When Multiple Perspectives and an Optimization Process Lead to Better Performance, an Automatic Sexism Identification on Social Media With Pretrained Transformers in a Soft Label Context

Johan Erbani, Előd Egyed-Zsigmond, Diana Nurbakova and Pierre-Edouard Portier

Univ Lyon, INSA Lyon, CNRS, UCBL, LIRIS, UMR5205, 20 Avenue Einstein, 69621 Villeurbanne, France

Abstract

Even if today, the sexism is socially widely disapproving, it remains an omnipresent phenomenon in our society. But faced with huge quantities of data, social platforms are struggling to identify it. This highlights the need to develop automatic detection tools that can subtly assess the sexistness of user-generated content. That's what sEXism Identification in Social neTworks (EXIST) is all about. The EXIST 2023 contest consists of three classification tasks : 1. detect sexism, 2. clarify the author's intention and 3. explicit the sexism type. Thanks to these three tasks, each data could be seen from three different points of view. This idea, combined with fine-tuned BERTs, model stacking and an optimization process, enabled us to rank 1st in the task 2 and 4th in the task 3 in a soft label context. This paper describes our approach, our negative results and some possible perspectives.

Keywords

Sexism Identification, Natural Language Processing, Transformer Models, BERT, Ensemble modeling, Sentiment Analysis, Sexism Detection, Twitter

1. Introduction

Identifying sexism automatically remains an open problem in Natural Language Processing (NLP). To address this issue, a series of scientific events called EXIST has been established with the objective of comprehending sexism in its widest scope. This includes not only explicit sexism but also more subtle forms of implicit sexist behavior. These scientific initiatives have the potential to raise awareness about women's rights issues and promote social cohesion. This paper describes the DRIM team's contribution to the three EXIST 2023 shared tasks.

This work is structured as follows: Section 2 briefly provides a description of several earlier studies. Section 3 will then present an explanation of tasks 1, 2 and 3, along with the corpus provided by the organizers. Following that, Section 4 and 5 will outline the experimental methodology and evaluation results respectively. Finally, in Section 6, we will present the key findings and conclusions of our studies, as well as some potential directions for future research.

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18-21, 2023, Thessaloniki, Greece

2. Related Works and Contributions

According to previous EXIST rapports, transformer-based models performed better than the other technique [1, 2]. Given this and the limited availability of labeled data, we employed a fine-tuning approach using models that were initially pre-trained in a self-supervised manner using extensive amounts of unlabeled data. More specifically, we have chosen to work with BERT which is commonly used in state-of-the-art approaches across various NLP problems.

As explain in [3], previous studies [4, 5, 3] have highlighted the issue of high performance dependency on the seed value when fine-tuning BERT for downstream tasks, particularly when the training data is limited. One way of reducing this undesirable effect is to use ensembles of models as in [3, 6]. Ensemble Modeling is an approach that encompasses the combination of multiple models. It is based on the premise that individual models may possess distinct strengths and limitations, and by merging them, an improved overall performance can be achieved. The key benefit of employing model ensembles lies in their capacity to diminish variance and enhance predictive accuracy.

The paper [7] used a Multi-Task Learning approach to solve a previous EXIST challenge edition. Multi-Task Learning is a machine learning method where a model is trained to perform several different tasks simultaneously. Instead of training separate models for each task, multi-task learning aims to share the knowledge and representations learned between the different tasks, which can potentially improve the overall performance of the model.

Our contributions relate to the development of a strategy that combines elements from both Ensemble Modeling and Multi-Task Learning methods. Our model is built upon three stacked BERT. Unlike previous approaches, our proposal focuses on observing the same object from multiple perspectives. The underlying idea was to provide the model with input various aspects to enhance its comprehension and interpretation abilities. Experimental results show that our strategy could outperform single-view models. Furthermore, we propose incorporating a prediction refinement mechanism on top of our models through an optimization process. This refinement process does not alter the model's weights but it has enabled us to outperform other models.

3. Datasets and Tasks

In 2023, the EXIST event involved the categorization of several thousands tweets written in English and Spanish. This categorization process encompassed three distinct tasks: detect sexism, clarify the author's intention and explicit the sexism type. Approximately, whatever the tasks, the Non-sexist class accounting for half the total annotator votes. The other classes are evenly distributed among the remaining votes. The data, the tasks and the classes are summarized in the tables 1, 2 and 3, respectively (see [8, 9] for more details). Each tweet is annotated by 6 annotators of different ages and genders, with the aim of 1. obtaining less biased labels and 2. learn more subtly to recognize the different categories. Indeed, the assumption that natural language expressions have a single and clearly identifiable interpretation in a given context is a convenient idealization, but it's far from reality, as explained in [10]. To cope with this, EXIST 2023 proposes to learn directly from different annotators' votes.

Table 1
An overview of EXIST 2023 data for both languages and both sets.

Language	Data set	Number of tweets
English	Train	3260
	Development	489
Spanish	Train	3660
	Development	549

Table 2
An overview of EXIST 2023 tasks.

Task n°	Question	Number of classes	Type
1	Sexist or not	2	Mono label
2	Author's intention	4	Mono label
3	Sexism type	6	Multi-label

The organisers leave participants free to choose the type of labels predicted:

- Soft-Soft. Provide the vote distribution;
- Hard-Soft. Provide the vote distribution and a hard-label;
- Hard-Hard. Provide only the hard-label.

We participated in all tasks in the Soft-Soft context.

For all tasks, whatever the context, the evaluation metric was the Information Contrast Measure (IMC) [11]. It is a similarity function that generalizes Pointwise Mutual Information (PMI).

4. Methodology

In this section, we describe our approach and the experimental framework employed. We present here only the attempts that we consider scientifically interesting, the successful ones as well as the unsuccessful ones, in order to maximize the usefulness of the paper for the reader.

We began by separating the dataset according to the language of the tweets. All the exploratory work was carried out on the English data. Once the best method had been determined, we replicated it on the Spanish data with BERT multilingual. In the following, as the procedures are identical, we will only describe the process of the English data.

4.1. Baseline

In order to compare our different attempts, we began by implementing a BERT base uncased baseline using a traditional approach. The architecture used is illustrated in Figure 1. The process involves taking the BERT classifier token [CLS], applying dropout, passing it through a dense layer, and finally using a softmax activation function. While using softmax for tasks 1 and

Table 3
An overview of the class statistics for both the training and development datasets at EXIST 2023.

Task n°	Class	Definition	Train Set	Dev. Set
1	Non-sexist	Not sexist	55%	52%
	Sexist	Sexist or about sexism (e.g. if the author denounces a sexist act or fact)	45%	48%
2	Non-sexist	Not sexist	55%	52%
	Direct	Sexist by itself	22%	22%
	Reported	Report a sexist situation	11%	12%
	Judgemental	Decrying a social injustice against women	12%	14%
3	Non-sexist	Not sexist	47%	43%
	Ideological Inequality	Discredits the feminist movement, rejects inequality between men and women	12%	13%
	Stereotyping Dominance	Promotes gender stereotypes, superiority of men, and limits women's abilities	14%	15%
	Objectification	Women are presented as objects apart from their dignity and personal aspects	11%	11%
	Sexual Violence	Sexual suggestions, requests for sexual favors or harassment of a sexual nature	07%	8%
	Misogyny Non-Sexual Violence	Expresses hatred and violence towards women	09%	10%

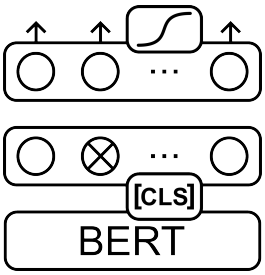

Figure 1: Architecture $\mathcal{B}i$ for task i where i could be $1, 2$ or 3. At the bottom, BERT base uncased from which we get the classifier, followed by dropout and linear layer. The output size matching with the number of categories of the task i. The top activation function is a softmax.

Figure 2: Illustration of the idea *When multiple perspectives lead to a better understanding.*

2 is appropriate because they are single-label tasks, it may not be suitable for task 3, which is a multi-label task. However, in the specific context of the competition, as explained in section 4.6, this issue is not problematic. Although the influence of the hyperparameter dropout is minor, we empirically determined that among the $\{0.1, 0.3, 0.5\}$ values the optimal rates were 0.5, 0.1 and 0.5 for task 1, 2 and 3 respectively. In accordance with the original BERT paper [12] and

with our experiments, we have chosen a batch size of 32. In order to increase the generalization capacity of the model, we preferred to optimize with AdamW rather than Adam as proposed in [13] with learning rate $lr = 1e - 5$. We use the cross-entropy loss as cost function. Training of the model was stopped at the peak of the ICM on the development set, which turned out to be 4, 7 and 9 epochs for tasks 1, 2 and 3 respectively.

For more stability in the initial phase of training, we applied linear learning rate warm-up during the first steps of the updates followed by a linear decay. However, we did not observe any positive effect. We also tried to apply the layer-wise learning rate decay technique. As explained in [14, 15] lower layers encode more general information and top layers are more specific to the training task. Consequently, the higher a layer is, the larger its learning rate should be and conversely, the lower it is, the smaller it should be. But again, even if the learning was faster during the first epochs at the end, we have not seen any improvement. Consequently, we did not retain these two strategies. However, a more detailed studies of hyperparameters might have led to further gains.

In the following, we will refer to $\mathcal{B}1$, $\mathcal{B}2$ and $\mathcal{B}3$ as the three BERTs fine-tuned according to the protocol described above on tasks 1, 2 and 3 respectively.

4.2. Preprocessing

We tried different preprocessing listed in the Table 4. When process involves new tokens, we update the BERT tokenizer to include new tokens, enabling the model to learn from them. In our study, we found that pre-processing had a minor influence on model performance. The best results were obtained using the first two approaches. Note that giving the meaning of hashtags and emojis as we did not seem to provide more information on average.

Table 4
Exploration of the various types of preprocessing techniques.

Idx	Description	Original tweet	Pre-processed tweet
1	Tokens hashtags, mentions, urls	@Bob43 #Metoo see u http...	<mention> <hashtag> see u <url>
2	Standardize text	I hate U & f∗ck you !!	I hate you and fuck you !!
3	Explicit hashtag	#Metoo	<hashtag> me too
4	Explicit emoji	I love U :)	I love U <emoji> happy

4.3. Different data perspectives

An intuitive idea illustrated in the Figure 2 is that multiple perspectives of an object can lead to a better understanding of it. We combined the models introduced in section 4.1 $\mathcal{B}1$, $\mathcal{B}2$, and $\mathcal{B}3$, to create a meta-model. $\mathcal{B}1$, $\mathcal{B}2$, and $\mathcal{B}3$ were trained on the same data but on different tasks 1, 2 and 3 respectively. After training the baseline models, we froze them to prevent further training and then stacked them together. On top of the stacked models, we added some additional layers. The resulting architecture is illustrated in Figure 3. This strategy worked well for tasks 1 and

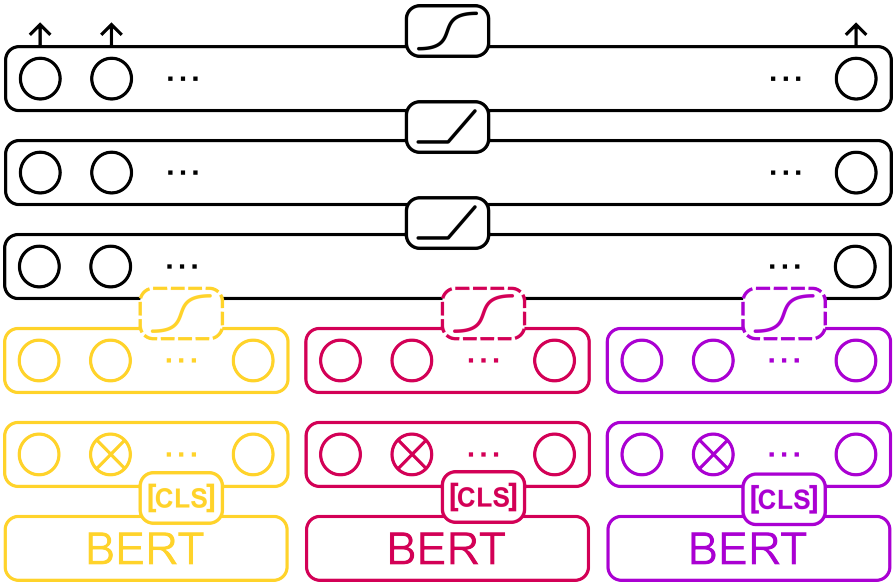

Figure 3: Architecture meta-model $\mathcal{M}i$ for task i where i could be $1, 2$ or 3. In yellow, $\mathcal{B}1$, in red $\mathcal{B}2$ and in purple $\mathcal{B}3$. On top, three linear layers with respectively a ReLU, a ReLU and a softmax as activation function. The optimal hidden size dimension was empirically determined as 12 i.e. the total number of categories all tasks combined. The output size is the number categories of the task i.

2, but it did not work for task 3. Additionally, the decision of whether to keep softmax on the top of the frozen networks depended on the task. It was better to keep softmax for task 1 and better to exclude it for task 2. In the following, we will refer to $\mathcal{M}1$ and $\mathcal{M}2$ to designate the two meta-models described above.

Based on our submission, it's difficult to pinpoint the exact reasons why the strategy worked for some tasks but not for others. However we assume that task 3 is the most difficult one and requires subtleties that are not adequately captured by the baseline models $\mathcal{B}1$ and $\mathcal{B}2$. A supporting evidence is that this strategy exhibited the highest effectiveness on task 1, which is considered the least complex among the tasks. Consequently, $\mathcal{B}2$ and $\mathcal{B}3$ could presumably transfer their captured nuances to $\mathcal{B}1$.

4.4. Manual features

In our effort to enhance the performance of our models, we sought out manual features that could uncover valuable information potentially overlooked by our existing models. These manual features were subjected to normalization before being stacked with the frozen networks as shown in Figure 4. Here is a list of the specific additional features we incorporated:

- **Tokens.** Give the number of <hashtag>, <mention> and <url>;
- **Text statistics.** Give the number of characters, upper characters, words, sentences, digits, citations, question marks and exclamation marks;
- **Sentiment analysis.** With the python library TextBlob, we provide the subjectivity and

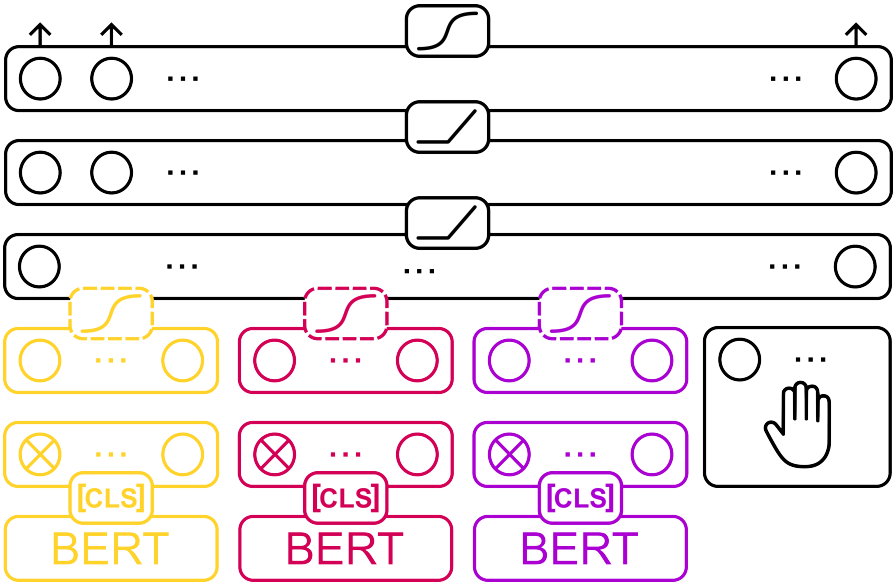

Figure 4: Modified architecture of $\mathcal{M}i$ for task i where i could be $1, 2$ or 3. Except for the inclusion of manual features, the only difference with the Mi architecture is the hidden size, which here is equal to $12+$*the number of manual features*.

the polarity;

- **Part-of-speech tagging.** Give the number of adjectives, noun, verb, adverbs...
- **Tenses and modal.** Give the number of verbs conjugated in the future, present or past tense and the number of modal verbs;
- **Complexity.** Provide the average word size, the average sentence size and a readability index (Flesch Reading Ease);
- **Latent Dirichlet Allocation (LDA).** We group test and train sets in a big corpus to identify main topics by the LDA process.

Regrettably, utilizing our current methodology, none of the manually engineered features demonstrated a discernible enhancement in the model's performance across various tasks. It is postulated that these features encompass information that has already been assimilated by our architecture.

4.5. Data augmentation

Three distinct approaches were implemented and evaluated. Unfortunately, none of them yielded desirable outcomes.

The first approach involved translating the Spanish data and augmenting the model's training set. Against intuition, this augmentation did not lead to performance improvements; instead, it resulted in decreased performance.

The second attempt aimed to test the assumption that the translated data was noisy. To counterbalance this noise, we have tried to select only tweets with a significant informational content. The methodology entailed training the model on the English data, subsequently testing it on the translated data, and identifying the instances where the model exhibited significant errors. These error-prone tweets were presumed to possess heavy information for the model's learning. Subsequently, the model's weights were reset, and the training was repeated on the base data alongside the translated data with the "most significant" information content. Unfortunately, this strategy failed to improve the model's performance.

Lastly, a third approach was undertaken by applying a similar strategy as described in section 4.3, but with other datasets. Two distinct datasets were utilized: the Spanish data translated from EXIST 2023 and the EXIST 2022 dataset. The methodology involved stacking $\mathcal{B}1$, $\mathcal{B}2$, and $\mathcal{B}3$ with additional frozen baseline models trained on specific tasks, such as task 1 of 2022 or the translated Spanish data of task 3. The underlying principle was to harness the collective insights of diverse people, represented by different models trained on distinct data types, in order to synthesize their outputs and generate an improved solution. Regrettably, this method has not produced convincing results.

These unsuccessful results led us to make two hypotheses. Firstly, we posited that tweets exhibit significant dissimilarities based on their cultural or temporal origins. Secondly, we hypothesized that the cultural disparity between Spanish and English annotators could lead to different evaluations.

4.6. Best possible distribution

The dataset construction results in non-continuous labels, despite their apparent continuity. This is due to the presence of six annotators for each task, which limits the set of possible label distributions to a finite number. Specifically, all label values are multiples of $1/6$. For instance, in task 1, there exist only seven feasible outputs denoted as $O = (o_1, o_2)$, where $o_1 = i/6$ for $i = 0, 1, \ \dots, 6$, and $o_2 = 1 - o_1$. Similarly, task 2 encompasses 84 possibilities, while task 3 encompasses a substantial number of $28\ 546$ possibilities due to its multi-label nature.

Exploiting this knowledge, after training the model, it becomes feasible to select, from the set of possible distributions, the closest one for each prediction. In particular, the multi-label task 3 could be solved with a softmax on the top of our model. The advantage is that the model benefits from the gradient backpropagation offered by the softmax function during its learning and has a good approximation to a label that does not sum to 1 thanks to the optimization trick. It is important to note that the process described is an optimization problem and not a deep-learning problem. Consequently, this optimization procedure does not impact the model during the training. This sneaky idea has considerably increased the model's performance. This optimization procedure will be noted $\mathcal{O}$ in the following.

5. Results

In this section, we provide a concise overview of the performance attained by our models configurations in the EXIST 2023 evaluation. Our rankings are indicated in the Table 5. It is noteworthy to point out that our current performance has improved slightly on that obtained

during the competition. This is due to the extra time we had to refine the hyperparameters of our models.

We can see in the Table 6 that the optimization process significantly improves model performance. Our meta-model strategy was also effective, but to a lesser extent.

Table 5
An overview of the ranking achieved at EXIST 2023.

Task n°	Language	Ranking	ICM	Cross Entropy
1	Both	22/56	0.54	0.89
	English	5/56	0.73	0.95
	Spanish	25/56	0.34	0.84
2	Both	1/27	-1.34	1.78
	English	1/27	-1.15	1.80
	Spanish	4/27	-1.54	1.77
3	Both	4/25	-3.68	-
	English	4/25	-3.08	-
	Spanish	4/25	-4.16	-

Table 6
An analysis of the performance exhibited by diverse models employed during EXIST 2023 on English data.

Task n°	Model	ICM	ICM Gain Over $\mathcal{B}i$	Cross Entropy
1	$\mathcal{B}1$	0.84	-	0.51
	$\mathcal{M}1$	1.01	+20%	0.52
	$\mathcal{M}1 + \mathcal{O}$	1.15	+36%	0.52
2	$\mathcal{B}2$	-0.93	-	0.97
	$\mathcal{M}2$	-0.88	+5%	0.98
	$\mathcal{M}2 + \mathcal{O}$	-0.52	+44%	0.98
3	$\mathcal{B}3$	-3.54	-	1.67
	$\mathcal{B}3 + \mathcal{O}$	-2.84	+20%	1.67

6. Conclusion

The submission made to the EXIST 2023 evaluation has yielded a considerable number of negative or neutral results. These outcomes are valuable as they indicate that specific approaches or hypotheses within our particular context did not produce favorable results. However, we have put forth two effective strategies.

Firstly, we proposed an innovative approach that combines ensemble modeling and the multi-learning model. This approach entails training the same architecture on multiple tasks using the same data and subsequently stacking the frozen sub-models in a meta-model. Our findings demonstrate that this architecture has the ability to surpass the limitations of single view models, leading to improved performance.

Furthermore, we suggested incorporating an optimization process into our model. Our ranking highlights that in a competition it is more important to be close to labels rather the ground truth. This optimization process plays a key role in refining the model's predictions and enhancing its overall performance.

Acknowledgments

Johan Erbani would like to sincerely thank his superiors and the entire DRIM team for the trust they have placed in him. Special thanks are due to Pierre-Yves Genest and Ousmane Touat for their precious advice and expertise, and to interns Maud Andruszak and Thanh Lam for their diligent research efforts, who made some contributions through their work on manual features.

References

[1] 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.

[2] 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) 229–240.

[3] E. Villa-Cueva, F. Sanchez-Vega, A. P. López-Monroy, Bi-ensembles of transformer for online bilingual sexism detection (2022).

[4] T. Zhang, F. Wu, A. Katiyar, K. Q. Weinberger, Y. Artzi, Revisiting few-sample bert fine-tuning, arXiv preprint arXiv:2006.05987 (2020).

[5] J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, N. Smith, Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping, arXiv preprint arXiv:2002.06305 (2020).

[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] F. M. Plaza-del Arco, M. D. Molina-González, L. López, M. Martín-Valdivia, Sexism identification in social networks using a multi-task learning system, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain, volume 2943, 2021, pp. 491–499.

[8] Laura Plaza, Jorge Carrillo-de-Albornoz, Roser Morante, Enrique Amigó, Julio Gonzalo, Damiano Spina, Paolo Rosso. Overview of EXIST 2023 – Learning with Disagreement for Sexism Identification and Characterization. Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023). Avi Arampatzis, Evangelos Kanoulas, Theodora Tsikrika,

Stefanos Vrochidis, Anastasia Giachanou, Dan Li, Mohammad Aliannejadi, Michalis Vlachos, Guglielmo Faggioli, and Nicola Ferro, Eds. September 2023, Thessaloniki, Greece.

[9] Laura Plaza, Jorge Carrillo-de-Albornoz, Roser Morante, Enrique Amigó, Julio Gonzalo, Damiano Spina, Paolo Rosso. Overview of EXIST 2023 – Learning with Disagreement for Sexism Identification and Characterization(Extended Overview). Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum. Mohammad Aliannejadi, Guglielmo Faggioli, Nicola Ferro and Michalis Vlachos, Eds.

[10] A. Uma, T. Fornaciari, A. Dumitrache, T. Miller, J. Chamberlain, B. Plank, E. Simpson, M. Poesio, Semeval-2021 task 12: Learning with disagreements, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), 2021, pp. 338–347.

[11] E. Amigo, A. Delgado, Evaluating extreme hierarchical multi-label classification, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 5809–5819. URL: https://aclanthology.org/2022.acl-long.399. doi:10.18653/v1/2022.acl-long.399.

[12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[13] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 (2017).

[14] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, arXiv preprint arXiv:1801.06146 (2018).

[15] C. Lee, K. Cho, W. Kang, Mixout: Effective regularization to finetune large-scale pretrained language models, arXiv preprint arXiv:1909.11299 (2019).