Prompt Engineering for Identifying Sexism using GPT Mistral 7B

Notebook for the EXIST Lab at CLEF 2024

Marco Siino[1,*], Ilenia Tinnirello[2]

[1]University of Catania, Piazza Università 2, Catania, 95131, Italy
[2]University of Palermo, Piazza Marina 61, Palermo, 90133, Italy

Abstract

EXIST is an ongoing series of scientific events and collaborative tasks dedicated to identifying sexism in social networks. The goal of EXIST - in this case hosted at CLEF 2024 - is to encompass the full spectrum of sexist expressions, ranging from overt misogyny to more subtle forms that include implicit sexist behaviours. A binary classification is the first task. Systems must determine whether a particular tweet includes statements or actions that are sexist. In this paper, we discuss the application of a Mistral 7B model to address the task in the hard labelling setup for English and Spanish. Our approach leverages a Mistral 7B model along with a few-shot learning strategy and prompt engineering. Thanks to our approach, on the English test set, our best run achieved an F1 of 0.56, and on the Spanish test set, it achieved an F1 of 0.51. In the global ranking, our approach was able to obtain an F1 of 0.53. Our selected approach is able to outperform some of the baselines provided for the competition while outperforming other LLM-based approaches.

Keywords

GPT, sexism, mistral 7B, LLM, prompt engineering

1. Introduction

In recent years, Natural Language Processing (NLP) has been reshaped by Generative Pre-trained Transformer (GPT) models [1, 2], by managing text across various applications. EXIST is a series of scientific events and shared tasks dedicated to identifying sexism in social networks. It aims to address sexism in a comprehensive manner, encompassing explicit misogyny and more subtle expressions of implicit sexist behaviours (EXIST 2021, EXIST 2022, EXIST 2023). The fourth edition of the EXIST shared task takes place as a Lab hosted at CLEF 2024.

Social networks serve as major platforms for social complaints, activism, and movements such as #MeToo, #8M, and #Time'sUp, which have rapidly gained traction. Many women throughout the world have been able to report sexist incidents in real life, including violence and discrimination, thanks to these sites. Social media platforms, however, also aid in the propagation of sexism and other rude, abusive, and offensive behaviours. In this situation, automated methods can be quite helpful in identifying and raising awareness of sexist discourses and behaviours. These methods can also be used to determine the most prevalent types of sexism, assess the frequency of abusive and sexist situations on social media platforms, and comprehend the ways in which sexism manifests itself in these medium. This lab helps with the creation of sexism detection applications. The activities in the latest version are centred upon visuals, namely memes, whereas the previous three editions were solely concerned with identifying and categorizing sexist text messages. Memes, which are usually funny pictures that become viral on the Internet and social media, are now included to encompass a wider range of sexist expressions, particularly those that pass for comedy. Consequently, the development of automated multimodal technologies that can identify sexism in text and memes is imperative.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
*Corresponding author.
marco.siino@unipa.it (M. Siino); ilenia.tinnirello@unipa.it (I. Tinnirello)
https://github.com/marco-siino (M. Siino)
0000-0002-4453-5352 (M. Siino); 0000-0002-1305-0248 (I. Tinnirello)

Detecting sexist content online is an increasingly complex challenge, requiring the development of automated tools for data extraction and categorization. These tools are essential for addressing both established and emerging societal concerns. Recent advancements in machine learning and deep learning architectures in almost every field [3, 4, 5] have driven a surge of interest also in natural language processing (NLP) techniques. Capitalizing on this momentum in NLP research, numerous text classification strategies have been proposed in the literature to automate the identification and categorization of online textual content. In the last fifteen years, some of the most successful strategies have been based on SVM [6, 7], on Convolutional Neural Network (CNN) [8, 9], on Graph Neural Network (GNN) [10], on ensemble models [11, 12] and, recently, on Transformers [1, 13, 14, 15].

Recently, the many approaches presented at SemEval 2024 - usually held before the CLEF conference - have further pushed the growing use of the large language model (LLM)-based architectures in academic research. LLM apps are used at SemEval to take on a variety of tasks and provide noteworthy outcomes. For example, T5 is used to the problem of determining the inference relation between plain language statements and Clinical Trial Reports [16] in Task 2 [17]. In Task 10, Hindi-English code-mixed conversations are subjected to emotion recognition in Conversation (ERC) using a Mistral 7B model [18]. Furthermore, a DistilBERT model is used in Task 8 [19] to recognize text generated by machines [20]. Inspired by the results provided by this last work, we decided to employ a Mistral 7B model to face the EXIST 2024 binary classification task (i.e., Task 1). EXIST 2024 [21, 22] finds its main basis on the previous editions of the same series [23, 24].

Classifying items in binary form is the first task. The systems must determine whether a particular tweet includes sexist language or actions (that is, if it is sexist in and of itself, portrays a sexist scenario, or disparages a sexist action). We provide a Transformer-based strategy that uses Mistral 7B to tackle the problem in both English and Spanish [25]. We applied the model in a specific few-shot manner that is covered in the remainder of this study. In particular, we provided samples from the English and Spanish training sets. We chose Mistral 7B because its comparison examination with two other state-of-the-art models—Llama 2 and Llama 1—shows significant improvements in common natural language processing tasks. Mistral 7B continuously performs better than Llama 2, a well-known open 13B model, in several benchmark tests. Additionally, as reported in its introductory paper, Mistral 7B performs better than Llama 1, a cutting-edge 34B model, not just equalling but surpassing its accomplishments in areas related to logic, maths, and coding.

The rest of the paper is developed in this manner. In Section 2, we provide some background on Task 1 hosted at EXIST 2024. An explanation of the employed technique is provided in Section 3. In Section 4, we detail the experimental configuration that was utilized to replicate our findings. Section 5 contains some discussions and the official task results. In section 6, we offer our conclusions and recommendations for additional study.

We make all the code publicly available and reusable on GitHub.

2. Background

This section furnishes background information regarding the Task 1, held at EXIST 2024. This text describes a challenge for participants to develop models that can detect sexism in tweets from Twitter. The challenge seeks creation of multilingual and monolingual models that can both detect sexism in terms of binary classification. These models, given a source content, need to identify if some form of sexism is present within the content.

For our submission, we only addressed the Task 1 (binary classification task), where we were asked to detect if a tweet contained some sexist content or not. An example from the official Task description is shown in the Figure 1.

Finally, the task organizers requested the submission of a JSON file. In our case, we submitted two runs. In one field of the JSON file is reported the ID of the test sample considered, in the second field it is reported the label (i.e., *YES* or *NO*).

- **SEXIST**:
  - *“Mujer al volante, tenga cuidado!”*
  - *“People really try to convince women with little to no ass that they should go out and buy a body. Like bih, I don’t need a fat ass to get a man. Never have.”*
- **NOT SEXIST**:
  - *“Alguien me explica que zorra hace la gente en el cajero que se demora tanto.”*
  - *"@messyworldorder it’s honestly so embarrassing to watch and they’ll be like “not all white women are like that”"*

Figure 1: In the Figure is shown a sample from the task description page. The output of the model for the task has to be one out of YES or NO.

3. System Overview

While it has been established that Transformers may not always be the optimal choice for text classification tasks [26], the efficacy of various strategies, such as domain-specific fine-tuning [27, 28] and data augmentation [29, 30], depends on the specific objectives.

The increasing adoption of Transformer-based architectures in academic research has also been bolstered by various methodologies showcased at SemEval 2024. These methodologies tackle diverse tasks and yield noteworthy findings. For instance, at the Task 2 [17], where to address the challenge of identifying the inference relation between a plain language statement and Clinical Trial Reports is used T5 [16]; Task 4 [31] where is employed a Mistral 7B model to detect persuasion techniques in memes [32]; and Task 8 [19], that utilizes a DistilBERT model to identify machine-generated text [20].

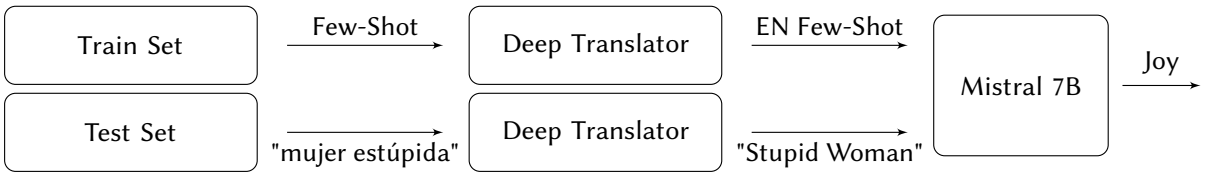

Figure 2: The system overview of our proposed approach in the case of the Spanish dataset. Given a set of Spanish samples from the train set, they are translated to English using Deep Translator. Then they are all provided as input - few-shot samples from the training set, together with a prompt question - to Mistral 7B. Following these few shot samples and the question as input, there is one sample from the test set for which the prediction has to be provided.

We employ Mistral 7B in our few-shot approach [33]. With seven billion parameters, Mistral 7B is a language model designed to be very performant. Mistral 7B outperforms the industry-leading open 13B model (Llama 2) on every benchmark that has been evaluated. Moreover, it outperforms the leading 34B model (Llama 1) in tasks pertaining to logic, math, and coding creation. The model leverages sliding window attention (SWA) to efficiently analyze sequences of different lengths while reducing inference costs, in addition to grouped-query attention (GQA) to accelerate inference. Furthermore, Mistral 7B - Instruct, a refined version designed for following instructions, performs better than Llama 2 13B - conversation model in both automatic and human assessments. Mistral 7B Instruct’s release highlights how easily the base model may be adjusted to provide considerable performance gains.

For our task, in the case of the Spanish language before prompting the model with the current sample from the test set, we made an online and real-time use of *Google Translator* from the *deep_translator* library. Then we randomly selected 10 samples from the provided labelled training set. Then we formatted the samples in each set in the following way:

Tweet1 // NO
Tweet2 // YES

...
TweetN // NO

After merging the formatted samples from the training set, we fed the model, appending to the few-shot samples the current unlabelled sample from the official test set. At this point, the full text containing the few-shot samples plus the sample to be classified were provided as prompts to Mistral.

Then the question provided as prompt to the model was: *" [INST] Is the following TWEET sexist, in any form, or does it describe situations in which such discrimination occurs (i.e., it is sexist itself, describes a sexist situation or criticizes a sexist behaviour)? Reply only with YES or NO."*. Where the *CONTEXT* were the few-shot samples provided. For all the samples from the test set, the model correctly predicted one of the response requested (i.e., YES or NO).

According to a recent study by [34], preprocessing usually has little effect on text classification tasks when employing Transformers. More precisely, when it comes to Transformers, the optimal set of preprocessing techniques does not really vary from doing none at all. We have not preprocessed the text in any way because of these factors, as well as the desire to maintain the speed and computational efficiency of our system.

4. Experimental Setup

Our model implementation was executed on Google Colab, utilizing the Mistral 7B library from Hugging Face, specifically the Mistral-7B-Instruct-v0.2-GGUF version from *TheBlock*. Additionally, we utilized the *deep_translator* package with Google Translator [1] for the translation task. The Mistral 7B version employed represents an enhanced iteration of the Mistral-7B-Instruct-v0.1 model, geared towards instruction fine-tuning. Instructions for instruction fine-tuning should be enclosed within [INST] and [/INST] tokens, with the initial instruction beginning with a sentence identifier, and subsequent instructions omitting this identifier. The generation process is terminated by the end-of-sentence token ID. Furthermore, we imported the Llama library [35] from *llama_cpp*, with comprehensive details available on GitHub.

All datasets required for the various phases of the experiment are accessible on the Official Competition page. No additional fine-tuning was conducted on the model. The experiment was executed using a T4 GPU provided by Google. Upon generating the predictions, the results were exported in the format specified by the organizers. As previously mentioned, our complete codebase is accessible on GitHub.

5. Results

To compile the final ranking, the official ranking metric used was the ICM normalized. Also, the ICM and the F1-score based on gold labels *YES* were reported. However, it is worth mentioning that also the ICM and the ICM normalized were provided in the final ranking.

In the Table 1, the global results obtained by the first three participants and by the last one are shown along with our submissions. While we do not know the details of other participants' implementations, we can notice that there is a relevant gap with our team. Also, the results of our two runs are comparable. As already stated, our approach is based on the application of prompt engineering using Mistral 7B. It is worth mentioning that the only difference between our two submissions is in the position of the tag $< s >$ used by Mistral. In the Table 2, we show the results for the English language obtained by the first three participants and by the last one are shown along with our submissions. In this case, we notice a greater gap with the last ranked submission. However, our best result is slightly better than the one obtained for the global ranking. Finally, in the Table 3, the results obtained by the first three participants and by the last one for the Spanish language are shown along with our submission. In this case, our approach obtained the worst results compared to the first positions.

[1] https://pypi.org/project/deep-translator/

Table 1
Performance of participant models for the global ranking in the *hard* setting. Results are sorted according to the ICM. Our two runs ranked 56 and 57 respectively.

Pos	Participant	ICM-Hard	ICM-Hard Norm	F1
1	NYCU-NLP_1.json	0.597	0.800	0.794
2	ABCD Team_1.json	0.596	0.799	0.783
3	CIMAT-CS-NLP_2.json	0.593	0.798	0.790
56	mc-mistral_2.json	0.061	0.531	0.532
57	mc-mistral_1.json	-0.009	0.495	0.478
70	The-Three-Musketeers_3.json	-0.464	0.266	0.300

Table 2
Performance of participant models for the English language in the *hard* setting. Results are sorted according to the ICM. Our two runs ranked 59 and 61 respectively.

Pos	Participant	ICM-Hard	ICM-Hard Norm	F1
1	EquityExplorers_2.json	0.618	0.815	0.761
2	EquityExplorers_1.json	0.595	0.804	0.749
3	I2C-UHU_2.json	0.580	0.796	0.763
59	mc-mistral_2.json	0.142	0.572	0.563
61	mc-mistral_1.json	0.076	0.539	0.519
66	shm2024_2.json	-0.367	0.313	0.462

Table 3
Performance of participant models for the Spanish language in the *hard* setting. Results are sorted according to the ICM. Our two runs ranked 58 and 59 respectively.

Pos	Participant	ICM-Hard	ICM-Hard Norm	F1
1	NYCU-NLP_1.json	0.621	0.811	0.824
2	ABCD Team_1.json	0.617	0.808	0.810
3	CIMAT-CS-NLP_2.json	0.610	0.805	0.815
58	mc-mistral_2.json	-0.012	0.494	0.507
59	mc-mistral_1.json	-0.087	0.456	0.444
66	UniLeon-UniBO_1.json	-0.511	0.245	0.607

Unfortunately, it is not easy from our perspective to motivate the actual gap with the best performing team. It is also worth noticing that our approach is ranked better in the case of the English language. This result is shown in the Table 2. Compared to the best performing models, our simple approach exhibits some room for improvements, although it is able to outperform some of the baseline provided. However, it is worth notice that it required no further pre-training and the computational cost to address the task is manageable with the free online resources offered by Google Colab. Furthermore, our approach made use of a quantized version of Mistral 7B available on Hugging Face and referenced in our code available on GitHub.

6. Conclusion

This paper presents the application of Mistral 7B-model for addressing the Task 1 at EXIST 2024 hosted at CLEF 2024. In our submission, we opted to adopt a few-shot learning strategy, utilizing an in-domain pre-trained Transformer model without modifications. Through numerous experimental iterations, we discovered the efficacy of constructing a prompt comprising examples extracted from the training dataset. Subsequently, we presented few-shot samples alongside a test sample as the prompt. The model's objective was to discern whether a sexist content is within a tweet. Undoubtedly, tackling this task presented considerable challenges, and despite our dedicated efforts, it's evident from the analysis of the final ranking that there is still considerable room for improvement. Potential

alternative avenues for exploration include leveraging the zero-shot capabilities of alternative models such as GPT and T5, expanding the training dataset by incorporating additional data sources, or implementing a novel approach to integrate domain-specific ontology-based knowledge, departing from the methodology outlined in our current work. These strategies hold promise for advancing the effectiveness and robustness of our model in identifying and addressing hallucinations [36]. Additional enhancements could be achieved through fine-tuning the model and reframing the problem as a distinct text classification task. Fine-tuning would allow the model to adapt more closely to the specific characteristics of our dataset and the nuances of the sexism identification task. By approaching the problem from a different classification perspective, we may uncover alternative feature representations or model architectures better suited to capturing the subtle distinctions between sexist and not sexist content. This strategic shift could potentially lead to more refined and accurate predictions, ultimately improving the overall performance of our system. Furthermore, given the interesting results recently provided on a plethora of tasks, also other few-shot learning [37, 38, 39, 40] or data augmentation strategies [41, 42, 43, 44] could be employed to improve the results. Upon reviewing the final ranking, it becomes apparent that our straightforward approach reveals areas where enhancements could be made. Nevertheless, it is noteworthy that our method necessitated no additional pre-training and remained computationally feasible using the resources provided by Google Colab. Furthermore, the proposed approach enabled us to surpass some baselines established by the task organizers. This achievement underscores the efficacy and accessibility of our methodology, despite its potential for further refinement.

Acknowledgments

We extend our gratitude to the anonymous reviewers for their insightful comments and valuable suggestions, which have significantly enhanced the clarity and presentation of this paper.

Authorship Contribution

Marco Siino: Conceptualization, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing - Original draft, writing - review & editing. **Ilenia Tinnirello:** Writing - review & editing, Methodology.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[2] G. Yenduri, M. Ramalingam, G. C. Selvi, Y. Supriya, G. Srivastava, P. K. R. Maddikunta, G. D. Raj, R. H. Jhaveri, B. Prabadevi, W. Wang, et al., Gpt (generative pre-trained transformer)–a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions, IEEE Access (2024).

[3] A. Sahu, P. K. Das, S. Meher, Recent advancements in machine learning and deep learning-based breast cancer detection using mammograms, Physica Medica 114 (2023) 103138.

[4] K. Sharifani, M. Amini, Machine learning and deep learning: A review of methods and applications, World Information Technology and Engineering Journal 10 (2023) 3897–3904.

[5] A. Nicosia, N. Cancilla, M. Siino, M. Passerini, F. Sau, I. Tinnirello, A. Cipollina, Alarms Early Detection in Dialytic Therapies via Machine Learning Models, in: T. Jarm, R. Šmerc, S. Mahnič-Kalamiza (Eds.), 9th European Medical and Biological Engineering Conference, Springer Nature Switzerland, Cham, 2024, pp. 55–66.

[6] F. Colas, P. Brazdil, Comparison of svm and some older classification algorithms in text classification tasks, in: IFIP International Conference on Artificial Intelligence in Theory and Practice, Springer, 2006, pp. 169–178.

[7] D. Croce, D. Garlisi, M. Siino, An SVM ensemble approach to detect irony and stereotype spreaders on twitter, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, volume 3180 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 2426–2432.

[8] Y. Kim, Convolutional neural networks for sentence classification, in: A. Moschitti, B. Pang, W. Daelemans (Eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, ACL, 2014, pp. 1746–1751. doi:10.3115/V1/D14-1181.

[9] M. Siino, E. Di Nuovo, I. Tinnirello, M. La Cascia, Detection of hate speech spreaders using convolutional neural networks, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021, volume 2936 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 2126–2136.

[10] F. Lomonaco, G. Donabauer, M. Siino, COURAGE at checkthat!-2022: Harmful tweet detection using graph neural networks and ELECTRA, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, volume 3180 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 573–583.

[11] M. Miri, M. B. Dowlatshahi, A. Hashemi, M. K. Rafsanjani, B. B. Gupta, W. Alhalabi, Ensemble feature selection for multi-label text classification: An intelligent order statistics approach, International Journal of Intelligent Systems 37 (2022) 11319–11341.

[12] M. Siino, I. Tinnirello, M. La Cascia, T100: A modern classic ensemble to profile irony and stereotype spreaders, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, volume 3180 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 2666–2674.

[13] M. Siino, M. La Cascia, I. Tinnirello, Mcrock at semeval-2022 task 4: Patronizing and condescending language detection using multi-channel cnn, hybrid lstm, distilbert and xlnet, in: G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, S. Ratan (Eds.), Proceedings of the 16th International Workshop on Semantic Evaluation, SemEval@NAACL 2022, Seattle, Washington, United States, July 14-15, 2022, Association for Computational Linguistics, 2022, pp. 409–417. doi:10.18653/V1/2022.SEMEVAL-1.55.

[14] M. Siino, DeBERTa at SemEval-2024 Task 9: Using DeBERTa for Defying Common Sense, in: A. K. Ojha, A. S. Doğruöz, H. Tayyar Madabushi, G. Da San Martino, S. Rosenthal, A. Rosá (Eds.), Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 291–297.

[15] M. Siino, All-Mpnet at SemEval-2024 Task 1: Application of Mpnet for Evaluating Semantic Textual Relatedness, in: A. K. Ojha, A. S. Doğruöz, H. Tayyar Madabushi, G. Da San Martino, S. Rosenthal, A. Rosá (Eds.), Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 379–384.

[16] M. Siino, T5-medical at semeval-2024 task 2: Using t5 medical embeddings for natural language inference on clinical trial data, in: Proceedings of the 18th International Workshop on Semantic Evaluation, SemEval 2024, Mexico City, Mexico, 2024, pp. 40–46.

[17] M. Jullien, M. Valentino, A. Freitas, SemEval-2024 task 2: Safe biomedical natural language inference for clinical trials, in: Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), Association for Computational Linguistics, 2024, pp. 1947–1962.

[18] M. Siino, Transmistral at semeval-2024 task 10: Using mistral 7b for emotion discovery and reasoning its flip in conversation, in: Proceedings of the 18th International Workshop on Semantic Evaluation, SemEval 2024, Mexico City, Mexico, 2024, pp. 298–304.

[19] Y. Wang, J. Mansurov, P. Ivanov, J. Su, A. Shelmanov, A. Tsvigun, C. Whitehouse, O. M. Afzal, T. Mahmoud, G. Puccetti, T. Arnold, A. F. Aji, N. Habash, I. Gurevych, P. Nakov, Semeval-2024 task 8: Multigenerator, multidomain, and multilingual black-box machine-generated text detection, in:

Proceedings of the 18th International Workshop on Semantic Evaluation, SemEval 2024, Mexico, Mexico, 2024, pp. 2057–2079.
[20] M. Siino, Badrock at semeval-2024 task 8: Distilbert to detect multigenerator, multidomain and multilingual black-box machine-generated text, in: Proceedings of the 18th International Workshop on Semantic Evaluation, SemEval 2024, Mexico City, Mexico, 2024, pp. 239–245.
[21] L. Plaza, J. Carrillo-de-Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of EXIST 2024 – Learning with Disagreement for Sexism Identification and Characterization in Social Networks and Memes, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), 2024.
[22] L. Plaza, J. Carrillo-de-Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of EXIST 2024 – Learning with Disagreement for Sexism Identification and Characterization in Social Networks and Memes (Extended Overview), in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.
[23] L. Plaza, J. Carrillo-de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2023–learning with disagreement for sexism identification and characterization, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2023, pp. 316–342.
[24] 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.
[25] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7b, arXiv preprint arXiv:2310.06825 (2023).
[26] M. Siino, E. Di Nuovo, I. Tinnirello, M. La Cascia, Fake news spreaders detection: Sometimes attention is not all you need, Information 13 (2022) 426. doi:10.3390/INFO13090426.
[27] C. Sun, X. Qiu, Y. Xu, X. Huang, How to fine-tune bert for text classification?, in: Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18, Springer, 2019, pp. 194–206.
[28] D. Van Thin, D. N. Hao, N. L.-T. Nguyen, Vietnamese sentiment analysis: An overview and comparative study of fine-tuning pretrained language models, ACM Transactions on Asian and Low-Resource Language Information Processing 22 (2023) 1–27.
[29] F. Lomonaco, M. Siino, M. Tesconi, Text enrichment with japanese language to profile cryptocurrency influencers, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 2708–2716.
[30] S. Mangione, M. Siino, G. Garbo, Improving irony and stereotype spreaders detection using data augmentation and convolutional neural network, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, volume 3180 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 2585–2593.
[31] D. Dimitrov, F. Alam, M. Hasanain, A. Hasnat, F. Silvestri, P. Nakov, G. Da San Martino, Semeval-2024 task 4: Multilingual detection of persuasion techniques in memes, in: Proceedings of the 18th International Workshop on Semantic Evaluation, SemEval 2024, Mexico City, Mexico, 2024, pp. 2009–2026.
[32] M. Siino, Mcrock at semeval-2024 task 4: Mistral 7b for multilingual detection of persuasion techniques in memes, in: Proceedings of the 18th International Workshop on Semantic Evaluation, SemEval 2024, Mexico City, Mexico, 2024, pp. 53–59.
[33] J. Littenberg-Tobias, G. R. Marvez, G. Hillaire, J. Reich, Comparing few-shot learning with GPT-3 to traditional machine learning approaches for classifying teacher simulation responses, in: AIED (2), volume 13356 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 471–474.
[34] M. Siino, I. Tinnirello, M. La Cascia, Is text preprocessing still worth the time? a comparative survey

on the influence of popular preprocessing methods on transformers and traditional classifiers, Information Systems 121 (2024) 102342.

[35] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. arXiv:2302.13971.

[36] M. Siino, BrainLlama at SemEval-2024 Task 6: Prompting Llama to detect hallucinations and related observable overgeneration mistakes, in: A. K. Ojha, A. S. Doğruöz, H. Tayyar Madabushi, G. Da San Martino, S. Rosenthal, A. Rosá (Eds.), Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 82–87.

[37] X. Wang, X. Wang, B. Jiang, B. Luo, Few-shot learning meets transformer: Unified query-support transformers for few-shot classification, IEEE Trans. Circuits Syst. Video Technol. 33 (2023) 7789–7802. doi:10.1109/TCSVT.2023.3282777.

[38] B. M. S. Maia, M. C. F. Ribeiro de Assis, L. M. de Lima, M. B. Rocha, H. G. Calente, M. L. A. Correa, D. R. Camisasca, R. A. Krohling, Transformers, convolutional neural networks, and few-shot learning for classification of histopathological images of oral cancer, Expert Systems with Applications 241 (2024) 122418. doi:https://doi.org/10.1016/j.eswa.2023.122418.

[39] M. Siino, M. Tesconi, I. Tinnirello, Profiling cryptocurrency influencers with few-shot learning using data augmentation and ELECTRA, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 2772–2781.

[40] Z. Meng, Z. Zhang, Y. Guan, J. Li, L. Cao, M. Zhu, J. Fan, F. Fan, A hierarchical transformer-based adaptive metric and joint-learning network for few-shot rolling bearing fault diagnosis, Measurement Science and Technology 35 (2024). doi:10.1088/1361-6501/ad11e9.

[41] F. Muftie, M. Haris, Indobert based data augmentation for indonesian text classification, in: 2023 International Conference on Information Technology Research and Innovation, ICITRI 2023, 2023, p. 128 – 132. doi:10.1109/ICITRI59340.2023.10250061.

[42] M. Siino, F. Lomonaco, P. Rosso, Backtranslate what you are saying and i will tell who you are, Expert Systems n/a (2024) e13568. doi:https://doi.org/10.1111/exsy.13568.

[43] J. M. Tapia-Téllez, H. J. Escalante, Data augmentation with transformers for text classification, in: L. Martínez-Villaseñor, O. Herrera-Alcántara, H. Ponce, F. A. Castro-Espinoza (Eds.), Advances in Computational Intelligence, Springer International Publishing, Cham, 2020, pp. 247–259.

[44] M. Siino, I. Tinnirello, XLNet with Data Augmentation to Profile Cryptocurrency Influencers, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 2763–2771.