# Efficient Multilingual Sexism Detection via Large Language Model Cascades

Notebook for the EXIST Lab at CLEF 2023

Lin Tian[1], Nannan Huang[1] and Xiuzhen Zhang[1,*]

[1]RMIT University, Melbourne, VIC 3000, Australia

Abstract

Sexism identification on social media platforms is an important task to promote gender equality by mitigating harmful stereotypes. In this report, we show how to leverage large language models for the EXIST challenge on all three tasks - automated detection of both English and Spanish tweets. Our submission, named Mario, is ranked first for the HARD label evaluation on both Task 1 and Task 2 and achieved the highest $F_1$ score of 0.8109 and 0.5711 respectively.

**Keywords**

Cascades Models, Automatic Sexism Categorisation, Automatic Sexism Detection, GPT-NeoX

## 1. Introduction

In this challenge, we explore different large language models based solutions for all three tasks of EXIST 2023 (sEXism Identification in Social neTworks) [1, 2], as part of CLEF 2023.

The challenge is divided into three tasks, namely Sexism Identification, Source Intention, and Sexism Categorisation, that collectively aim to classify, understand the intention, and categorise the facets of sexism in tweets to gain insights on the various forms of sexist expressions and behaviours on social media. For task 1, the objective is to perform a binary classification on tweets, segregating them into ones that manifest sexist expressions or behaviours and ones that do not. Task 2 aims to discern the underlying intent in tweets classified as sexist, categorising them into three classes: direct perpetration of sexism, reporting of experienced or observed sexism, and judgemental commentary on sexist situations or behaviours. The goal of the third task is to stratify the identified sexist tweets into five distinct categories reflective of the forms of sexism they exhibit: ideological and inequality, stereotyping and dominance, objectification, sexual violence, and misogyny and non-sexual violence. In this work, we attempted three tasks both in English and Spanish. Our approach utilises large language models with ensembling and cascading strategies for sexism identification on social media based on the text content.

In the provided dataset, there are two types of labels - hard and soft. For the hard labels, they are assigned by a majority vote of the annotations. Soft labels, on the other hand, are the entire

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

*Corresponding author.

lin.tian2@student.rmit.edu.au (L. Tian); amber.huang@student.rmit.edu.au (N. Huang); xiuzhen.zhang@rmit.edu.au (X. Zhang)

**Table 1**
Overall statistics of Task 1 Training Data.

| Language | Yes | No | Total |
|---|---|---|---|
| English | 1,331 | 1,983 | 3,314 |
| Spanish | 1,821 | 1,863 | 3,684 |
| All | 3,152 | 3,846 | 6,998 |

**Table 2**
Overall statistics of Task 2 Training Data.

| Language | Direct | Judgement | Reported | No | Total |
|---|---|---|---|---|---|
| English | 632 | 176 | 229 | 1,983 | 3,020 |
| Spanish | 866 | 283 | 305 | 1,863 | 3,317 |
| All | 1,498 | 459 | 534 | 3,846 | 6,337 |

set of human annotations with their variability, which is determined using the likelihood of each class. Note that, in tasks 1 and 2, which are mono label issues, the sum of the probabilities is equal to one. Whereas task 3 is a multi-label task, the probability sum there can be greater than one. We focus on the Hard labels for all three tasks - automated detection of multilingual sexism in social media posts. We design a system of cascades of language models for sexism detection. We also demonstrate an efficient way to utilise large language models, designed to speed up the inference time and maintain competitive performance. Our submission is ranked first among all 74 runs for tasks 1 and 2 on hard label evaluations and achieved a $F_1$ score of 0.8109 and a $F_1$ score for task 2 of 0.575. This shows that large language based cascade models are able to handle sexism identification and categorisation tasks confidently.

## 2. Related Work

Fundamentally, sexism identification is categorised as a subtask of abusive language detection. It shares a close relationship with a number of abusive language detection, including racism, hate speech, personal attacks, and others. We consider sexism identification a problem of text classification, where the models will classify which predefined labels a given text or tweet belongs to.

There have been studies done to identify sexism in the text on social media platforms [3, 4, 5, 6, 7]. Most approaches use deep learning based methods to tackle this task. Jha and Mamidi [7] included simple machine learning baselines (Support Vector Machines and FastText classifier) to classify tweets into three categories (hostile, benevolent and others). Sharifirad et al. [5] adopted text augmentation techniques and text generation data from ConceptNet and Wikidata to boost the model performance. Some related datasets have also been released to promote further research in this line [8, 9, 10].

**Table 3**
Task 3 training data hard labels distribution.

| Label | Total |
|---|---|
| No | 3,846 |
| Objectification | 1,286 |
| Sexual-violence | 798 |
| Stereotyping-dominance | 1,664 |
| Ideological-inequality | 1,325 |
| Misogyny-non-sexual-violence | 1,014 |

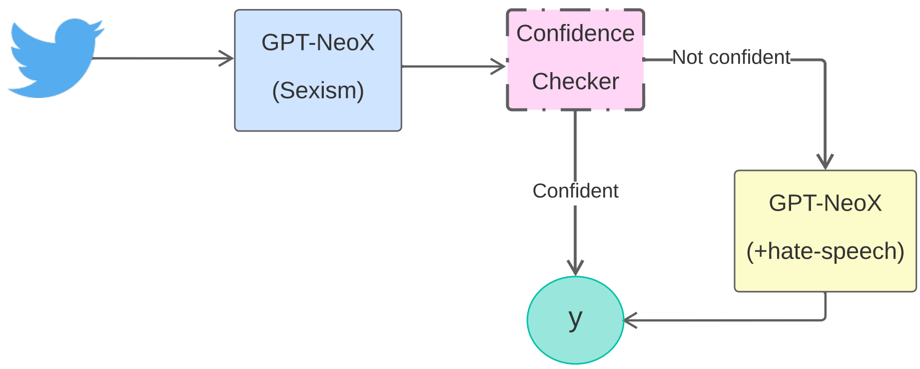


**Figure 1:** Overall Model Architecture.

## 3. Dataset

The given dataset contains both English and Spanish tweets for all three tasks. Since we only focus on the hard label identification task, the problem is formulated as a binary text classification task, a multi-class classification task, and a multi-label multi-class text classification task for tasks 1, 2 and 3.

For task 1, a total of 6,998 tweets are used for training. Among them, we randomly sampled 700 instances to use as a development dataset and the rest as a training dataset. Table 1 shows the overall statistics of the training data we used for task 1. Noted that we removed the "NOT_FOUND" instances when we trained all the models for task 1. For this task, both "Yes" and "No" labels are well balanced in distribution as shown in Table 1.

Task 2 was formulated as a four-class text classification problem with assigned hard labels. We follow the same approach by removing "NOT_FOUND" hard labelled instances from the training set, detailed data statistics can be referred to in Table 2. For both English and Spanish data, it has an unbalanced distribution across the "Direct", "Judgement", and "Reported" labels. The "No" instances are shared across all jobs because they all use the same source data.

Task 3 focuses on the classification of sexism. We treated it as a multi-label multi-class text classification task, focusing on the hard labels. Five different categories are given ("Ideological and inequality", "Stereotyping and dominance", "Objectification", "Sexual violence" and "Misogyny and non-sexual violence"). The label distribution is shown in Figure 3.

# 4. Methodology

## 4.1. Model Training

Compared with GPT-3 [11], ChatGPT [12] and GPT-4 [13], we adopt open-source GPT-NeoX [14] and BERTIN-GPT-J-6B [15] as our backbone models for all the experiments. The GPT-J model is a GPT-2-like causal language model trained on the Pile dataset [16]. The BERTIN-GPT-J-6B model shares the same model architecture with training data in Spanish [17].

## 4.2. Cascade Models

As shown in Figure 1, two GPT based large language models are included in our cascades. One model is fine-tuned with in-domain training data for three tasks, and the other hate-speech boosted model is sequentially fine-tuned on several hate speech datasets [18, 19, 20, 21] and an open-sourced hate-speech tweets dataset from the huggingface library [22] in the target language (English or Spanish) and then fine-tuned with in-domain task specific training data.

The confidence checker is working as the confidence-score based filter to distinguish the hard samples from the easy ones. We use a threshold on the confidence score to determine when to exit the cascade. The confidence threshold is one of the hyper-parameters in our settings. The final confidence threshold is picked depending on the best performance on our development set.

To highlight the practical benefit of cascades, it saves computation costs and improves inference speed compared to ensemble models. Based on our experiments, the cascade models yielded the best performance on the development dataset.

## 4.3. Label Smoothing

One of the common problems with large language models is their overconfidence in prediction tasks. Label smoothing prevents the network from becoming overconfident and has been used in many state-of-the-art models, including image classification, language translation, and speech recognition. Label smoothing is a simple yet effective regularisation tool that operates on labels. The intuition behind label smoothing is to not let the model learn that a specific input results in only a specific output.

Instead of using one-hot encoded vectors ([0,1] in this case), we introduce noise distribution $u(y|x)$. Our new ground truth label for data $(x_i, y_i)$ would be

$$\begin{aligned} p'\left(y \mid x_i\right) &= (1-\varepsilon) p\left(y \mid x_i\right)+\varepsilon u\left(y \mid x_i\right) \\ &= \begin{cases}1-\varepsilon+\varepsilon u\left(y \mid x_i\right) & \text{if } y=y_i \\ \varepsilon u\left(y \mid x_i\right) & \text{otherwise}\end{cases} \end{aligned} \tag{1}$$

where $\varepsilon$ is a weight factor, $\varepsilon \in [0,1]$ and note that $\sum_{y=1}^{K} p'\left(y \mid x_i\right)=1$.

**Table 4**
Approach tested in each run.

| Run | Task 1 | Task 2 | Task 3 |
|---|---|---|---|
| Mario_1 | binary | - | confidence threshold as 0.4 |
| Mario_2 | multi-class | multi-class | confidence threshold as 0.3 |
| Mario_3 | combination of experts | - | confidence threshold as 0.25 |

**Table 5**
Hyper-parameters for model training.

| Name | Value |
|---|---|
| # epoch | 4 |
| batch size | 4 per device |
| learning rate | [1e-5, 1e-4, 2e-3] |
| temperature | 0 |
| training steps | [3420, 3850, 4250] |
| confidence threshold | [0.85, 0.92, 0.95] |

By applying this technique, the model becomes less confident with extremely confident labels. This is exactly what we wanted to avoid. As our cascade models are selected purely based on the confidence score, this leads to better estimates on easy and hard sample selections.

# 5. Experiments

## 5.1. Settings

For the GPT models, we used the huggingface library for our experiments [23]. For the GPT-Spanish model, we adopt the version from huggingface library [24].

Furthermore, inspired by Do and Ng [25], we included five more public datasets to improve the models' robustness and mitigate the performance variance. Models that achieved the best performance on our development dataset are used. Hyper-parameters are shown in Table 5. Data preprocessing has been applied for all the runs. We removed the @username when we preprocessed the tweet text.

As shown in Table 4, we submit 3 runs for task 1, 1 run for task 2, and 3 runs for task 3. For the Mario_2 in Task 1, we fine-tuned the base language model with multi-class classification loss and gold labelled data from task 2 to help the model better understand the difference between the true identified sexism tweets and the non-sexism tweets. For task 2, only Mario_2 fine-tuned with multi-class classification supervised learning loss was applied and submitted. For Task 3, the cascades models were trained with multi-label loss, and different confidence thresholds were set up for selecting the labels as the final hard labels. Note that all the soft labels are reported with the model's confidence scores.

**Table 6**
Results in the test data.

| Task | Lang | Model | Ranking | ICM-H | $F_1$ | ICM-S | Ranking |
|---|---|---|---|---|---|---|---|
| Task 1 | All | Mario_1 | 2 | 0.654 | 0.8058 | 0.4507 | 4 |
| Task 1 | All | Mario_2 | 3 | 0.612 | 0.8029 | 0.3634 | 3 |
| Task 1 | All | Mario_3 | 1 | 0.6575 | 0.8109 | 0.4719 | 2 |
| Task 1 | English | Mario_1 | 3 | 0.588 | 0.7626 | 0.1009 | 4 |
| Task 1 | English | Mario_2 | 10 | 0.5459 | 0.765 | 0.0038 | 17 |
| Task 1 | English | Mario_3 | 2 | 0.5996 | 0.7734 | 0.128 | 3 |
| Task 1 | Spanish | Mario_1 | 1 | 0.6995 | 0.8383 | 0.6826 | 3 |
| Task 1 | Spanish | Mario_2 | 3 | 0.6552 | 0.83 | 0.6071 | 4 |
| Task 1 | Spanish | Mario_3 | 2 | 0.6959 | 0.8387 | 0.6998 | 2 |
| Task 2 | All | Mario_2 | 1 | 0.4887 | 0.5715 | -5.8157 | 7 |
| Task 2 | English | Mario_2 | 1 | 0.3677 | 0.5224 | -7.1029 | 9 |
| Task 2 | Spanish | Mario_2 | 1 | 0.5711 | 0.6059 | -5.1329 | 6 |
| Task 3 | All | Mario_1 | 9 | 0.0896 | 0.5011 | -9.1398 | 3 |
| Task 3 | All | Mario_2 | 8 | 0.1228 | 0.5145 | -9.6735 | 5 |
| Task 3 | All | Mario_3 | 6 | 0.17 | 0.5323 | -10.2297 | 6 |
| Task 3 | English | Mario_1 | 10 | -0.0269 | 0.4595 | -10.8847 | 7 |
| Task 3 | English | Mario_2 | 9 | 0.0133 | 0.4772 | -11.4612 | 8 |
| Task 3 | English | Mario_3 | 7 | 0.0568 | 0.4971 | -11.9003 | 11 |
| Task 3 | Spanish | Mario_1 | 7 | 0.1779 | 0.5305 | -7.797 | 2 |
| Task 3 | Spanish | Mario_2 | 6 | 0.204 | 0.5405 | -8.2903 | 3 |
| Task 3 | Spanish | Mario_3 | 4 | 0.2562 | 0.5578 | -8.9369 | 4 |

### 5.2. Results

To evaluate the performance of the models, the official results are based on normalised ICM [26] and F1 scores. Our models performance results are included in Table 6.

As we only trained with hard labels, we focused on the performance on Hard-hard evaluation and Hard-soft evaluation. The ICM-hard and $F_1$ scores for Hard-hard evaluation and ICM-soft are included for Hard-soft evaluation. The soft labels are reported based on model confidence scores. Thus, the rankings are dropped for Hard-soft evaluations compared to Hard-hard evaluations. It also proves that there is no correlation between human disagreement and large language model confidence, as shown in [27].

## 6. Conclusion

In this paper, we propose a text-based sexism classifier with simple cascade models. We show the effectiveness of using large language models as the backbone and simple confidence-based cascade models for quicker inference. The utilisation of the cascade model further shows the benefits of filtering out the hard samples over the label smoothed confidence scores and achieving the best performance in sexism detection task 1 and sexism categorisation task 2.

In future work, we plan to explore the given soft labels and better understand how to leverage large language models and learn from human disagreements.

## References


[1] 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).

[2] 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. (2023).

[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. Samory, I. Sen, J. Kohne, F. Flöck, C. Wagner, "call me sexist, but...": Revisiting sexism detection using psychological scales and adversarial samples, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 15, 2021, pp. 573–584.

[5] S. Sharifirad, B. Jafarpour, S. Matwin, Boosting text classification performance on sexist tweets by text augmentation and text generation using a combination of knowledge graphs, in: Proceedings of the 2nd workshop on abusive language online (ALW2), 2018, pp. 107–114.

[6] 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.

[7] 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.

[8] S. Karlekar, M. Bansal, Safecity: Understanding diverse forms of sexual harassment personal stories, arXiv preprint arXiv:1809.04739 (2018).

[9] P. Parikh, H. Abburi, N. Chhaya, M. Gupta, V. Varma, Categorizing sexism and misogyny through neural approaches, ACM Transactions on the Web (TWEB) 15 (2021) 1–31.

[10] P. Parikh, H. Abburi, P. Badjatiya, R. Krishnan, N. Chhaya, M. Gupta, V. Varma, Multi-label categorization of accounts of sexism using a neural framework, arXiv preprint arXiv:1910.04602 (2019).

[11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[12] OpenAI, Chatgpt, https://openai.com/blog/chatgpt/ (2023a).
[13] OpenAI, Gpt-4 technical report, https://cdn.openai.com/papers/gpt-4.pdf (2023b).
[14] S. Black, S. Biderman, E. Hallahan, Q. Anthony, L. Gao, L. Golding, H. He, C. Leahy, K. McDonell, J. Phang, M. Pieler, U. S. Prashanth, S. Purohit, L. Reynolds, J. Tow, B. Wang, S. Weinbach, GPT-NeoX-20B: An open-source autoregressive language model, in: Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models, Association for Computational Linguistics, virtual+Dublin, 2022, pp. 95–136. URL: https://aclanthology.org/2022.bigscience-1.9. doi:10.18653/v1/2022.bigscience-1.9.
[15] bertin-project/bertin-gpt-j-6b, https://huggingface.co/bertin-project/bertin-gpt-j-6B, 2023. Accessed: 2023-05-01.
[16] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, et al., The pile: An 800gb dataset of diverse text for language modeling, arXiv preprint arXiv:2101.00027 (2020).
[17] bertin-project/mc4-es-sampled, https://huggingface.co/datasets/bertin-project/mc4-es-sampled, 2023. Accessed: 2023-05-01.
[18] E. Fersini, P. Rosso, M. Anzovino, Overview of the task on automatic misogyny identification at ibereval 2018., Ibereval@ sepln 2150 (2018) 214–228.
[19] M. Á. Álvarez-Carmona, E. Guzmán-Falcón, M. Montes-y Gómez, H. J. Escalante, L. Villasenor-Pineda, V. Reyes-Meza, A. Rico-Sulayes, Overview of mex-a3t at ibereval 2018: Authorship and aggressiveness analysis in mexican spanish tweets, in: Notebook papers of 3rd sepln workshop on evaluation of human language technologies for iberian languages (ibereval), seville, spain, volume 6, 2018.
[20] A. Gautam, P. Mathur, R. Gosangi, D. Mahata, R. Sawhney, R. R. Shah, # metooma: Multi-aspect annotations of tweets related to the metoo movement, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 14, 2020, pp. 209–216.
[21] P. Sachdeva, R. Barreto, G. Bacon, A. Sahn, C. von Vacano, C. Kennedy, The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism, in: Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022, European Language Resources Association, Marseille, France, 2022, pp. 83–94.
[22] datasets/tweets_hate_speech_detection, https://huggingface.co/datasets/tweets_hate_speech_detection, 2023. Accessed: 2023-05-01.
[23] Eleutherai/gpt-neox-20b, https://huggingface.co/EleutherAI/gpt-neox-20b, 2023. Accessed: 2023-05-01.
[24] bertin-project/bertin-gpt-j-6b-alpaca, https://huggingface.co/bertin-project/bertin-gpt-j-6B-alpaca, 2023. Accessed: 2023-05-01.
[25] C. B. Do, A. Y. Ng, Transfer learning for text classification, Advances in neural information processing systems 18 (2005).
[26] E. Amigó, 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), 2022, pp. 5809–5819.
[27] Y. Qu, K. Roitero, D. L. Barbera, D. Spina, S. Mizzaro, G. Demartini, Combining human and machine confidence in truthfulness assessment, ACM Journal of Data and Information Quality 15 (2022) 1–17.