

# Leveraging MiniLMv2 Pipelines for EXIST2023

Notebook for the EXIST Lab at CLEF 2023

Alexandru Petrescu<sup>1,\*</sup>

<sup>1</sup>*Politehnica University of Bucharest, Splaiul Independenței 313, București 060042, Romania*

## Abstract

This paper introduces our methodology for addressing the three EXIST2023 tasks focused on Sexism Identification in Social Networks. Our proposed solution leverages advanced multi-lingual transformers, specifically XLMR and MiniLMV2, which serve as state-of-the-art frameworks for handling the provided data. Additionally, we employ a data-processing pipeline and task-specific metrics to fine-tune the pre-trained model. The evaluation of our solution demonstrates promising outcomes on the testing set and attains a commendable overall ranking within the competition, particularly excelling in the Soft-Soft evaluation. Our proposed architecture successfully tackles all tasks with favorable outcomes, while also offering opportunities for further enhancements. While the Hard evaluation could benefit from improvements, it is worth noting that our models exhibited signs of overfitting when trained on the available data. In light of this, we made the decision to provide them with more flexibility in the classification process. This approach allowed for increased adaptability and potentially better generalization when handling unseen data instances.

## Keywords

NLP, MiniLMv2, Transformers, Text Classification, Learning with disagreements, Sexism detection

## 1. Introduction

The following document serves as the working notes for our submission to EXIST 2023, described in [1] and, representing the efforts of the AlexPUPB team. EXIST is a renowned series of scientific events and shared tasks focused on the identification of sexism in social networks. The objective of EXIST is to encompass sexism in its entirety, ranging from overt misogyny to more subtle manifestations involving implicit sexist behaviors.

For this particular event, we tackled three interrelated tasks:

- TASK 1: Sexism Identification - This task involved a binary classification approach.
- TASK 2: Source Intention - We employed a multi-class (4) classification technique to discern the intentions behind the identified sexism.
- TASK 3: Sexism Categorization - To achieve a comprehensive understanding, we employed a multi-label classification strategy for categorizing sexism.

---

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

\*Corresponding author.


✉ [alex.petrescu@upb.ro](mailto:alex.petrescu@upb.ro) (A. Petrescu)

🌐 <https://alexpetrescu.net/> (A. Petrescu)

🆔 0000-0002-7731-2403 (A. Petrescu)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

While addressing each task, we leveraged a state-of-the-art pipeline that utilizes transformers. This pipeline demonstrates adaptability, allowing us to tailor our approach to the specific requirements of each task, including the generation of labels and the choice of evaluation metrics.

Transformers have emerged as the leading methodology for text-related operations, particularly in the realm of classification. We take advantage of the remarkable capabilities of transformers, making use of industry-trained models facilitated by the Hugging Face platform. Furthermore, we fine-tune these models to optimize their performance for our particular task.

## 2. Related Work

In the current literature, many solutions have been proposed for detecting harmful content. Some solutions focus on how embeddings affect the classification results [2, 3, 4], others propose new stacked deep neural networks [5] or transformer-based ensemble models [6]. Furthermore, there are methods that also methods that consider network-dependent information [7]. Another research direction deals with network immunization [8, 9, 10]. These methods consider different information diffusion strategies [11, 12, 13] to identify harmful nodes and stop the spread within the social network of harmful content.

Sexism, a type of harmful content, is the act of discriminating against another person based on their gender. For the task of EXIST2023 we will be leveraging social-media data and tackling it in 3 NLP-oriented tasks, namely classification tasks. With previous experience and promising results using transformers [13] and confirmed in EXIST2021 by [14], we plan using a more-task optimized transformer, to tackle the multi-lingual problem, namely XLM-RoBERTa [15]. While classical ML models offer good results, as shown in [16] or [14] where the authors obtain better results with them on the test set, transformers manage to obtain a better representation.

## 3. Experiments

XLM-RoBERTa [15] is a pre-trained transformer model that leverages 2.5TB of filtered CommonCrawl data containing 100 languages and is the baseline for our experiments. After experimenting with it and using multiple metrics we have decided to use a more powerful model, MiniLMv2[17], which offers a boosted performance at a smaller size. This multilingual model can perform natural language inference (NLI) on 100+ languages and is therefore also suitable for multilingual zero-shot classification.

The final pipeline for the task uses MiniLMv2 with distinct data-processing modules and metric functions. The choosing of the best parameters is done manually, at this stage, because the fine-tuning process uses all the training data, without splits, and the model overfits after too many epochs.

### 3.1. Metrics used by the Event

As mentioned on the official site, the metrics used in the competition, for which the engine will be optimized, are:

- ICM-Hard
- ICM-Hard Norm
- F1
- Cross Entropy
- Majority class
- Minority class
- Oracle most voted

### 3.2. Experimental Setup

After experimenting with XLM-Roberta we have decided to move on to MiniLMv2 as the results for 3 epochs with the best setups in both cases can be seen in table ??, as in terms of performance XLM-Roberta can achieve 1 epoch in 35 minutes, while MiniLMv2 can in 1.5 minutes

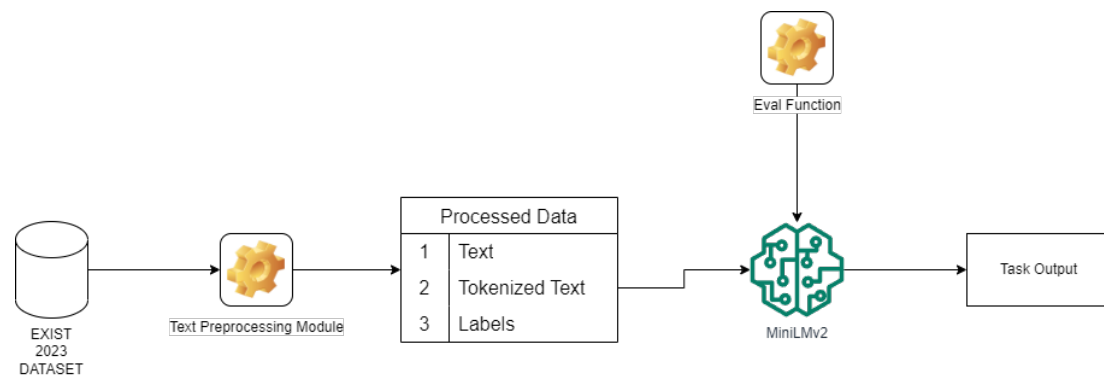
**Table 1**

MiniLM and xlm-roberta after 3 epochs with best parameters

	Train Loss	Test Loss	Test F1
<b>MiniLM-L12-H384</b>	0.4417	0.5266	0.7538
<b>xlm-roberta-base</b>	0.5234	0.5565	0.7475

Our proposed solution, as can be seen in the following figure, uses the same model, MiniLMv2, for all tasks the 2 components that are changed are:

1. A Text Processing Module: for each task this results in a list of pairs of: Text, Tokenized Text and Label(s)
2. The evaluation function for which the model will improve



**Figure 1:** Generic Solution Architecture

In what follows we will describe these modules for each task and if there is any particular setup used for the fine-tuning, other than the recommended one for Hugging Face Pipelines. While the dataset contains much meta-data information regarding the annotators, that information

was not used in the current form of the solution and a point of future improvement for this will be considering the labels weighed task-specific.

For all the tasks the following hyper-parameters have been used:

- $learning\_rate = 2e^{-5}$
- $per\_device\_train\_batch\_size = 32$
- $per\_device\_eval\_batch\_size = 32$
- $weight\_decay = 0.01$

### 3.2.1. Task 1

The first task is a binary classification. The systems has to decide whether or not a given tweet contains sexist expressions or behaviours and the metric used for the evaluation is F1. For picking the label, we are using the majoritarian label, equally weighting all the annotations.

For this task we have noticed that the model overfits hard resulting in a more categorical class probability (the classes either 100% probability or 0% probability, nothing in between), so we have decided to allow it to train only for 10 epochs.

### 3.2.2. Task 2

Source Intention: Once a message has been classified as sexist, the second task aims to categorize the message according to the intention of the author, which provides insights in the role played by social networks on the emission and dissemination of sexist messages. For this task we are using a new model over the samples labeled as sexist by the classifier for Task 1.

We are using the same idea for obtaining the label, equal weight for all annotations, but this time F1 is weighted. This time we have managed to get 20 epochs without overfitting.

For this task we could have gone further with the training but we have noticed that some classes started to lose all their probability.

### 3.2.3. Task 3

The third task was a multi-label classification for each tweet labeled as sexist, by task1. For this part both modules are a bit more complicated.

For obtaining the label the text module computes the probability of each label regarding the number of annotators, with the same weight for each of the once more. This time we will predict the probability of each label.

For the metric we are using Mean Squared Error compared to the ground-truth probabilities trying to minimize the error. This resulted in us being able to run for about 20 epochs without major classes overfitting.

We have tried to use the pipeline-specific F1, but we have encountered issue with the current version of it, so we have decided to implement the metric ourselves. After a few experiments we have migrated to MSE as it behaves better.

### 3.3. Results

We have compiled the results in table 2 with the previous mentions, Tasks 1 and 2 use F1 as metric and Task 3 uses accuracy.

**Table 2**

Model Stats at the End of Training

	<b>Task 1</b>	<b>Task 2</b>	<b>Task 3</b>
<b>Loss</b>	0.1508	0.4756	0.3814
<b>Best Metric</b>	0.9539	0.7027	47.9045

## 4. Conclusions and Results

The outcomes of our approach can be observed in Table 3 of the official leaderboard. For a more comprehensive analysis, please refer to the Results Chapter available on the official site. As anticipated, our transformer models demonstrated favorable performance, as reflected in the achieved results. Notably, our models excelled in the Hard-Hard evaluation and outperformed others in the Soft-Soft evaluation. This discrepancy can be attributed to the relatively higher freedom allowed to certain scenarios, leading to an increased probability assigned by our model. Task 3 yielded the highest ranking; however, it is important to note that this result is contingent upon only labeling tweets identified as sexist by the Task 1 model, indicating room for improvement in subsequent steps.

Furthermore, it is necessary to adopt a different evaluation method now that the test dataset is provided. In the initial setup, we utilized the entire dataset for training purposes to maximize the model's specialization. However, this approach yielded unfavorable training evaluations and introduced ambiguity in the model evaluation process.

**Table 3**

Ranking in EXIST2023 competition

<b>Task</b>	<b>Rank</b>
Task 1 Soft-Soft	25
Task 1 Hard-Hard	40
Task 1 Hard-Soft	41
Task 2 Soft-Soft	11
Task 2 Hard-Hard	22
Task 2 Hard-Soft	27
Task 3 Soft-Soft	5
Task 3 Hard-Hard	12
Task 3 Hard-Soft	9

## 5. Further Improvements

As mentioned before, the model used for task 1 will be the key point in improving the other tasks. On top of that, we consider using transfer-learning techniques in order to further improve the task 1 model and to use it in the following tasks.

Another thing that can be done is to use the meta-data for the annotators to weigh each label considering that this is a sexist task.

One more direction to look at is using a different model, specialized for the 2 languages that the tweets were in. Additionally, we can leverage ensembles as they usually provide better results than just the models in their components.

One final thing that could improve is using a more specific type of transformer, one that uses tweet data, such as BERTweet[18].

## 6. Acknowledgements

I want to thank Ciprian-Octavian Truică, [ciprian.truica@upb.ro](mailto:ciprian.truica@upb.ro), and Elena-Simona Apostol, [elena.apostol@upb.ro](mailto:elena.apostol@upb.ro), for previous guidance and help with cosmetic changes in my documents.

## References

- [1] L. Plaza, J. Carrillo-de Albornoz, R. Morante, J. Gonzalo, E. Amigó, D. Spina, P. Rosso, Overview of exist 2023: sexism identification in social networks, in: Proceedings of ECIR'23, 2023, pp. 593–599. doi:10.1007/978-3-031-28241-6\_68.
- [2] V.-I. Ilie, C.-O. Truică, E.-S. Apostol, A. Paschke, Context-Aware Misinformation Detection: A Benchmark of Deep Learning Architectures Using Word Embeddings, IEEE Access 9 (2021) 162122–162146. doi:10.1109/access.2021.3132502.
- [3] C.-O. Truică, E.-S. Apostol, A. Paschke, Awakened at CheckThat! 2022: fake news detection using BiLSTM and sentence transformer, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF2022), 2022, pp. 749–757.
- [4] C.-O. Truică, E.-S. Apostol, It's All in the Embedding! Fake News Detection Using Document Embeddings, Mathematics 11 (2023) 508. doi:10.3390/math11030508.
- [5] E.-S. Apostol, C.-O. Truică, A. Paschke, ContCommRTD: A Distributed Content-based Misinformation-aware Community Detection System for Real-Time Disaster Reporting, ArXiv preprint (2023). URL: <https://arxiv.org/abs/2301.12984>. arXiv:2301.12984.
- [6] C.-O. Truică, E.-S. Apostol, MisRoBÆRTa: Transformers versus Misinformation, Mathematics 10 (2022) 1–25(569). doi:10.3390/math10040569.
- [7] C.-O. Truică, E.-S. Apostol, P. Karras, DANES: Deep Neural Network Ensemble Architecture for Social and Textual Context-aware Fake News Detection, ArXiv preprint (2023). URL: <https://arxiv.org/abs/2302.01756>. arXiv:2302.01756.
- [8] A. Petrescu, C.-O. Truică, E.-S. Apostol, P. Karras, Sparse Shield: Social Network Immunization vs. Harmful Speech, in: ACM International Conference on Information and Knowledge Management (CIKM2021), ACM, 2021, pp. 1426–1436. doi:10.1145/3459637.3482481.

- [9] Ö. Coban, C.-O. Truică, E.-S. Apostol, CONTAIN: A Community-based Algorithm for Network Immunization, ArXiv preprint (2023). URL: <https://arxiv.org/abs/2303.01934>. arXiv:2303.01934.
- [10] C.-O. Truică, E.-S. Apostol, R.-C. Nicolescu, P. Karras, MCWDST: a Minimum-Cost Weighted Directed Spanning Tree Algorithm for Real-Time Fake News Mitigation in Social Media, ArXiv preprint (2023). URL: <https://arxiv.org/abs/2302.12190>. arXiv:2302.12190.
- [11] A. Petrescu, C.-O. Truică, E.-S. Apostol, Sentiment Analysis of Events in Social Media, in: 2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP), IEEE, 2019, pp. 143–149. doi:10.1109/iccp48234.2019.8959677.
- [12] C.-O. Truică, E.-S. Apostol, T. Ştefu, P. Karras, A Deep Learning Architecture for Audience Interest Prediction of News Topic on Social Media, in: International Conference on Extending Database Technology (EDBT2021), 2021, pp. 588–599. doi:10.5441/002/EDBT.2021.69.
- [13] A. Petrescu, C.-O. Truică, E.-S. Apostol, A. Paschke, EDSA-Ensemble: an Event Detection Sentiment Analysis Ensemble Architecture, 2023. arXiv:2301.12805.
- [14] S. Butt, N. Ashraf, G. Sidorov, A. F. Gelbukh, Sexism identification using bert and data augmentation-exist2021., in: IberLEF@ SEPLN, 2021, pp. 381–389.
- [15] 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.
- [16] I. Ameer, N. Ashraf, G. Sidorov, H. Gómez Adorno, Multi-label emotion classification using content-based features in twitter, *Computación y Sistemas* 24 (2020) 1159–1164.
- [17] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, M. Zhou, Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020. arXiv:2002.10957.
- [18] D. Q. Nguyen, T. Vu, A. T. Nguyen, Bertweet: A pre-trained language model for english tweets, 2020. arXiv:2005.10200.