masked sentence through the model and has to predict the masked words.
- Next sentence prediction (NSP): the models concatenates two masked sentences as inputs during pretraining. Sometimes they correspond to sentences that were next to each other in the original text, sometimes not. The model then has to predict if the two sentences were following each other or not.

The BERT raw model can be used for masked language modeling or next sentence prediction, but this model is primarily aimed at being fine-tuned on tasks that use the whole sentence (potentially masked) to make decisions, such as sequence classification, token classification or question answering.

The BERT model is available in two versions BERT base and BERT large. BERT base has 12 transformer blocks, 768 hidden layers, 12 bidirectional self-attention heads and 110 Million parameters. BERT large has 24 transformer blocks, 1024 hidden layers, 16 bidirectional self-attention layers and 340M parameters.

A single sentence was fed into the model at a time. The input sentences were split into tokens and mapped to their indexes using the BERT tokenizer library, indicated as input_ids. The [CLS] (classification token) and [SEP] (separate segment token) were appended at the beginning and end of every sentence, respectively. An input attention mask of fixed length with 0 indicating padded tokens, and 1, indicating unpadded tokens was applied. Each of the transformers indicated received a list of token embeddings and produced a feature vector of the same length at the output. The output of [CLS] for the last transformer layer containing vector transformations of prediction probabilities was used as aggregated sequence representation from which classifications were made.

mBERT is based on the original BERT model with 12 self-attention layers, 12 attention heads each, a hidden size of 768 and a total of 178M parameters. It was pretrained on 104 languages with the Wikipedia dataset.

According to the authors, the BETO model has 12 self-attention layers, 16 attention heads each, a hidden layer of size 1024, and a total of 110M parameters. However, the actual version uploaded to HuggingFace has a BERT-base-like architecture with 12 self-attention layers, 12 attention heads each, a hidden size of 768, and a total of 110M parameters. It was pretrained with text from different sources: all the Spanish data from Wikipedia and the Spanish portion of the OPUS project.

2.3.2. RoBERTa and XLM-RoBERTa

Roberta is a pretrained model on English language using a masked language modeling (MLM) objective. The RoBERTa model was pretrained on the BookCorpus, English Wikipedia, CC-News, OpenWebText and Stories datasets. The size of the full dataset is about 160 GB of text. The configuration of the base and large versions is as follows:

- The RoBERTa-base model was made up of 12 transformer layers with 768-hidden layers, 12 attention heads, and 125 million parameters used in the experiment.
- The RoBERTa-large model was made up of 24 transformer layers with 1024-hidden layers, 16 attention heads, and 355 million parameters used in the experiment.

The RoBERTa tokenizer was used to encode the input texts into tokens and designated them as the input_ids. The [CLS] (classification token) and [SEP] (separate segment token) were appended at the beginning and end of every sentence, respectively. An input attention mask of fixed length with 0 indicating padded tokens, and 1, indicating unpadded tokens was applied. Each of the transformers indicated received a list of token embeddings and produced a feature vector of the same length at the output. The output of [CLS] for the last transformer layer containing vector transformations of prediction probabilities was used as aggregated sequence representation from which classifications were made.

XLM-RoBERTa (XLM-R) is a multilingual version of RoBERTa. XLM-R is a transformer-based multilingual masked language model pre-trained on hundred languages, using more than 2 Terabytes of filtered CommonCrawl [15] data. We used Hugging Face's implementation of XLM-

RoBERTa model, which inherits the XLM training method and draws on the ideas of RoBERTa. We also tried to use different size models: XLM-RoBERTa-base and XLM-RoBERTa-large.

2.4. Experimental Setup

For both tasks, we have separated the dataset between languages (English and Spanish). We have truncated all texts that had more than 200 tokens (for BERT) and 128 tokens (for Roberta) and we have added two tokens to mark the beginning and the end of the sequence to each input text, [CLS] and [SEP] respectively. We have padded texts shorter than 200 or 128 tokens with zeroes. We have not performed any text augmentation or pre-processing besides standard byte-pair encoding. We split the training data into 80% for train and 20% for validation. The validation data was applied to determine the model selection and a set of hyper-parameters.

In the fine-tuning stage, we adjusted several hyperparameters to improve our trained model and obtain better results. We used the following set of hyper-parameters to train models:

- The output of [CLS] for the last transformer layer containing vector transformations of prediction probabilities was used as aggregated sequence representation from which classifications were made. We probed with two sequential models: i) a Linear layer, ReLU activation function and Linear layer model. The input dimension of the first Linear layer was 768 for the base type model and 1024 for the large type model, and the output dimension of the first Linear layer was always 50 while the input dimension of the second Linear was 50 and the output dimension was 2 or 6 depending on the number of classes (named A Classifier in this research); ii) a Linear layer, a Dropout layer to fight overfitting and Linear layer model. The input and output dimension of the first Linear layer was 768 for the base type model and 1024 for the large type model. The Dropout probability was equal to 0.1. On top of the Dropout Layer, a Linear layer was added. The input dimension of the Linear layer was 768 or 1024 and the output was 2 or 6 depending on the number of classes (named B Classifier in this research)
- The training batch size was made equal to 16, 32 or 64 for BERT and 8 for Roberta
- The model was trained for 10 epochs using early stopping technique
- We set the max length as 200 for BERT and 128 for Roberta
- The learning rate to 5e-6, 5e-5, 5e-4 and 5e-3. The learning rate that controls how much the model changes in response to the estimated error each time the model weights are updated. Choosing the optimal learning rate is a difficult task, as a small learning rate may result in a slow training process and a value that is too large can cause the model to diverge instead of converging to the solution.
- We use the cross entropy loss function
- We use Adam optimizer [16] to update the parameters

We have selected the best monolingual models using the best hyperparameters in the validation data.

Table 1 provides the most relevant information about the best monlingual models obtained using the validation data. In the first row, for fine-tuning M1 model on English and task 1 labels, we add B classifier (see Subsection 2.4) on top of Roberta Large model with the hyperparameters

Table 1
Information of the best monolingual models

Id.	trained on	Model	Best hyperp. values
M1	EN and Task1 labels	Roberta Large	Mtl:128; Ne:2; Lr:5e-06; Bs=8; C=B
M2	ES and Task1 labels	XLM-R Large	Mtl:128; Ne:3; Lr:5e-06; Bs=8; C=B
M3	EN and Task2 labels	Roberta Large	Mtl:128; Ne:2; Lr:5e-06; Bs=8; C=B
M4	ES and Task2 labels	BETO Base	Mtl:200; Ne:3; Lr:5e-05; Bs=32; C=A

specified in the Table 1. The abbreviations we use are: Mtl: max tweet length, Lr: learning rate, Bs: batch size, Ne: number of epochs and C: classifier.

In the second row, for fine-tuning M2 model on Spanish and task 1 labels, we add B classifier (see Subsection 2.4) on top of XLM-R Large model with the hyperparameters specified in Table 1.

In the third row, for fine-tuning M3 model on English and task 2 labels (including non-sexist texts), we add B classifier (see Subsection 2.4) on top of Roberta Large model with the hyperparameters specified in Table 1.

Finally, the fourth row describes M4 model. For fine-tuning M4 model on Spanish and task 2 labels (including non-sexist texts), we add A classifier (see Subsection 2.4) on top of BETO base model with the hyperparameters specified in Table 1.

3. Results

This section contains the main research results and is divided into 2 subsections: 3.1 contains a description of the best monolingual approaches used and 3.2 contains the comparison between the final multilingual approaches on the test subset.

All the results have been obtained using only the dataset provided by the organizers for both tasks and, later, during the training phase, the validation set was merged with the training set.

3.1. Results of Monolingual Approaches

In Table 2 we present the results of our best monolingual approaches and for comparison purposes, we include the data of the best system in each task. We participate with one run in task 2 and two runs in task 1.

The second row presents our best results obtained for task 2 in English using M3 model (Roberta Large model). The value obtained for macro f-measure is 0.4689, which was rated in twelfth position. The first position in the ranking obtained a value of 0.5337 for macro f-measure.

The fourth row details our best results obtained for task 2 in Spanish using M4 model (BETO Base model). The value obtained for macro f-measure is 0.4679, which was rated in twelfth position. The first position of the ranking obtained a value of 0.4867 for macro f-measure.

In Spanish, for the first task, we have participated with two runs: an XLM-R Large model trained on task 1 labels and a BETO Base model trained on sexism categorization labels and incorporating post-processing that turned any sexism category into a "sexist" label. The sixth row details our best results obtained for task 1 in Spanish using M2 model (XLM-R Large model)

Table 2
Results of monolingual approaches in the test data

Rank	Team	Task	Lang.	Model	Accuracy	Precision	Recall	$F1$
1	avacaondata_1	task2	EN	-	0.7471	0.6184	0.5532	0.5337
12	multiaztertest_1	task2	EN	M3	0.7110	0.5520	0.4789	0.4689
1	ELiRF-VRAIN_3	task2	ES	-	0.6786	0.5891	0.4881	0.4867
12	multiaztertest_1	task2	ES	M4	0.6466	0.5457	0.4863	0.4679
1	CIMATCOLMEX_1	task1	ES	-	0.7801	0.7808	0.7805	0.7801
2	multiaztertest_1	task1	ES	M2	0.7744	0.7753	0.7749	0.7744
20	multiaztertest_2	task1	ES	M4	0.7444	0.7443	0.7441	0.7442
1	avacaondata_3	task1	EN	-	0.8422	0.8388	0.8365	0.8376
10	multiaztertest_2	task1	EN	M3	0.8023	0.7981	0.794	0.7958
12	multiaztertest_1	task1	EN	M1	0.7928	0.7888	0.7951	0.7901

trained on task 1 labels. The value obtained for accuracy was 0.7744, and this system ranked second best Spanish model. The system in the first position obtained an accuracy value of 0.7801. The seventh row shows the results obtained for task 1 in Spanish using M4 model (BETO Base model) and applying sexism categorization labels. The accuracy value is 0.7444, which placed the system twentieth.

As to English language, for task 1, we participated with two runs too, with Roberta Large model trained on task 1 and task 2 labels. The ninth row presents our best results obtained for task 1 in English using M3 model (Roberta Large) trained on sexism categorization labels and incorporating post-processing that turned any sexism category into a "sexist" label. Accuracy is at 0.8023, which placed the system in the tenth position whereas the accuracy of the best system is at 0.8422. Finally, the tenth row shows the second-best result for task 1 in English produced using M1 model (Roberta Large model) trained on task 1. The accuracy is at 0.7928 which was rated in twelfth position.

3.2. Results of the Multilingual Approaches

For a multilingual approach, two multilingual models were proposed for task 1 and one multilingual model was presented for task 2.

Regarding task 1, we had four monolingual models and, therefore, four multilingual models could have been created by combining an English and a Spanish model as follows: M3-M2, M1-M2, M3-M4 and M1-M4. but only two different approaches were presented. On the one hand, the first approach combined M3 and M4 models, which are the best models obtained for task 1 using task 1 labels. On the other hand, the second one combined M1 and M2 models, which are the best models for task 2 trained on task 2 labels, and, besides, it incorporated post-processing that turned any sexism category into a "sexist" label.

Concerning task 2, M1 and M2, which are the best models obtained for this task, have been combined.

In Table 3 we present the results of our best multilingual approaches with the results provided by the organisers. For comparison purposes, we include the data of the best system in each task and two baselines proposed by organizers.

Table 3
Results of the multilingual approaches in the test data

Rank	Team	Task	Model	Accuracy	Precision	Recall	$F1$
1	avacaondata_1	task2	-	0.7013	0.5907	0.5351	0.5106
9	multiaztertest_1	task2	M1-M2	0.6786	0.5451	0.4826	0.4706
26	organizers	task2	baseline	0.5784	0.4299	0.3395	0.342
30	organizers	task2	Majority Class	0.5539	0.5539	0.1429	0.1018
1	avacaondata_1	task1	-	0.7996	0.7982	0.7975	0.7978
9	multiaztertest_1	task1	M1-M2	0.7836	0.7831	0.7853	0.783
12	multiaztertest_2	task1	M3-M4	0.7110	0.5520	0.4789	0.4689
39	organizers	task1	baseline	0.6928	0.6919	0.685	0.6859
44	organizers	task1	Majority Class	0.5444	0.5444	0.5	0.3525

Organizers proposed two different baselines so that we could establish the expected performance of the submitted runs. Firstly, they provided a benchmark (Baseline svm tfidf) based on Support Vector Machine (linear kernel) trained on tf-idf features built from the texts unigrams. Secondly, they supplied a model that labels each record based on the majority class (Majority Class).

Our best multilingual approaches were ranked ninth in both tasks.

In task 2, we combined a pre-trained Roberta Large model for English and BETO for Spanish, which obtaining a macro f-measure of 0.47. Remarkably, both approaches perform similarly in macro f-measure. The highest-rated system obtained a value of 0.5106 for macro f-measure.

In task 1, we combined a pre-trained Roberta Large model for English and XLM-R Large for Spanish, and the accuracy obtained was 0.7836. In this case, the English model results are 2 points above the Spanish model. The accuracy of the best-rated system was at 0.7996. Remarkably, all our models are above the baseline and majority class systems.

4. Conclusion and Future Work

In this paper, we have presented the results of the MultiAzterTest team at the first and second tasks of the Exist-IberLEF 2022 shared task. These attempts are aimed at exploring the different types of monolingual Bert and Roberta models to fight inequality and discrimination against women. The combination of the best monolingual approaches in Spanish and English confirms that sexism detection in social networks has been reasonably well addressed. While in task1 our best system obtained 0.7836 in accuracy the best model presented in the shared task 2022 obtained an accuracy of 0.7996, only +0.016 more than our system. In task 2, the difference between our best system (0.4706 in macro f-measure) and the best system presented (0.5106 in macro f-measure) was +0.04 more than our system.

However, in the future, we intend to experiment with ensemble models to provide a better generalization such as the weighted average ensemble method.

Additional models such as multimodal transformers could also be implemented to incorporate tabular data with HuggingFace transformers. Therefore, we could combine the output of a HuggingFace transformer with the output of other tools such as MultiAzterTest [17], which analyses 163 linguistic and stylistic features in English, 141 in Spanish and 125 in Basque or MultiAzterTest-Social [7], which includes features to analyse social media texts inspired by Fersini et al. [18] and other improvements covering 280 features for English and 244 for Spanish.

Additionally, the combination of two or more HuggingFace transformers using a simple linear layer on top of them could be implemented. Since in many cases one single model might not give the best results, we could intend to combine several "weak" classifiers together and somehow combine the results of each of these weak classifiers in a meaningful way to achieve improved results.

Acknowledgments

This work has been partially funded by the DeepText (KK-2020/00088) project.

References

[1] A. Jha, R. Mamidi, When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data, in: Proceedings of the second workshop on NLP and computational social science, 2017, pp. 7–16.

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

[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] M. Montes, P. Rosso, J. Gonzalo, E. Aragón, R. Agerri, M. Á. Álvarez Carmona, E. Álvarez, J. Carrillo-de Albornoz, L. Chiruzzo, L. Freitas, H. Gómez, Y. Gutiérrez, S. M. Jiménez, S. Lima, F. M. Plaza-de Arco, M. Taulé, Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021, CEUR Workshop Proceedings, 2021.

[5] 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).

[6] 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, arXiv preprint arXiv:2111.04551 (2021).

[7] K. Bengoetxea, I. Gonzalez-Dios, MultiAzterTest@Exist-IberLEF 2021: Linguistically Motivated Sexism Identification, 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), volume 2943, CEUR-WS, 2021, pp. 449–457. URL: http://ceur-ws.org/Vol-2943/.

[8] 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), 2019, pp. 4171–4186.

[9] 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.
[10] 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).
[11] 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.
[12] 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).
[13] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., HuggingFace's Transformers: State-of-the-art natural language processing, arXiv preprint arXiv:1910.03771 (2019).
[14] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, S. Fidler, Aligning books and movies: Towards story-like visual explanations by watching movies and reading books, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 19–27.
[15] G. Wenzek, M.-A. Lachaux, A. Conneau, V. Chaudhary, F. Guzmán, A. Joulin, E. Grave, Ccnet: Extracting high quality monolingual datasets from web crawl data, arXiv preprint arXiv:1911.00359 (2019).
[16] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).
[17] K. Bengoetxea, I. Gonzalez-Dios, MultiAzterTest: a Multilingual Analyzer on Multiple Levels of Language for Readability Assessment, Manuscript from author, 2021.
[18] E. Fersini, D. Nozza, G. Boifava, Profiling Italian Misogynist: An Empirical Study, in: Proceedings of the Workshop on Resources and Techniques for User and Author Profiling in Abusive Language, 2020, pp. 9–13.