GPT Hallucination Detection Through Prompt Engineering

Notebook for the ELOQUENT 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

Detecting hallucinated or factually inaccurate information from GPT models can be challenging for humans. Consequently, it is crucial to thoroughly test Large Language Models (LLMs) for their accuracy before deployment. One potential method for identifying hallucinated content, which is explored at the ELOQUENT 2024 Lab hosted at CLEF 2024, involves using LLMs to assess the output of other LLMs. In this paper, we discuss the application of a Mistral 7B model to address the task in the hard labelling setup for English and Swedish. 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 proposal achieved an F1 of 0.72, and on the Swedish test set, it achieved an F1 of 0.75. Our selected approach is able to outperform some of the baselines provided for the competition while outperforming other LLM-based approaches.

Keywords

GPT, hallucinations detection, 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 generating human-like text across various applications. Despite their impressive capabilities, these models exhibit a phenomenon known as "hallucination" [3, 4, 5] where they produce text that is plausible but factually incorrect or nonsensical. Understanding the hallucination phenomenon is crucial for improving the reliability and trustworthiness of these AI systems. Hallucination in the context of GPT models refers to the generation of content that is not grounded in the input data or real-world knowledge. This can manifest itself in several ways: a) *Factual Errors*: The model generates incorrect information about well-known facts, such as stating that "Paris is the capital of Italy" instead of France, b) *Incoherent Responses*: The model produces text that does not logically follow from the input, resulting in gibberish or irrelevant content, c) *Invented Details*: The model creates details or events that did not occur, which can be problematic in contexts requiring accuracy, such as news articles or scientific reports.

The detection of hallucinated content online presents a growing challenge, necessitating the development of automated tools for data extraction and categorization. These tools can address established and emerging societal concerns. Recent advancements in machine and deep learning architectures have fuelled a surge in interest towards NLP techniques. Building upon this surge in NLP research, several text classification strategies have been proposed in the literature to automate the identification and categorization of online textual content [6, 7]. In the last fifteen years, some of the most successful strategies have been based on SVM [8, 9], on Convolutional Neural Network (CNN) [10, 11], on Graph Neural Network (GNN) [12], on ensemble models [13, 14] and, recently, on Transformers [1, 15, 16, 17].

The surge in the adoption of LLM-based architectures within academic research has been further propelled by diverse methodologies showcased at SemEval 2024 where, regarding the Task 6 (namely, *SHROOM*), the ELOQUENT Task finds its main foundation. ELOQUENT and SHROOM were proposed

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)

independently around the same time, making them contemporaries. HalluciGen was not based on the SHROOM task; however, due to the timing of the two tasks, SHROOM's data was leveraged for the paraphrase scenario. It is important to note that SHROOM data was not used for the translation scenario. Additionally, SHROOM and ELOQUENT have different aims: ELOQUENT focuses on both generation and detection, with a specific interest in LLM systems, whereas SHROOM is solely a detection task and does not specify LLM systems. Also at SemEval, LLM applications address a range of tasks and yield notable outcomes. For instance, in Task 2 [18], T5 is utilized to confront the challenge of identifying the inference relation between plain language statements and Clinical Trial Reports [19]. In Task 10, a Mistral 7B model is employed to perform emotion Recognition in Conversation (ERC) within Hindi-English code-mixed conversations [20]. Additionally, in Task 8 [21], a DistilBERT model is leveraged to identify machine-generated text [22].

Finally, for the Task 2 at ELOQUENT 2024 – HalluciGen Detection – the organizers aim to develop robust LLM-based detectors for hallucinated content. To facilitate a cross-model evaluation, the first objective focuses on creating evaluators that can both identify and generate hallucinations. The second objective involves testing these evaluators on challenging hallucination cases to reveal their strengths and weaknesses. To face with the paraphrase scenario of the HalluciGen task, here we describe a system submitted for the English and the Swedish language, proposing a Transformer-based approach which made use of Mistral 7B [23]. We used the model in a particular few-shot way described in the rest of this paper. Specifically, we provided 16 and 20 samples from the English and from the Swedish training set, respectively. We opted for Mistral 7B because the comparative analysis between Mistral 7B and other leading models, namely Llama 2 and Llama 1, reveals noteworthy advancements in common NLP tasks. Across multiple benchmark evaluations, Mistral 7B consistently exhibits superior performance in comparison to Llama 2, a prominent open 13B model. Moreover, its efficacy extends beyond mere parity with, but rather exceeds, the achievements of Llama 1, a state-of-the-art 34B model, particularly in tasks pertaining to reasoning, mathematics, and code generation. Our findings are also supported by the final ranking where several Llama-based approaches underperform when compared to our approach which makes use of Mistral 7B.

This is how the remainder of the paper is developed. We give some background information on Task 2 hosted at ELOQUENT 2024 in Section 2. Section 3 offers an explanation of the methodology utilized. We describe the experimental setup used to reproduce our work in Section 4. The official task results and some discussions are given in Section 5. We provide our conclusion and suggestions for further research in section 6.

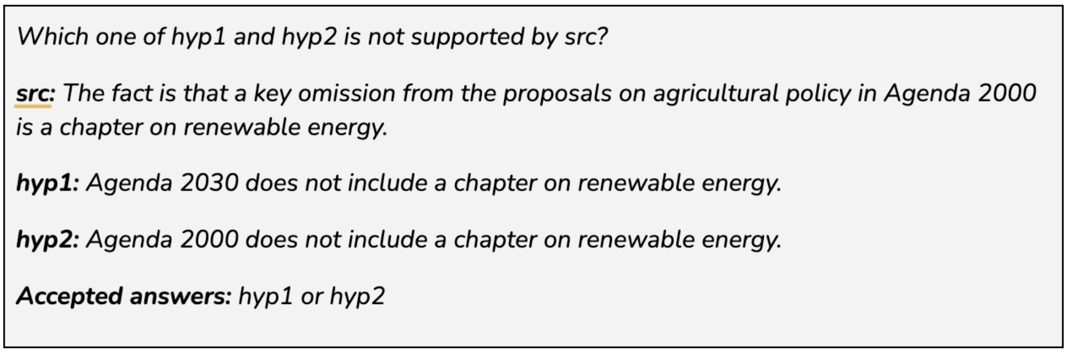
We make all the code publicly available and reusable on GitHub.

2. Task Description

This section furnishes background information regarding the Task 2, held at ELOQUENT 2024. This task describes a challenge for participants to develop models that can detect hallucinations in machine translation and paraphrase generation tasks. The challenge seeks creation of multilingual and monolingual models that can both detect and generate hallucinations in machine translation (evaluating two translations) and paraphrase generation (assessing a single paraphrase). These models, given a source sentence and potential outputs (hypotheses), need to identify nonsensical content (hallucinations) even without a reference translation or paraphrase for comparison.

For our submission, we only addressed the detection step, where we were asked to select which one out of two hypotheses provided was a hallucination. An example from the official Task description is shown in the Figure 1.

Finally, the task organizers requested the submission of a CSV file in the format shown in the Figure 2. In the first column is reported the ID of the test sample considered, in the second column it is reported which one of the two hypotheses is the hallucinated one and in the last column there is an optional field where it is possible to report any possible explanation provided by the model.



Which one of hyp1 and hyp2 is not supported by src?

src: *The fact is that a key omission from the proposals on agricultural policy in Agenda 2000 is a chapter on renewable energy.*

hyp1: *Agenda 2030 does not include a chapter on renewable energy.*

hyp2: *Agenda 2000 does not include a chapter on renewable energy.*

Accepted answers: *hyp1 or hyp2*

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 hyp1 or hyp2. In this case, the hallucinated hypothesis is hyp1.

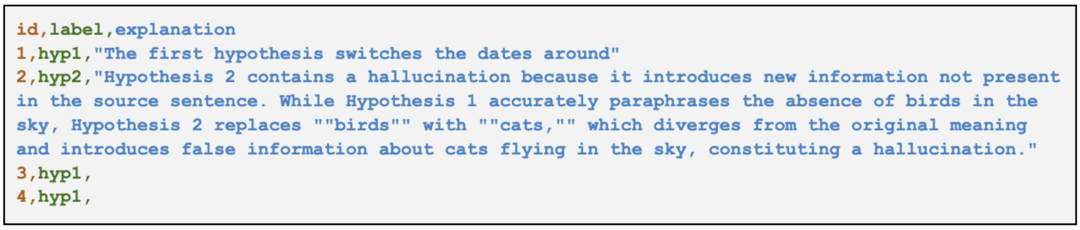
3. System Overview

Even if it has already been empirically shown on a few tasks (e.g., text classification, author profiling etc. [24, 25, 26, 27]) that Transformers alone are not necessarily the best option for performing text classification, depending on the goal some strategies like domain-specific fine-tuning [28, 29], or data augmentation [30, 31] can be beneficial for several applications.

As a starting point, we tried to leverage Mistral 7B in a zero-shot way. However, the zero-shot prompting with Mistral 7B Instruct likely resulted in hallucinations rather than providing the expected output labels (hyp1, hyp2) due to several factors: the model may not have been sufficiently trained on the specific task without additional context or examples, leading to reliance on its prior knowledge, which may not align perfectly with the task's requirements. Additionally, the prompts used may not have been clear or specific enough, resulting in open-ended responses instead of precise labels. The complexity of the task might involve nuances requiring a more guided or fine-tuned approach, and even state-of-the-art models like Mistral 7B Instruct have limitations in zero-shot scenarios, struggling without sufficient context and examples. Considering this preliminary findings, our approach is a few-shot one [32] and makes use of the above-mentioned Mistral 7B. Mistral 7B - specifically *Mistral-7B-Instruct-v0.2* from *Hugging Face* - is a language model equipped with 7 billion parameters, is designed to excel in both performance and efficiency. Compared to the leading open 13B model (Llama 2), Mistral 7B demonstrates superior performance across all evaluated benchmarks [23]. Moreover, it outperforms the top released 34B model (Llama 1) in tasks related to reasoning, mathematics, and code generation. The model leverages grouped-query attention (GQA) to expedite inference, along with sliding window attention (SWA) to efficiently process sequences of varying lengths while minimizing inference costs. Additionally, a fine-tuned variant, Mistral 7B – Instruct, tailored for adhering to instructions, surpasses the Llama 2 13B – chat model across both human and automated benchmarks. The introduction of Mistral 7B Instruct underscores the ease with which the base model can be fine-tuned to achieve notable performance enhancements. The Mistral 7B Instruct variant requires a specific input format, as stated below:

```
<s>[INST] Instruction [/INST] Model answer</s>[INST] Follow-up instruction [/INST]
```

Instruction, along with the following *Model answer*, can be a single sample with the related label or a set of sample/label pairs (realizing, in this case, a few-shot use of the model). Then, *Follow-up instruction* is the current sample for which the prediction has to be provided by the model. Specifically, we have prepared a few-shot text string containing samples from the training set along with their respective labels. At this point, the full text containing the training samples plus the sample to be classified were provided as prompt to Mistral. Then the question provided as prompt to mistral was: *"Which one of hyp1 and hyp2 is not supported by src?"*. To this request, the model replied with one option between *hyp1* or *hyp2*.



```
id,label,explanation
1,hyp1,"The first hypothesis switches the dates around"
2,hyp2,"Hypothesis 2 contains a hallucination because it introduces new information not present
in the source sentence. While Hypothesis 1 accurately paraphrases the absence of birds in the
sky, Hypothesis 2 replaces ""birds"" with ""cats,"" which diverges from the original meaning
and introduces false information about cats flying in the sky, constituting a hallucination."
3,hyp1,
4,hyp1,
```

Figure 2: In the Figure, it is shown the output format requested by the organizers for the detection task. One file is requested for each of the two languages (i.e., English and Swedish).

So, as an example from the test set, to the sentence: "Mr President, the approach adopted by the rapporteur to the Commission's 1999 annual economic report is comprehensive and also sensible." and the hypothesis 1: "The approach taken by the rapporteur to the 1997 annual economic report is comprehensive and sensible." and the hypothesis 2: "The approach taken by the rapporteur to the 1999 annual economic report is comprehensive and sensible." the model replied to the prompt: *hyp1*. It is important to mention that we also tried to use the model in a zero-shot configuration. In this case, we just asked the model to pick one of the two hypotheses as hallucinated content. Unfortunately, the model usually developed discussions as answers that, in most cases, did not identify the hallucinated hypothesis.

Finally, we collected all the predictions provided on the test set to into a CSV file with the required format to submit our predictions.

The one just discussed is the approach followed for the English test set. In the case of the Swedish test set, we made use of *deep_translator* from the *Google Translator* library, to translate samples into English before feeding Mistral 7B. Our preliminary experiments on feeding Mistral with the original Swedish samples did not provide relevant results.

As noted in the recent study by [26], the contribution of preprocessing for text classification tasks is generally not impactful when using Transformers. More specifically, the best combination of preprocessing strategies is not significantly different from performing no preprocessing at all in the case of the LLMs evaluated. For these reasons, and to keep our system fast and computationally light, we have not performed any preprocessing on the text. The low impact of the best preprocessing techniques - or combinations of techniques - using Transformers, as reported in the study, is due to several factors like preserving the quantity and the quality of the original information available.

4. Experimental Setup

We implemented our model on Google Colab. The library we used comes from Hugging Face and as already mentioned is Mistral 7B. We employed the v0.2 iteration of Mistral 7B, which represents an enhanced version of the Mistral-7B-Instruct-v0.1 model. To harness the capabilities of instruction fine-tuning, prompts must be enclosed within [INST] and [/INST] tokens. Additionally, the initial instruction should commence with a sentence identifier. The next instructions should not. The assistant generation will be ended by the end-of-sentence token ID. We also imported the Llama library [33] from *llama_cpp*. We did not perform any additional fine-tuning on the model. To run the experiment, a T4 GPU from Google has been used. After the generation of predictions, we exported the results on the format required by the organizers. As already mentioned, all of our code is available on GitHub.

5. Results

To compile the final ranking, the evaluation used the F1-score based on gold labels indicating which hypothesis contains the hallucination. These labels are human-annotated for the paraphrase task. However, it is worth mentioning that also the accuracy, the precision and the recall were provided in the final ranking.

Table 1
Performance of participant models for the English language. Results are sorted according to the F1-score. Our model ranked 6th.

Pos	Participant	acc	f1	prec	rec	Model
1	final_gpt4_en_v2_detection - Narjes Nikzad.csv	0.91	0.91	0.91	0.91	gpt-4-turbo
2	test_pg_english - harika vuppala.csv	0.90	0.90	0.91	0.90	Majority voting of different finetuned LLMs
3	majority_vote_result_en_narjes - Narjes Nikzad.csv	0.85	0.85	0.86	0.85	Majority vote on google/gemma-7B-it, meta-llama/Meta-Llama-3-8B-Instruct, gpt-3.5-turbo, gpt-4-turbo
4	final_llama3_en_v1_detection - Narjes Nikzad.csv	0.80	0.80	0.81	0.80	meta-llama/Meta-Llama-3-8B-Instruct
5	final_gpt_en_v1_detection - Narjes Nikzad.csv	0.73	0.73	0.83	0.73	gpt-3.5-turbo
6	eloquent2024_mc_mistral_en_prediction - Marco Siino.csv	0.73	0.72	0.73	0.73	Mistral 7B
7	final_gemma_en_v1_detection - Narjes Nikzad.csv	0.71	0.71	0.77	0.71	google/gemma-7B-it
8	final_llama3_prompt_narjes_en_v1_detection - Narjes Nikzad.csv	0.69	0.69	0.81	0.69	meta-llama/Meta-Llama-3-8B
9	final_gpt_en_narjes_detection - Narjes Nikzad.csv	0.68	0.68	0.75	0.68	gpt-3.5-turbo
10	final_gemma_en_vnarjes - Narjes Nikzad.csv	0.54	0.49	0.73	0.54	gemma-7B-it

In Table 1, the results obtained by the participants are shown along with the models used. While we do not know the details of other participants' implementations, we can notice that most of the submissions were made by the same team. Furthermore, it is not clear and fully explained the gap between the team at the position 5 and the team at position 9. While it seems that both used the same model (i.e., *GPT-3.5*) it is not easy from our perspective to motivate the actual gap. It is also worth noting that our approach ranked better in Swedish than in English. This result is shown in the Table 2 and can be motivated with the findings reported in [34], where the authors state that during the translation process some relevant semantic can be made more explicit. For both languages, GPT 4 Turbo appeared to be the best performing model according to the final ranking. 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 noticing 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 Task 2 at ELOQUENT 2024 hosted at CLEF 2024. For our submission, we decided to follow a few-shot learning approach, employing as-is, an in-domain pre-trained Transformer. After several experiments, we found it beneficial to build a prompt containing samples from the training set. Then we provide, as a prompt, the few-shot samples together with a test sample. The model was asked to select which hypotheses were an actual hallucination. The task is challenging, and there is still opportunity for improvement, as can be noted

Table 2
Performance of participant models for the Swedish language. Results are sorted according to the F1-score. Our model ranked 3rd.

Pos	Participant	acc	f1	prec	rec	Model
1	final_gpt4_se_v1_detection - Narjes Nikzad.csv	0.81	0.81	0.81	0.81	gpt-4-turbo
2	test_pg_swedish - harika vuppala.csv	0.80	0.79	0.82	0.80	Majority voting of different finetuned LLMs
3	eloquent2024_mc_mistral_sv_prediction - Marco Siino.csv	0.76	0.75	0.78	0.76	Mistral 7B
4	final_gpt_se_v1_detection - Narjes Nikzad.csv	0.71	0.70	0.76	0.71	gpt-3.5-turbo
5	majority_vote_result_se_narjes - Narjes Nikzad.csv	0.67	0.66	0.72	0.67	Majority vote on google/gemma-7B-it, meta-llama/Meta-Llama-3-8B-Instruct, gpt-3.5-turbo, gpt-4-turbo
6	final_gpt_se_narjes_detection - Narjes Nikzad.csv	0.61	0.60	0.65	0.61	gpt-3.5-turbo
7	final_llama3_se_v1_detection - Narjes Nikzad.csv	0.60	0.59	0.60	0.60	meta-llama/Meta-Llama-3-8B-Instruct
8	final_gemma_se_v1_detection - Narjes Nikzad.csv	0.59	0.52	0.71	0.59	gemma-7B-it
9	final_llama3_prompt_narjes_se_v2_detection - Narjes Nikzad.csv	0.57	0.48	0.77	0.57	meta-llama/Meta-Llama-3-8B
10	final_gemma_se_vnarjes - Narjes Nikzad.csv	0.07	0.11	0.47	0.07	google/gemma-7B-it

looking at the final ranking. Possible alternative approaches include utilizing the zero-shot capabilities of other models like GPT and T5, increasing the size of the few-shot set by using further data from the training set, or directly integrating ontology-based domain knowledge differently than what has been proposed in our work. Further improvements could be obtained with a fine-tuning and modelling the problem as a different text classification task. Furthermore, given the interesting results recently provided on a plethora of tasks, also other few-shot learning [35, 36, 37, 38] or data augmentation strategies [39, 34, 40, 41] could be employed to improve the results. Looking at the final ranking, our simple approach exhibits some room for improvements. However, it is worth noticing that required no further pre-training and the computational cost to address the task is manageable with the free online resources offered by Google Colab. Also, thanks to the proposed approach, we have been able to outperform the baseline provided by the task organizers.

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.

CRediT Authorship Contribution Statement

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] M. Lee, A mathematical investigation of hallucination and creativity in GPT models, Mathematics 11 (2023) 2320.

[4] S. A. Athaluri, S. V. Manthena, V. K. M. Kesapragada, V. Yarlagadda, T. Dave, R. T. S. Duddumpudi, Exploring the boundaries of reality: Investigating the phenomenon of artificial intelligence hallucination in scientific writing through chatgpt references, Cureus 15 (2023) e37432.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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. URL: https://doi.org/10.3115/v1/d14-1181. doi:10.3115/V1/D14-1181.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] M. Siino, McRock at SemEval-2024 task 4: Mistral 7B for multilingual detection of persuasion techniques in memes, 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. 53–59.

[17] M. Siino, Mistral at SemEval-2024 task 5: Mistral 7B for argument reasoning in civil procedure, 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. 155–162.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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).

[24] 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.

[25] J. Pizarro, Profiling Bots and Fake News Spreaders at PAN'19 and PAN'20: Bots and Gender Profiling 2019, Profiling Fake News Spreaders on Twitter 2020, in: Proceedings - 2020 IEEE 7th International Conference on Data Science and Advanced Analytics, DSAA 2020, 2020, p. 626 – 630. doi:10.1109/DSAA49011.2020.00088.

[26] 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.

[27] R. O. Bueno, B. Chulvi, F. Rangel, P. Rosso, E. Fersini, Profiling irony and stereotype spreaders on twitter (IROSTEREO). overview for PAN at CLEF 2022, 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. 2314–2343.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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`.

[34] 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`.

[35] 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. URL: https://doi.org/10.1109/TCSVT.2023.3282777. doi:`10.1109/TCSVT.2023.3282777`.

[36] 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. URL: https://www.sciencedirect.com/science/article/pii/S0957417423029202. doi:`https://doi.org/10.1016/j.eswa.2023.122418`.

[37] 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.

[38] 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). URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85180156886&doi=10.1088%2f1361-6501%2fad11e9&partnerID=40&md5=5cf48be6a9dc20b836051fb5c8a4c47b. doi:`10.1088/1361-6501/ad11e9`.

[39] 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`.

[40] 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.

[41] 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.

A. Online Resources

The source code of our submission is available via

- https://github.com/marco-siino/eloquent2024