Transfer Learning for Automatic Sexism Detection with Multilingual Transformer Models

AIT_FHSTP@EXIST2022

Daria Liakhovets[1,*,†], Mina Schütz[1,†], Jaqueline Böck[2,†], Medina Andresel[1], Armin Kirchknopf[2], Andreas Babic[2], Djordje Slijepčević[2], Jasmin Lampert[1], Alexander Schindler[1] and Matthias Zeppelzauer[2]

[1] *Austrian Institute of Technology, 1210 Vienna, Austria*

[2] *St. Pölten University of Applied Sciences, 3100 St. Pölten, Austria*

Abstract

In recent years sexism has become an increasingly significant problem on social networks. In order to address this problem, the sEXism Identification in Social neTworks (EXIST) challenge has been launched at IberLEF in 2021. In this international benchmark, sexism detection is formulated as a Natural Language Processing (NLP) task with the aim to automatically identify sexism in social media content (binary classification) and to classify statements into different categories such as dominance, stereotyping or objectification. In this paper we present the contribution of team AIT_FHSTP for the EXIST challenge at IberLEF in 2022. To solve the two related tasks we applied two multilingual transformer models, one based on a multilingual BERT and one based on an XLM-RoBERTa architecture, and a monolingual (English) T5 model. Our approach uses two different strategies to adapt the transformers to the detection of sexist content: first, unsupervised pre-training with additional data and second, supervised fine-tuning with additional as well as augmented data. For both tasks the XLM-RoBERTa model, which applies a combination of the two strategies, outperforms the other two models. The best run for the binary classification (task 1) achieves a macro F1-score of 74.96% and scores the 26^{th} rank in the benchmark; for the multi-class classification (task 2) our best submission scores the 13^{th} rank with a macro F1-score of 46.75%.

Keywords

Sexism Detection, Sexism Identification, Social Media Retrieval, Transformer Models, Natural Language Processing,

1. Introduction

Discriminatory views and statements in particular against women are unfortunately a common phenomenon in online and social media. Typically, this is often an indica25tion of other toxic content categories such as hate speech [1] or disinformation [2]. Detection of such comments is

IberLEF 2022, September 2022, A Coruña, Spain.

[*] Corresponding author.

[†] These authors contributed equally.

daria.liakhovets@ait.ac.at (D. Liakhovets); mina.schuetz@ait.ac.at (M. Schütz); jaqueline.boeck@fhstp.ac.at (J. Böck); medina.andresel@ait.ac.at (M. Andresel); armin.kirchknopf@fhstp.ac.at (A. Kirchknopf); Andy.babic@bandy.at (A. Babic); djordje.slijepcevic@fhstp.ac.at (D. Slijepčević); jasmin.lampert@ait.ac.at (J. Lampert); alexander.schindlerait.ac.at (A. Schindler); matthias.zeppelzauer@fhstp.ac.at (M. Zeppelzauer)

challenging, since sexism and misogyny may appear in various forms and differ over language and cultural barriers. The shared task on sEXism Identification in Social neTworks (EXIST) at IberLEF 2022 [3] provides a systematic benchmark for the structured evaluation of machine learning and/or natural language understanding approaches. The benchmark covers a wide spectrum of sexist content and aims to differentiate different types of sexism. It incorporates English and Spanish content from Twitter and Gab, as well as the categorization that was provided by experts in gender issues. This paper presents our contribution to the benchmark, describes our approach and summarises the obtained results for both tasks, i.e., the binary sexism identification task (task 1) and the sexism categorization task (task 2). Our specific approach is characterized by a comprehensive use of data augmentation and the integration of external (unlabeled) data to make the classification models more robust. In order to account for the bilingual dataset we employ multilingual models (except for T5, which is monolingual). Our paper is structured as follows: Section 2 describes our methodological approach with a focus on the employed datasets and models. Our experimental setup is outlined in Section 3, which is followed by the presentation of the results (Section 4) as well as the discussion and final conclusions (Section 5).

2. Method

The methodological approach follows the same structure as presented in [4] and is based on the EXIST2022 benchmark dataset, which contains 11,345 instances of annotated social media messages. The EXIST2022 dataset is approximately 4,000 instances larger than the previous dataset from 2021. Nevertheless, the dataset is rather small for training complex NLP models. This circumstance is mitigated by applying pre-trained transformers, which have been subsequently fine-tuned on the EXIST2022 dataset. Due to this approach, we achieve competitive results even on low-resource data. Following our successful strategy of the last challenge ([4]), we employ this time different transfer learning strategies. We apply three pre-trained transformer models: multilingual BERT (mBERT) [5], XLM-RoBERTa (XML-R) [6] and T5 [7]. Our last year's contribution to the benchmark yielded the 3^{rd} best team result with different data augmentation strategies [4]. This year we adapt these strategies and enhance them with an even larger set of extended datasets in the field of hate speech and sexism detection for pre-training and fine-tuning transformer models. In this paper we use the term "pre-training" to refer to the unsupervised re-training of a transformer model and "fine-tuning" to refer to the supervised training for the specific classification task. Our respective strategies are the following:

- **Pre-training strategy:** Pre-trained models, which are usually pre-trained on large generic datasets, often overfit on small datasets [8]. Therefore, we experiment with pre-training available models with task-specific data.

- **Fine-tuning strategy:** To use the pre-trained models for the given downstream task, the models have to be trained (i.e. fine-tuned) on labeled task-specific data. This can be done either on only the upper layers or all layers of the transformer model.

Table 1
Additional datasets used for pre-training and/or fine-tuning

dataset name	pre-trainig	fine-tuning	Source
EXIST2022 (equals EXIST2021 training data)	YES	YES	[9]
Call me sexist but (CMSB)	YES	YES	[10]
Sexismo en código binario (SCB)	YES	YES	[11]
MeTwo	YES	YES	[12]
SexismOnTwitch (SOT)	YES	NO	[13]
Sexist Stereotype Classification (SSC)	YES	NO	[14]
Sexist Workplace Statements (ISEP)	YES	NO	[15]
Urban Dictionary definitions dataset for misogyny speech detection (MTM)	YES	NO	[16]
HatEval2019	YES	NO	[17]

A more detailed description of the implementation of the strategies is provided in Section 3 and the results obtained with these approaches are presented in Section 4.

2.1. EXIST2022 Data

The *EXIST2022* dataset includes postings from social media platforms such as Twitter and Gab, as well as annotations for different categories of sexism. The dataset is split into train and test partitions. The training set consists of 11,345 instances in English (5.644) and Spanish (5.701) language. The test set contains 1.058 instances (526 English and 532 Spanish postings). Each data instance is assigned a binary label (for task 1) indicating whether it is *sexist* or *non-sexist*. In addition, a multi-class categorization is provided for task 2: *ideological-inequality*, *objectification*, *stereotyping-dominance*, *misogyny-non-sexual-violence*, *sexual-violence*, *non-sexist*. We additionally augmented the EXIST2022 dataset and additional datasets by translating each post into the respective other language (i.e., from English to Spanish and vice versa). Due to this procedure, an English and a Spanish version of each dataset was created. The online tool DeepL (https://www.deepl.com/translator) was used for this purpose.

2.2. External Data

Data augmentation is one of the two strategies being pursued with our contribution to the challenge. In addition to the *EXIST2022* dataset provided by the organizers, we pre-train different models on additional datasets which are semantically related to the *EXIST2022* dataset. The intention is to learn additional patterns from semantically similar texts and tasks and to transfer them to the EXIST tasks. We conduct experiments using several additional datasets for pre-training and fine-tuning, which are shown in Table 1.

2.2.1. Datasets for Pre-Training and Fine-Tuning

- **CMSB** is an English dataset that is an aggregation of three different datasets, consisting of three types of content: social media posts (tweets), psychological survey items and synthethic adversarial modifications of both. More precisely, the Twitter data can be divided into three further datasets including: *the hostile sexism dataset* [18], *the benevolent sexism dataset* [19] and the *call me sexism dataset* [20] which has been collected by the authors. The entire dataset contains 13,634 tweets.

- **SCB** the information on this Spanish dataset consisting of about 5520 instances, was retrieved from Twitter and then manually classified according to the presence of violent and misogynistic content.

- **MeTwo** is a Spanish dataset that consists of 3,600 tweets that can be used to detect sexist innuendo, behaviors, and expressions. The labels of the tweets are: *SEXIST*, *NON_SEXIST* and *DOUBTFUL*. The original dataset consists of tweet-IDs labeled as *status_id* and the associated label for the category. Content and metadata of the corresponding tweets were provided by the creator of the dataset upon request.

2.2.2. Datasets for Pre-Training Only

- **SOT** is a Spanish dataset that contains comments which were scraped from Twitch using the Twitch API for Python. For this dataset, a list of female Twitch streamers was selected based on the followers and the topic the streamer was dealing with. The data got first classified into "innocuous" and "inappropriate" tweets. In a second step, the inappropriate data got further divided into the classes *love-stuck* and *strongly sexist.* Approximately 300 tweets have been classified as inappropriate and 3,000 were innocuous. Inappropriate tweets consist of about 50% *love-stuck* and 50% *strongly-sexist* tweets.

- **SSC** is an English dataset consisting of Instagram posts scraped searching for the hashtags "bloodymen", "boys", "everydaysexism", "girls", "guys", "manspalining", "metoo", "sexism", "sexist" and "slutshaming". The data got annotated into sexist and non-sexist statements. A text was classified as sexist if it was about sexism (for example: "I was told to shut up because women don't know science") or was sexist itself (for example: "Women belong in the kitchen"). The annotated dataset includes 6,238 posts.

- **ISEP** is an english dataset that contains more than 1100 examples of sexist comments in the workplace, which more or less balance between unambiguous sexist comments and ambiguous or neutral cases (which are labeled with a *1* or *0* respectively).

- **MTM** is an English dataset that was gathered from 1999 to 2006 from the Urban Dictionary platform. It is composed of 2,285 definitions which were categorized as *misogynistic* and *non-misogynistic.*

- **HatEval2019** is a dataset that can be used for detecting hate speech against women and immigrants. It is composed of 13,000 English tweets and 6,000 Spanish tweets. From a total of 19,600 tweets, 9,091 have a negative relation towards immigrants and 10,509 against women. Furthermore, the tweets are divided into 3 categories: 1) *Hate Speech* (against women or immigrants), 2) *Target Range* (against a generic group or individual), and 3) *Aggressiveness.*

2.2.3. Twitter Dataset for mBERT Only

Specifically for the training of the mBERT model, we extracted a total of about 40 million messages in English and Spanish from the full COVID-19 Twitter stream (https://developer.twitter.com/en/docs/twitter-api/tweets/covid-19-stream/overview), which Twitter made accessible for the research community in response to the pandemic in March 2020. With the help of hashtag filters related to COVID-19 and the Coronavirus, the stream was designed such that it follows the ongoing discussions in real-time. In order to obtain a comparable dataset, we filtered out similar hashtags as were used for the EXIST2022 challenge dataset and collected a total of 40 million tweets. Filtering reduced the number of tweets further to about 10 million.

2.3. Models

To model the textual data, we employ three different transformers [21]: multilingual BERT (mBERT), XLM-RoBERTa (XLM-R) and T5:

- **mBERT** is a multilingual transformer model based on the original BERT (Bidirectional Encoder Representations from transformers) architecture [22]. However, in comparison to the original transformer [21], BERT is only based on the encoder part of the transformer and does not make use of the decoder. BERT is pre-trained on data from Wikipedia and the BookCorpus and can be fine-tuned with two methods: with capturing a sentence in a bidirectional way with the attention mechanism, either with Masked Language Modelling or Next Sentence Prediction. The multilingual version of BERT has been additionally trained with Wikipedia content in 100 languages [23].

- **XLM-R** is also a multilingual model and, similarly to mBERT, is trained on 100 languages with a dataset containing CommonCrawl data. The architecture consists of a combination of the XLM [24] and RoBERTa [25] transformer models. The latter is an enhanced version trained with Masked Language Modelling of the original BERT architecture [26]. In general XLM-R has shown to outperform mBERT on several NLP benchmark tasks [6].

- **T5** is a monolingual (English) model pre-trained on a mixture of unsupervised and supervised tasks. For this challenge, the smaller version of T5 (T5-small [7]) is used. This model is pre-trained on the English C4 dataset [27] and the Wiki-DPR [28] dataset. For more information on these datasets and the model, refer to the T5-small model documentation on Huggingface [7]. A T5 peculiarity is that all downstream tasks are executed are performed in a sequence-to-sequence manner (using a string format).

3. Experimental Setup

Figure 1 provides a graphical overview of our experimental setup and the different training strategies. The main focus is on the two investigated approaches, i.e., unsupervised pre-training and supervised fine-tuning, and the datasets that are utilised.

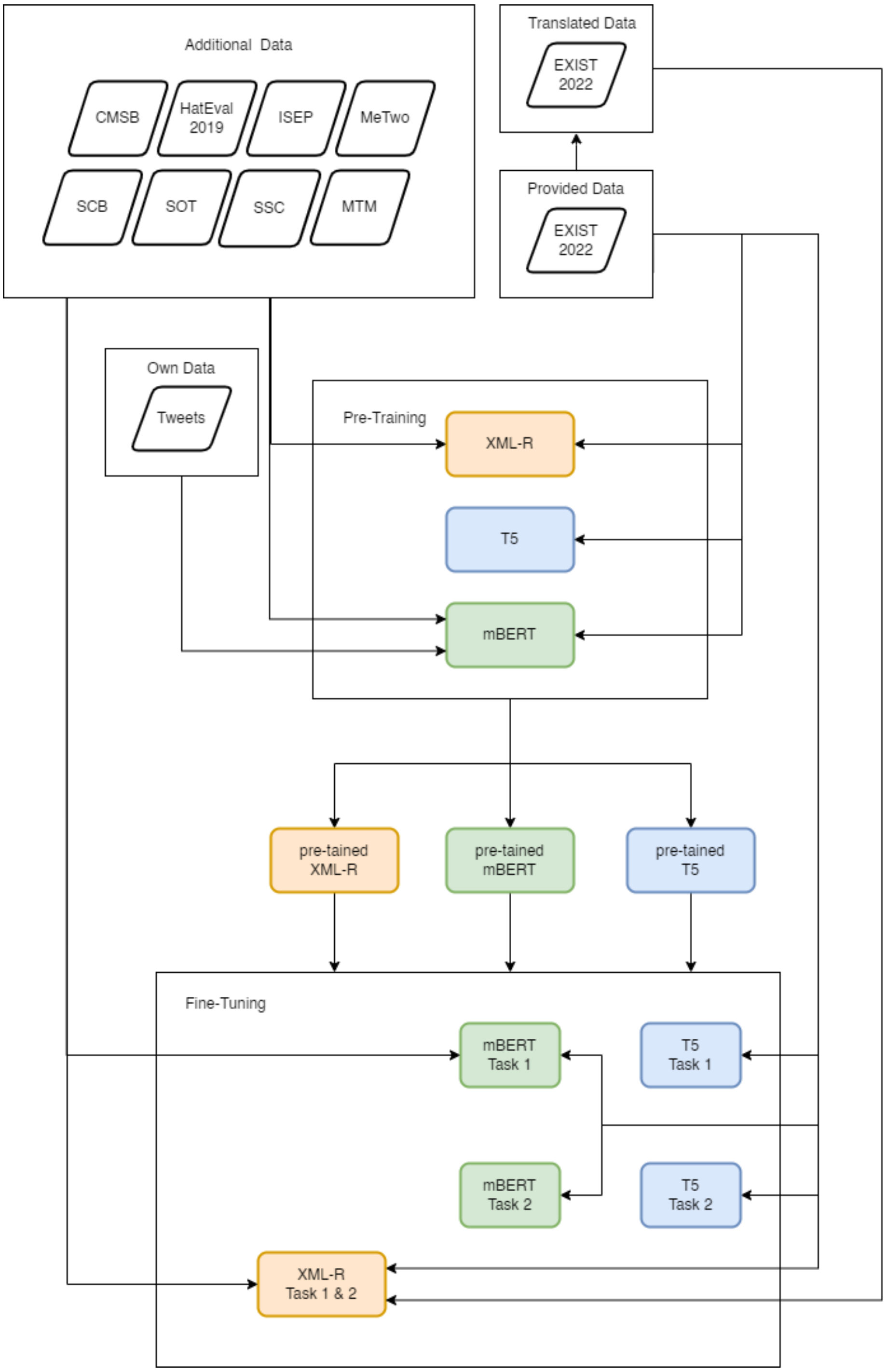

Figure 1: Overview of the experimental setup for training the three transformer models, including both training strategies, i.e., unsupervised pre-training and supervised fine-tuning, as well as the employed datasets.

To evaluate the different hyperparameters and dataset combinations, we first conducted several initial experiments on the EXIST data with multiple pre-trained transformer models provided by the HuggingFace [29] library. We started our experiments with pre-training as well as fine-tuning all three models (mBERT, XLM-R, and T5) on the EXIST2022 training data for both languages with the provided training set. Furthermore, we examine whether there was any benefit to aggregating the labels from task 2 to obtain labels for task 1, i.e., fine-tuning the models to task 2 and then aggregating the labels for the binary classification task (task 1). This aggregation leads to a better result only in the case of the XLM-R. Our best result with respect to the validation performance is obtained with the T5 model pre-trained and fine-tuned only on the EXIST2022 training set. Similarly to our last years approach we investigate three different pre-processing pipelines, since we gained better results last year:

- **Version 1:** removing hashtags, mentions, links and whitespace.
- **Version 2:** removing hashtags, links, whitespace characters and replacing mentions with *@USER*.
- **Version 3:** removing non-ASCII and whitespace characters.

We assume that cleaning the social media data would lead to less overfitting on the provided training set. However, this year the pre-processing steps worsened the classification performance, which lead to keeping the datasets in their original form. We conduct further experiments with pre-training the models with the additional datasets and the EXIST2022 training data. This results in better predictions than only fine-tuning the models on the EXIST2022 data. The combination of pre-training only on the EXIST2022 and fine-tuning with additional data leads to comparable results, however most predictions were better with pre-training it on more than just the EXIST2022 data. Overall, the results of the different trials were very similar, varying by only 1-2%. Finally, we also examine the impact of the translated EXIST2022 data for fine-tuning, which only slightly improved the results and even worsened them for mBERT. Unfortunately for the T5 model, during inference, in some cases random strings instead of the learned 0 and 1 strings are predicted by some of our models. One of the experiments included a pre-trained and fine-tuned model using the additional and translated datasets. This model performed slightly better on task 1 with an accuracy of 83,58% compared to our submitted model which reached an accuracy of 83,54% on our validation data. Since this model tends to fail in predicting the labels of some samples, it was not used for our final submission.

3.1. Unsupervised Pre-Training of XLM-R

For this system we use the already pre-trained XLM-R [6] and re-train the model with additional epochs using the RoBERTa Masked Language Modeling (MLM) on the following data: EXIST2022, CMSB [10], HatEval2019 [17], ISEP [15], MeTwo [30], MTM [16], SCB [11], SOT [13], and SSC [14]. We train the model on the MLM task with MLM probability of 0.15 for 25 epochs with a batch size of 16 and a learning rate of $5e^{-5}$. Sequence length is set to 128 tokens.

3.2. Unsupervised Pre-Training of mBERT

We trained the mBERT model on the MLM task with an MLM probability of 0.15. The model was trained for 10 epochs with a batch size of 16 and a learning rate of $2e^{-5}$. In contrast to the XLM-R model, we trained this model not only on the EXIST2022 data and the additional datasets (CMSB, HateEval, ISEP, MeTwo, MTM, SCB, SOT, and SSC), but also on a large dataset of tweets we sampled from the Covid-19 stream (see Section 2.2.3). To reduce the amount of training time, we only used 10 million samples from the complete total of gathered tweets for pre-training. This resulted in an overall training time of about seven days.

3.3. Unsupervised Pre-Training of T5

Our final T5-small model is re-trained on the original EXIST2022 dataset with an learning rate of $3e^{-4}$, a batch size of 8 and a maximum buffer size of 64.

3.4. Supervised Fine-Tuning of XLM-R

Since the aggregation strategy (i.e., train the transformer for task 2 first and then aggregate the labels for task 1) results in better performance for XLM-R, we use the same XLM-R model for both tasks. The model is fine-tuned on the original EXIST2022 training dataset and on the translated data. We train the model for 3 epochs with a learning rate of $2e^{-5}$ and then for 3 epochs with a learning rate of $1e^{-5}$. Batch size is set to 8 and sequence length to 256. This is the only model, where the predictions after fine-tuning for task 2 were used for the binary classifications in task 1.

3.5. Supervised Fine-Tuning of mBERT

We use an already pre-trained multilingual, cased BERT model, without lower-casing (model size: L=12, H=768, A=12; number of total parameters = 110M) [5], which we pre-train as stated above, and fine-tune for the binary classification task (task 1) on the EXIST2022 training data as well as the additional datasets (CMSB, SCB, MeTwo) for 7 epochs, with a batch size of 16, a maximum sequence length of 256 and a learning rate of $2e^{-5}$. However, the best results for task 2 with mBERT are achieved with fine-tuning it only on the EXIST2022 training data and slightly changed hyperparameters: 8 epochs, a batch size of 16, a maximum sequence length of 256 and a learning rate of $2e^{-5}$.

3.6. Supervised Fine-Tuning of T5

The T5 transformer model is able to learn any downstream task by providing the model the specific input sequences (text) and the desired targets (classes). Also, a specific prefix is added which enables the model to remember which downstream task should be done in inference. Since the T5-Model is only able to accomplish sequence-to-sequence tasks, the labels are first mapped to *0* and *1* and then transformed into strings. Our submitted T5-small models for task 1 and task 2 are pre-trained and fine-tuned only on the original EXIST2022 dataset. For both tasks, the models are trained with a maximum sequence length of 512 and a learning rate of $1e^{-4}$. The model for task 1 is trained for 9 epochs and the model for task 2 for 10 epochs.

Table 2
Macro-averaged F1-scores (F1) and classification accuracies (CA) split by language. Abbreviation "val" stands for results on the validation set and "test" for official test data. The performance measures are expressed in percent (%).

Task	Run	Approach	CA (val)	F1 (val)	CA (test)	F1 (test)	Ranking
1	1	mBERT	85.29	84.37	74.20	74.10	29^{th}
1	2	T5	83.54	83.52	71.83	71.81	36^{th}
1	3	XLM-R	85.90	86.48	75.05	74.96	26^{th}
2	1	mBERT	77.06	64.74	64.18	43.66	19^{th}
2	2	T5	83.54	83.82	52.55	35.71	25^{th}
2	3	XLM-R	77.63	72.46	65.22	46.75	13^{th}

4. Results

The overall best results are obtained using the EXIST2022 training data as well as the additional datasets for pre-training in all experiments, regardless of the employed model. However, the final data augmentation strategies and hyperparameters for fine-tuning slightly differ for each model. In the following, we present the setup of the final approaches submitted to the benchmark for evaluation. For calculating the evaluation metrics in the development phase we randomly split the provided EXIST2022 training set into a 90% training and 10% validation set. The validation and test results for both tasks are presented in Table 2. The last column in Table 2 lists the ranking of our submissions in the EXIST2022 benchmark. The top ranked submission in the overall benchmark achieved an accuracy of 79.96% and a macro-averaged F1-score of 79.78% for task 1 (team: avacaondata) and an accuracy of 70.13% and a macro-averaged F1-score of 51.06% for task 2 (team: avacaondata). Our strategies are briefly summarized in the following, before we discuss the models in detail comparing the performance metrics for both languages.

Pre-training strategy: The XLM-R and mBERT are both trained on the original EXIST2022 training set and additional data (i.e., CMSB, HatEval, ISEP, MeTwo, MTM, SCB_ES, SOT_ES, SSC_EN). Additionally, the mBERT is trained on 10 million tweets. The T5 is only trained on the original EXIST2022 training set. The pre-training strategy does not differ for both given tasks.

Fine-tuning strategy: Based on the experiments, we applied different fine-tuning strategies for each pre-trained model. mBERT was fine-tuned for both tasks on the EXIST2022 data, as well as the additional datasets for task 1. In task 2 the model performed better with only fine-tuning it on the EXIST2022 data. The T5 was only fine-tuned on the EXIST2022 training set in both tasks. The XLM-R on the other hand was fine-tuned with the EXIST2022 data and its translations. The model was only trained on task 2, which means that the predictions for task 1 were aggregated across the different sexism sublasses of task 2.

In our approach in task 1, all of our models seem to overfit on the training data, since the validation accuracy and F1-score are significantly higher than on the final testdata. However, the results for the binary classification all of our models have a very similar performance and only slight differences in the overall metrics, when compared directly. This is not the case for our results for the fine-granular 6-way classification in task 2. mBERT and XLM-R have similar

Table 3
Macro-averaged F1-scores (F1) and classification accuracies (CA) split by language. Abbreviation “EN” stands for the English test data and “ES” for the Spanish test data. The performance measures are expressed in percent (%).

Task	Run	Approach	CA (EN)	F1 (EN)	CA (ES)	F1 (ES)
1	1	mBERT	76.43	76.24	71.99	71.96
1	2	T5	74.90	74.87	68.80	68.52
1	3	XLM-R	76.05	75.61	74.06	74.05
2	1	mBERT	67.30	44.41	61.09	42.35
2	2	T5	53.23	35.41	51.88	35.87
2	3	XLM-R	66.35	45.45	64.10	47.75

results, while XLM-R performed slightly better. However, the T5 model fails at the multiclass task with a difference between the validation F1-score of 83.82% and test F1-score of only 35.71%, even though this model had the best overall performance of all of our approaches during the validation experiments. In general, it seems that pre-training the models on a large set of data enhanced the performance and make the models more robust in a test setting. Since XLM-R outperformed mBERT in their original papers - as we stated in the model descriptions - and also in our last years approach, we assumed to have similar results this year. We believed that using 10 million additional tweets - based on the last years keywords - for pre-training the mBERT would outperform the XLM-R by chance. Even though they have comparable results, the XLM-R was still our best model in that case. This could be due to hyperparameter settings during the fine-tuning strategy.

4.1. Language-Specific Results

In this section the detailed results for English and Spanish are shown in Table 3. In general, all models performed better for English data. Since the training data is split evenly across languages, we believe this is due to the pre-trained multilingual models that we used in our experiments. Usually large models are mostly trained only on English data and for additional languages there is often not enough training data available. For task 1 the results show that pre-training mBERT and XLM-R on both English and Spanish additional data slightly led to outperforming the T5, even though it had a higher performance on the validation set. This is confirming the observation we had based on the overall results. For task 2, mBERT and XLM-R have again a very similar output as shown in the general results - with less precision on the Spanish data. However, the T5 performed much worse for Spanish than for English data, which explains the overall results.

5. Discussion & Conclusion

In this paper, we provided the details on our submission to the EXIST2022 benchmark, which consists of two tasks on the classification of sexist content. In our experiments we found that the unsupervised pre-training strategy of the XLM-R model [6] with additional external data is the most promising strategy. This year the XLM-R achieved an F1-score of 74.96% in task 1 and

46.75% in task 2. The T5 model only being pre-trained on the EXIST2022 data is outperformed by the former strategy and shows great signs of overfitting. As we already concluded in our first approach last year, the use of additional data (either external datasets or translations) resulted in improvement for all strategies. Similarly, our experiments again reveal that the pre-training of the whole model on domain-specific data was more effective compared to only using the EXIST2022 data. As a final remark, depending on the model, the 10 million additional tweets for pre-training mBERT did not outperform the XLM-R. In future work, we want to compare the pre-training strategies of XLM-R and mBERT in detail.

Acknowledgments

This contribution has been funded by the FFG Project "Defalsif-AI" (Austrian security research programme KIRAS of the Federal Ministry of Agriculture, Regions and Tourism(BMLRT), grant no. 879670) and Project "Young People Against Online Hate: Computer-assisted Strategies for Facilitating Citizen-generated Counter Speech", WWTF Austria, grant no. ICT-20-016.

A. Online Resources

Our implementation and datasets used are available at: https://github.com/fhstp/EXIST2022

References

[1] C. Demus, J. Pitz, M. Schütz, N. Probol, M. Siegel, D. Labudde, Detox: A comprehensive dataset for german offensive language and conversation analysis, in: Proceedings of the 6th Workshop on Online Abuse and Harms (WOAH 2022), Association for Computational Linguistics, Online, 2022, pp. 54–61.

[2] M. Schütz, A. Schindler, M. Siegel, K. Nazemi, Automatic fake news detection with pre-trained transformer models, in: A. Del Bimbo, R. Cucchiara, S. Sclaroff, G. M. Farinella, T. Mei, M. Bertini, H. J. Escalante, R. Vezzani (Eds.), Pattern Recognition. ICPR International Workshops and Challenges, Springer International Publishing, Cham, 2021, pp. 627–641.

[3] 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 69 (2022).

[4] M. Schütz, J. Boeck, D. Liakhovets, D. Slijepcevic, A. Kirchknopf, M. Hecht, J. Bogensperger, S. Schlarb, A. Schindler, M. Zeppelzauer, Automatic sexism detection with multilingual transformer models, CoRR abs/2106.04908 (2021). URL: https://arxiv.org/abs/2106.04908. arXiv:2106.04908.

[5] I. Turc, M.-W. Chang, K. Lee, K. Toutanova, Well-read students learn better: On the importance of pre-training compact models, arXiv preprint arXiv:1908.08962 (2019).

[6] 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.

[7] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, 2020. URL: https://huggingface.co/t5-small, accessed on 18.05.2022.

[8] A. Schindler, T. Lidy, A. Rauber, Comparing shallow versus deep neural network architectures for automatic music genre classification, in: In Proceedings of 9th Forum Media Technology (FMT2016), St. Pölten University of Applied Sciences, Institute of Creative Media Technologies, 2016, pp. 17–21.

[9] F. Rodríguez-Sánchez, J. C. 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).

[10] M. Samory, I. Sen, J. Kohne, F. Floeck, C. Wagner, The 'call me sexist but' dataset (cmsb), 2021. URL: https://doi.org/10.7802/2251, accessed: 2022-03-09.

[11] R. I. Medina, Sexismo en código binario: Violencia digital y política contra las mujeres en méxico, 2021. URL: https://github.com/RMedina19/sexismo_codigo_binario, accessed: 2022-03-09.

[12] 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, accessed: 2022-03-09.

[13] D. G. Ibáñez, R. V. Puig, Sexism bot classifier on twitch: Moderating sexist comments on twitch streamings., 2020. URL: https://github.com/VPRamon/SexismOnTwitch, accessed: 2021-05-04.

[14] A. Debnath, S. S, N. Bhakt, K. Garg, P. Parikh, Sexist stereotype classification on instagram data, 2020. URL: https://github.com/djinn-anthrope/Sexist_Stereotype_Classification, accessed: 2021-05-04.

[15] D. Grosz, P. Conde-Cespedes, sexist-workplace-statements, 2020. URL: https://www.kaggle.com/datasets/dgrosz/sexist-workplace-statements, accessed: 2022-03-09.

[16] T. Lynn, P. T. Endo, P. Rosati, I. Silva, G. L. Santos, D. . Ging, Urban dictionary definitions dataset for misogyny speech detection, 2019. URL: https://data.mendeley.com/datasets/3jfwsdkryy/3. doi:10.17632/3jfwsdkryy.3, accessed: 2021-05-04.

[17] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, M. Sanguinetti, 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, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 54–63.

[18] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on Twitter, in: Proceedings of the NAACL Student Research Workshop, Association for Computational Linguistics, San Diego, California, 2016, pp. 88–93. URL: https://aclanthology.org/N16-2013. doi:10.18653/v1/N16-2013.

[19] 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, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 7–16. URL: https://aclanthology.org/W17-2902. doi:10.18653/v1/W17-2902.

[20] M. Samory, I. Sen, J. Kohne, F. Flöck, C. Wagner, "unsex me here": Revisiting sexism detection using psychological scales and adversarial samples, CoRR abs/2004.12764 (2020).

URL: https://arxiv.org/abs/2004.12764. arXiv:2004.12764.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. ukasz 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 30, Curran Associates, Inc., 2017, pp. 5998–6008. URL: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

[22] 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, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://www.aclweb.org/anthology/N19-1423. doi:10.18653/v1/N19-1423.

[23] J. Devlin, S. Petrov, Bert multilingual models, 2019. URL: https://github.com/google-research/bert/blob/master/multilingual.md, accessed: 2010-06-02.

[24] A. Conneau, G. Lample, Cross-lingual language model pretraining, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 32, Curran Associates, Inc., 2019, p. 11. URL: https://proceedings.neurips.cc/paper/2019/file/c04c19c2c2474dbf5f7ac4372c5b9af1-Paper.pdf.

[25] 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.

[26] 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.

[27] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, arXiv e-prints (2019). arXiv:1910.10683.

[28] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W. tau Yih, Dense passage retrieval for open-domain question answering, 2020. arXiv:2004.04906.

[29] 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.

[30] 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.