Can Hallucination Reduction in LLMs Improve Online Sexism Detection?

Leyuan DING[1], Praboda Rajapaksha[1,2], Aung Kaung Myat[1],

Reza Farahbakhsh[1], and Noel Crespi[1]

[1]Samovar,Telecom SudParis,Institut Polytechnique de Paris,91120 Palaiseau,France.
[2]Department of Computer Science,Aberystwyth University,SY23 3DB Ceredigion,UK.
{leyuan.ding, praboda_rajapaksha, aung-kaung.myat,
reza.farahbakhsh, noel.crespi}@telecom-sudparis.eu

Abstract. Online sexism is a pervasive problem with a significant impact on the targeted individuals and social inequalities. Automated tools are now widely used to identify sexist content at scale, but most of these tools do not provide any further explanations beyond generic categories such as 'toxicity', 'abuse' or 'sexism'. This paper explores the impact of hallucination reduction in LLMs on enhancing sexism detection across three different levels: binary sexism, 4-categories of sexism, and fine-grained vectors, with a focus on explainability in sexism detection. We have successfully applied Neural Path Hunter (NPH) to GPT-2, with the purpose of "teaching" the model to hallucinate less. We have used hallucination-reduced GPT-2, achieving accuracy rates of 83.2% for binary detection, 52.2% for 4-categories classification and 38.0% for the 11-vectors fine-grained classification, respectively. The results indicate that: i) While the model performances may slightly lag behind the baseline models, hallucination-reducing methods have the potential to significantly influence LLM performance across various applications, beyond just dialogue-response systems. Additionally, this method could potentially mitigate model bias and improve generalization capabilities, based upon the dataset quality and the selected hallucination reduction technique.

Keywords: LLM, GPT-2, RoBERTa, Hallucination, Sexism Detection

1 Introduction

In response to the imperative need for enhancing harmony within online social platforms, the detection and classification of hate speech on social media have garnered significant attention. One distinctive sector of hate speech, sexism, has particularly emerged as a focal point of concern. Addressing this challenge, the task of detecting and classifying sexism has emerged as a prominent area of research due to its significant impact on victims. Sexist content can be subtle, implicit, and nuanced, and therefore is it difficult to distinguish it from non-sexist content. As a result, it is more challenging to detect sexist content when it comes to online platforms.

Automated tools are currently extensively utilized for large-scale identification of sexist content, mainly in online social media, providing support for tasks such as content moderation, monitoring, and research. However, many detection tools do not detect sexist content at a granular level and focus only on the main categories, such as toxicity, abuse, or sexism. Furthermore, there is a need for existing tools and research to place a greater emphasis on the explainability of sexism detection. To address this issue, the SemEval-2023 Task 10 [1] on the Explainable Detection of Online Sexism (EDOS) was organized hierarchically, which had been divided into three subtasks as a taxonomy; i) Binary sexism, ii) Category of sexism and iii) Fine-grained vectors.

We observed that almost all participants in this task used Transformer-based models, except one group which had applied GPT-3 for in-context learning. Even though BERT and GPT are both transformer-based, the majority of participants used BERT-based models. These models exhibited higher detection capabilities thanks to their bidirectional contextual understanding and task-specific embedding after pre-training with labelled data. On the other hand, GPT-based models are not as well-suited for understanding the meaning of the text, rather they generate the text. Moreover, GPT-based models are larger compared to BERT-based models, resulting in extended training times.

Hence, in this study, we try to explore whether we can use LLMs, particularly GPT-based models, for sexism detection. However, hallucination is a common problem in large language models, including GPT models. It occurs when the model generates text that is not factually accurate or realistic, as shown in figure 1 where we wanted a short explanation of the Large Language Model (LLM) while the answer generated is Master of Laws. This can be a problem when using LLMs for detection tasks.

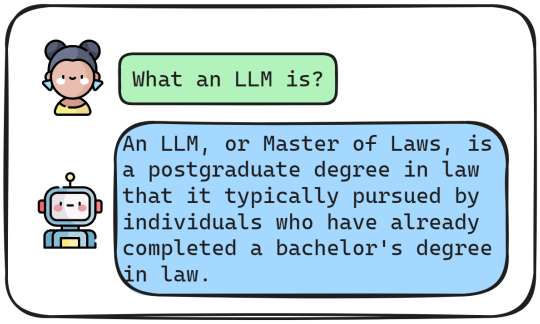

Fig. 1: Example of a hallucinated response

Hallucination of LLMs has been observed in multiple scenarios such as text, image, video, and audio as summarized in this survey [2]. The detection and reduction of models' hallucinations are becoming increasingly significant in improving model performance. ChatLaw [3], a combination of vector database re-

trieval with keyword retrieval, overcomes the hallucination problem and reduces the inaccuracy; Statement Accuracy Prediction, based on Language Model Activations (SAPLMA) [4], a trained classifier achieves an average of 71% to 83% accuracy on the binary true-false detection of generated contents. Since our task is online sexism detection, we want to figure out whether hallucination reduction can help in reducing false positives leading to improved detection performances and model explainability. Inspired by this work, adapting these LLMs to the social and cultural context of sexism is challenging without domain adaptation. Hence, in this research, we try to address this challenge by implementing GPT-2 with hallucination-reducing methods for the online sexism detection and classification tasks, to see the feasibility and effectiveness.

To address the EDOS challenge, we specifically used Generative Pre-Trained Transformer (GPT) [5]-based language models, as it is shown that GPT models have not exhibited better performances for this task [1]. Even though the BERT-based models outperformed GPT models in this EDOS challenge, the exploration of GPT models for this task and enhancement for other detection tasks remains valuable. This is primarily because GPT has a generative nature, allowing it to capture implicit and context-dependent expressions related to sexism. In addition, these LLMs significantly improved in large-scale context understanding, associated with continuous research and advancements in GPT architectures. In our experiments, we used Transformer-based models Roberta, Roberta-large [6] and Electra [7] as baseline models to compare the GPT-based model.

Our work comprehensively addresses all subtasks, aiming to achieve a more scalable perspective. The experimental results show that the GPT-2 model does not outperform other baseline models, however, exhibited equal false positive rates and not-sexist classification rates as Roberta, which is the best-performed model in our experiments. This is interesting as generative model performances for NLP detection tasks exhibit reasonable performances, and we can improve the classification by pre-training with more sexist content to reduce the false negative rate. Additionally, we plan to enhance these models by applying more hallucination-reducing methods and using them in broader natural language processing scenarios in the future.

The rest of the paper is organized as follows: Section 2 describes several novel hallucination-reducing frameworks, to provide an overview of methodologies facilitating the better performance of LLMs, especially in dialogue systems, chatbots and smart assistants for instance. Section 6 presents more precisely the 'learning patterns' of LLM proposed in our work. After that we have looked into the SemEval task from a deeper perspective, starting with several BERT-based [8] language models which are generally recognized as the "First choice" when doing text classification and analysis tasks, also the most used to address the problem proposed by competition task, this section also contains the most challenging part of our work, the implementation of the hallucination-less LLMs and apply it to the sexism detection task. Then the last part (Sec. 4) comes the interpretation of the results we have obtained, and last but not least, a discus-

sion of the limitations of this work and suggestions for further investigations. Our source codes are available in GitHub (i) BERT-bases LLMs and GPT-2, (ii) Hallucination-less GPT-2

2 Related work

In this section, we have studied the state-of-the-art existing sexism detection and hallucination-reducing methodologies.

2.1 Sexism detection

As a challenging natural language processing task, sexism detection has attracted a lot of attention with the development of social networks for the past few years. Existing research focuses on different aspects when detecting sexism, including:
– Ambiguity and Context Dependency: sexism elements often rely on implicit cues, humour, sarcasm etc, where both BERT [9] and GPT [5] are capable of making predictions with bidirectional and unidirectional understanding of contexts.
– Cultural and Linguistic Variation: low-resources in other languages than Latin languages, Chinese for example, a zero-shot cross-lingual method [10] has been proposed.
– Data Imbalance and Annotation Challenges: data are often imbalanced and it is sometimes very subjective when annotating data collected from social media platforms, focal loss and other data augmentation strategies have been applied in this research [11] to improve online sexism detection.
– Lack of Interpretability: Complex deep learning models, especially transformer-based ones, result in difficulties for people to comprehend what is happening inside the black box. [1]

2.2 Hallucination

The issue of **Hallucination** of Large Language Models (LLMs), normally refers to the generation of plausible yet incorrect factual information [12]; Including but not limited to providing false citations [13] and invent scientific study that does not exist [14], as the example shown in the figure 1, where the generated content was not desired.

Hallucinations can be categorised into two groups: intrinsic hallucinations and extrinsic hallucinations. [15]. Intrinsic hallucination, which indicates the generated contents are not in line with the ground truth while extrinsic hallucination indicates what LLMs are generating seems coherent but cannot be traced anywhere. Recently, many new models have been proposed for evaluating LLMs' hallucinations, **HaluEval** [16], a benchmark which can be employed to distinguish which types of hallucinated contents tend to be generated, the ability of LLMs to be aware of the hallucinations, also, it can be further paired with human annotations for verification.

2.3 Hallucination Reduction Methods

This section briefly introduces several technical methods proposed aiming to detect, evaluate and reduce Large Language Models' hallucination issues, models are listed in this section organized in chronological order, covering various technologies such as contrastive learning, contrastive parameter assembling and bringing in synthetic tasks etc.

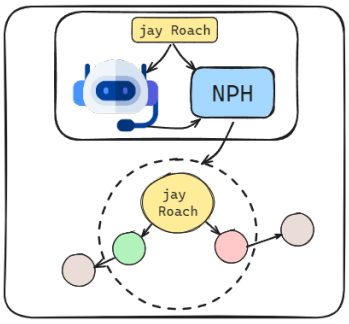

Fig. 2: Simplified NPH system

- **Neural Path Hunter (NPH)** [17] Highlighting the faithfulness of answers when designing dialogue systems, NPH, a refinement strategy consists of two models: A token-level hallucination critic and an entity mention retriever, which is capable of reducing the hallucination of answers generated by all LLMs, especially those applied in *Knowledge Graph-grounded* dialogue systems trained on Wikidata.
- **Mixed Contrastive Learning (MixCL)** [18] A framework based on the principle of Contrastive Learning, which indicates that the distance of similar samples should be drawn as close as possible, MixCL consists of two main steps: negative sampling and mixed contrastive learning aim to reduce the probability of generating negative tokens, to make LLMs less hallucinated.
- **Reducing Hallucination in Opendomain dialogue systems RHO (ρ)** [19] A model utilizing the representations of linked entities and relation predicates from a knowledge graph (KG), using two types of knowledge grounding: both locally and globally, to generate more accurate responses.
- **Contrastive Parameter Ensembling (CaPE)** [20] A method availing of training data more efficiently, reducing hallucination by variating the noise in training examples. Training **expert** and **anti-expert** models additionally on clean and noisy subsets of the entire dataset.
- **Chain-of-VErification (CoVE)** [12] A hallucination-reducing method has been proposed to improve language model performances by verifying the initial response with generated topic-related verification question-answer pairs before giving the final response.

– **Synthetic Transfer (SynTra)** [21] As a way of "teaching" LLMs to hallucinate less, SynTra first designs synthetic tasks where hallucination can be easily evaluated, and then it optimizes the system message, rather than the model weights. And transfer the system message to the realistic task at the final step.

2.4 Hallucination-less GPT-2

In this paper, we have selected the first hallucination-reducing method: **Neural Path Hunter (NPH)** [17], the structure of the model is shown in figure 2, because of the feasibility of adapting the trained model to the online sexism detection and classification task after numerous attempts.

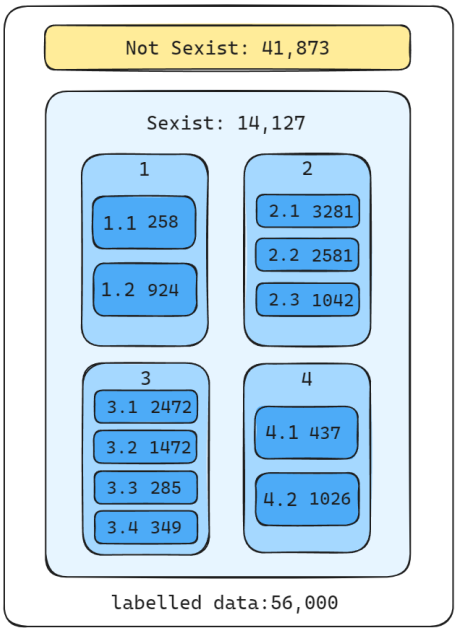

Fig. 3: Distribution of the labelled data.
The entire dataset contains two million unlabeled and 56,000 well-annotated data samples collected from two popular social media platforms: Gab and Reddit. Dataset consists of 4 main categories: (1) threats (2) derogation, (3) Animosity or (4) Prejudice. Each category is classified into more refined categories: (1.1) threats of harm, (1.2) incitement and encouragement of harm, (2.1) descriptive attacks, (2.2) aggressive and emotive attacks (2.3) dehumanizing attacks & overt sexual objectification (3.1) casual use of gendered slurs, profanities, and insults, (3.2) immutable gender differences and gender stereotypes, (3.3) backhanded gendered compliments, (3.4) condescending explanations or unwelcome advice (4.1) supporting mistreatment of individual women, (4.2) supporting systemic discrimination against women as a group.

3 Methodology and Experiments

This section explains more details about the dataset considered in this study and the methodology we adopted for sexist detection.

3.1 Dataset Description

To address sexism detection using LLMs, we used the taxonomy and dataset introduced by SemEval-2023 Task 10 on the Explainable Detection of Online Sexism (EDOS) [1]. The EDOS task collected datasets from two different social media platforms: Gab and Reddit. It consists of attributes such as "label_sexist" and "label_category", in which "label_vector" corresponds to the three subtasks, respectively and this guides them to build their three-level taxonomy; i) Binary sexism, ii) Category of sexism and iii) Fine-grained vectors. This hierarchical taxonomy helps to make the sexism model more explainable; if the data sample is considered as "sexist" in the first sub-task, it will be passed into the 4-category check to see which category it belongs to (1) threats (2) derogation, (3) Animosity or (4) Prejudice. The distribution of the finer-level categories is mentioned in figure 3. We have used all labelled data samples for task A, 80,000 in total. However, only 25% of the data was being used for the sub-tasks B and C since the *'not sexist'* was being removed after the detection task.

3.2 Preprocessing

During the data preprocessing phase, we first split the dataset into sub-datasets for each sub-task, then utilized a series of specialized functions to meticulously clean and normalize the textual data. These functions included:

1. **Text Cleaning**: This function applies regular expressions to remove usernames, URLs, non-alphanumeric characters (excluding basic punctuation), and multiple spaces, ensuring a cleaner text. It also eliminates specific unwanted characters and symbols such as degrees, euros, accented letters, and miscellaneous symbols (e.g., °, €, è, é), providing a more standardized textual dataset.
2. **Noise Removal**: By removing content enclosed within square brackets, this step further purifies the text from any residual noisy data or metadata annotations that could skew the analysis.
3. **Case Normalization**: Converting all text to lowercase ensures uniformity across the dataset, eliminating case sensitivity from impacting the data processing and analysis stages.

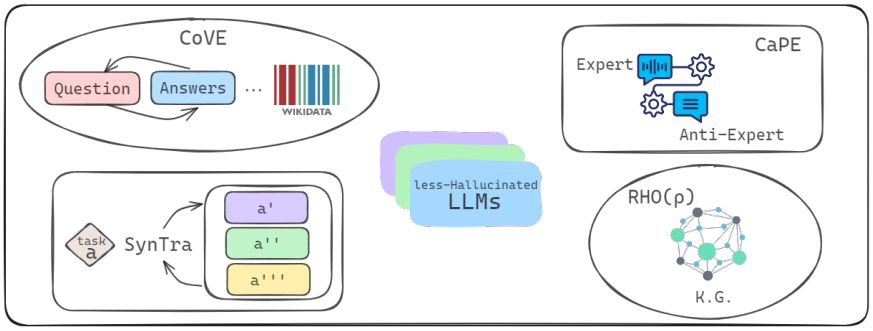

Fig. 4: Hallucination reducing methods examples.

CoVE: short for 'Chain of VErifications', a hallucination-reducing framework that compares the responses of generated questions and the baseline responses before providing the final answer. SynTra: stands for 'Synthetic Transfer', a novel model that aims to reduce LLMs' hallucination by generating synthetic tasks and shows cooperation with synthetic data can help mitigate undesired behaviors. While CaPE (Contrastive Parameter Assembling), training an 'expert' and 'anti-expert' model on a clean dataset and a noisy one. Reducing Hallucination in Open-domain dialogue systems (RHO), make the LLMs hallucinate less with the help of knowledge graphs.

3.3 Teach LLMs to hallucinate less

Algorithm 1: Reduced-Hallucination GPT-2 via NPH Integration

Input: dataset_csv: Path to OpenDialKG CSV dataset.
dataset_json: Path for JSON-converted dataset.
model_path: Pre-trained GPT-2 model identifier.
kge_path: Knowledge graph embeddings directory.
graph_path: Knowledge graph data directory.
output_dir: Directory for the fine-tuned model.
Output: Fine-tuned GPT-2 model with reduced hallucinations.

1 Convert Dataset from CSV to JSON format;
2 Initialize GPT-2 model;
3 Integrate NPH configurations using kge_path and graph_path for contextual grounding;
4 Prepare training and evaluation data from dataset_json;
5 Fine-tune GPT-2 model with NPH enhancements;
6 Rebuild the less-hallucinated GPT-2 model for Online Sexism detection and classification task;

As shown in figure 4, different hallucination-reducing methods can be applied to LLMs, to reduce hallucination to have better performances even if the models were proposed and trained grounded on dialogue-response systems.

In our study, we utilized the Neural Path Hunter (NPH) methodology to mitigate hallucinatory outputs in dialogue systems, refining the GPT-2 model's performance on the OpenDialog dataset for enhanced effectiveness in online sexism detection and classification.

We conducted a comparative analysis of the GPT-2 model's performance on this task before and after implementing NPH to explore the potential utility of hallucination reduction techniques beyond dialogue systems. By examining the model's consistency across different conditions, we aimed to illuminate the broader implications of employing such technologies. We selected NPH for its robust framework aligning with our objectives in sexism detection and classification. Figure 2 provides an overview of the NPH model's architecture. Following Algorithm 1, we adapted the GPT-2 framework using post-NPH training configuration files, crucial for fine-tuning the model to generate outputs with reduced hallucinatory content.

3.4 Experimental Results

Our experiments were conducted on "imtserver02", which was located in our laboratory, a server equipped with an Intel(R) Xeon(R) W-2245 CPU at 3.90GHz, operating on Ubuntu 22.04.4 LTS with a Linux 6.5.0-26-generic kernel. The system features 16 CPU cores, 16.5 MiB of L3 cache, and advanced VT-x virtualization, ensuring high performance and efficient parallel processing.

LMs	**Task A**	**Task B**	**Task C**
Roberta	**0.892**	**0.651**	0.522
Roberta_large	0.837	0.483	**0.524**
Electra	0.882	0.647	0.501
GPT-2	0.850	0.562	0.434
Less-hallucinated GPT-2 with NPH	0.841 (↓ **0.009**)	0.549 (↓ **0.013**)	0.430 (↓ **0.004**)

Table 1: Accuracy of the LMs for i) Binary sexism (Task A), ii) Category of sexism (Task B) and iii) Fine-grained vectors (Task C)

Note: the down arrows indicate a decrease in accuracy when comparing the Less-hallucinated GPT-2 to the standard GPT-2.

We have first re-applied the BERT-based models such as Roberta, Roberta-large [6] and Electra [7] for the first sub-task (Task A). These models are proven to have the best performances according to the final results [1] published at the SemEval competition. We used these BERT-based models to have a better understanding of the sexism multi-classification task and provide comparative results as baseline models. Our experimental setup involved the preprocessing3.2 of text data to fit the models' input requirements, followed by the application of a customized focal loss function during training to address class imbalance.

The models had been trained using a batch size of 6 and employed the one-cycle policy for learning rate adjustment to optimize performance. Validation was conducted on a separate test dataset to assess the models' accuracy and generalization capability, ensuring a comprehensive evaluation of their effectiveness in the targeted multi-classification task.

As the results show in table 1, BERT-based models are proven to have stable and excellent performances for the binary detection task. As illustrated in figure 5, both Roberta and Electra are proved to have good performances while Roberta-large has the best performance for the third sub-task (Task C), **Fine-grained Category** of Sexism detection.

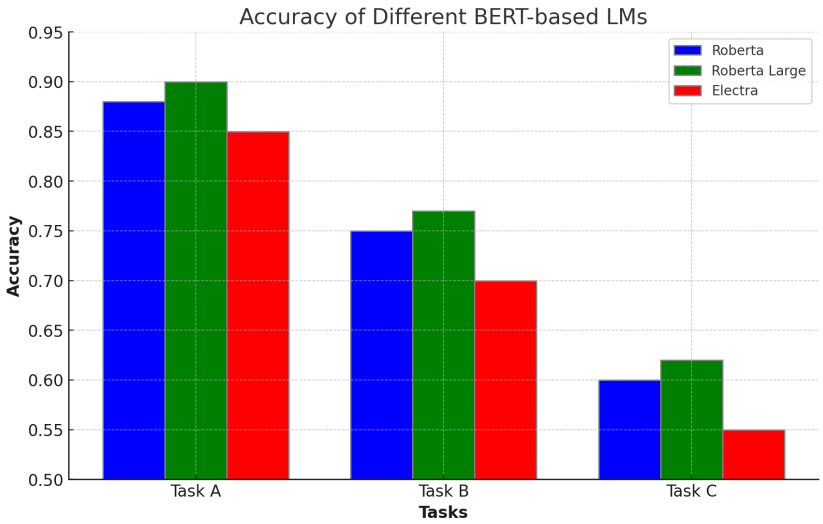

Fig. 5: BERT-based LMs trained on tasks A, B and C

To see the impacts of different hallucination reduction techniques, we first apply the initial version of the GPT-2 [22] model for all online sexism detection and classification tasks. Then, the GPT-2 models were applied to one hallucination-reducing method to get hallucination-less GPT models, and after that re-applied for the same tasks to observe if there was any difference in the performances.

As demonstrated in Table 1 and figure 6 the accuracy of the GPT-2 Model for Task A, Task B and Task C is 85.0%, 56.2% and 43.4%, respectively. In each experiment, the batch size was 32 and the number of epochs was 5 for Task A and 25 for both Task B and Task C. As depicted in the last two rows of the Table 1, the results obtained when re-building the GPT-2 model inside the Neural Path Hunter (NPH) [17] on a dataset of conversations: **OpenDialKG** [23]. The less-hallucinated GPT-2 model obtained 84.1% accuracy for the binary detection task, with a reduction of performances of 0.9% compared to GPT-2. Same for

the second sub-task and third sub-task, GPT-2 has higher performances than the less-hallucinated GPT-2, with a difference of 1.3% and 0.4%, respectively. The lack of improvement in accuracy can be attributed to the model being extensively overfitted, as it was trained solely on the OpenDialKG dataset. To address this issue, implementing hallucination reduction techniques with training on varied datasets, particularly those focused on sexism or hate speech detection and classification, Hate Speech and Offensive Language(HSOL)[24] for instance, might be beneficial. Also, we can notice that the accuracy for three sequential tasks has slightly changed before and after applying the hallucination reduction method chosen.

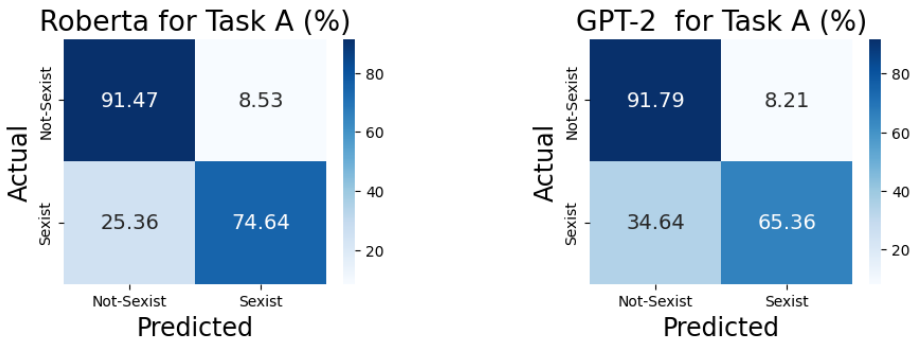

Fig. 6: Confusion matrix of the Roberta and GPT-2 model for the binary detection of Sexism

As shown in figure 6, we generated a confusion matrix for the best-performed BERT-based model and GPT-2 to understand whether the misclassification rates can be reduced or similar in GPT-2 models. We observed that the false positive rate of the GPT-2 model (8.21%) and not-sexist content classification for the GPT-2 model (91.79%) are very similar to that of Roberta. However, we observed that higher false negative rate in the GPT-2 model compared to Roberta. We aim to reduce the false negative rates using a large sexism dataset and more regularization techniques in the future. In addition, we have noticed that the time consumed for training the GPT-2 model when applying the online sexism classification on the dataset provided in EDOS task [1] is much less than the BERT-based models in the same execution environment.

3.5 Discussion

In this section, we interpret the results we have obtained in the context of existing literature and mainly focus on exploring potential avenues for future research. We begin by examining new hallucination-reducing methods specifically designed for and applicable to *dialogue-response systems*.

We then apply several of these methods on the first sub-task of SemEval-2023 Task 10, which involves detecting sexism in social media data collected from two popular platforms: Gab and Reddit. Our findings demonstrate that applying hallucination-reducing methods on GPT-based LLMs, such as GPT-2, is shown not to be an effective strategy for improving their performance in classification tasks. However, several limitations in our study need further investigation. First, the number of hallucination-reducing methods applied to "Teach" LLMs to hallucinate less is relatively small, future studies with more diverse technologies could provide more comprehensive and detailed impacts than we have observed. Additionally, our research primarily focused on online sexism detection and classification tasks using datasets provided by the organization [1]. It would be valuable for future research to explore other available datasets, such as the 'Call me sexist but' Dataset (CMSB) [25], and consider different multilingual datasets that are publicly accessible.

In summary, while our study highlights the impact of the hallucination-reducing methods for LLMs beyond dialogue-response scenarios, particularly in classification tasks, further research is needed to delve deeper into this complex and diverse area. Addressing the limitations of our work and conducting a more extensive exploration of the suggested avenues will significantly contribute to the optimizing of LLMs by leveraging hallucination-reduction methods across a broad range of natural language processing scenarios.

4 Conclusion

In this study, we explored various novel frameworks for reducing hallucinations in the context of online sexism detection and classification, extending beyond the realm of dialogue-response systems. These tasks presented significant challenges for several reasons: (i) The hallucination-reducing methods are not originally designed or trained for the classification tasks, making the implementation of code both challenging and time-consuming. (ii) Online sexism detection and classification pose inherent complexities due to the sheer volume of posts on popular social platforms and the wide diversity of sexist content. (iii) While the majority of participants opted for BERT-based language models, known for their superior performance in binary and multi-classification tasks, our research has shown that the application of hallucination-reduction methods can enhance the performance of GPT-based language models. This suggests the potential for achieving better results in various other natural language processing scenarios.

These findings underscore the importance of further research into the adaptation of hallucination-reduction techniques for diverse NLP tasks, to enhance model performance and applicability across a broader range of applications.

References

1. H. R. Kirk, W. Yin, B. Vidgen, and P. Röttger, "SemEval-2023 Task 10: Explainable Detection of Online Sexism," in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics, 2023. [Online]. Available: http://arxiv.org/abs/2303.04222
2. V. Rawte, A. Sheth, and A. Das, "A survey of hallucination in large foundation models," 2023. [Online]. Available: https://arxiv.org/pdf/2309.05922.pdf
3. J. Cui, Z. Li, Y. Yan, B. Chen, and L. Yuan, "Chatlaw: Open-source legal large language model with integrated external knowledge bases," 2023. [Online]. Available: https://arxiv.org/pdf/2306.16092.pdf
4. A. Azaria and T. Mitchell, "The internal state of an llm knows when it's lying," 2023. [Online]. Available: https://arxiv.org/pdf/2304.13734.pdf
5. A. Radford and K. Narasimhan, "Improving language understanding by generative pre-training," 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:49313245
6. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019. [Online]. Available: https://arxiv.org/pdf/1907.11692.pdf
7. K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators," 2020. [Online]. Available: https://arxiv.org/pdf/2003.10555.pdf
8. J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805
9. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019. [Online]. Available: https://arxiv.org/pdf/1810.04805.pdf
10. G. Li, P. Rajapaksha, R. Farahbakhsh, and N. Crespi, "Emotionally-bridged cross-lingual meta-learning for chinese sexism detection," in *The 12th CCF International Conference on Natural Language Processing and Chinese Computing (NLPCC)*, Foshan, China, Oct. 2023. [Online]. Available: https://hal.science/hal-04168449
11. S. Rallabandi, S. Singhal, and P. Seth, "Sss at semeval-2023 task 10: Explainable detection of online sexism using majority voted fine-tuned transformers," 2023. [Online]. Available: https://arxiv.org/pdf/2304.03518.pdf
12. S. Dhuliawala, M. Komeili, J. Xu, R. Raileanu, X. Li, A. Celikyilmaz, and J. Weston, "Chain-of-verification reduces hallucination in large language models," 2023. [Online]. Available: https://arxiv.org/pdf/2309.11495.pdf
13. N. F. Liu, T. Zhang, and P. Liang, "Evaluating verifiability in generative search engines," 2023. [Online]. Available: https://arxiv.org/pdf/2304.09848.pdf
14. H. Thorp, "Chatgpt is fun, but not an author," 2023. [Online]. Available: https://www.science.org/doi/10.1126/science.adg7879
15. Neeraj Varshney. (2023) The hallucination problem of large language models. https://medium.com/mlearning-ai/the-hallucination-problem-of-large-language-models-5d7ab1b0f37f.
16. J. Li, X. Cheng, W. X. Zhao, J.-Y. Nie, and J.-R. Wen, "Halueval: A large-scale hallucination evaluation benchmark for large language models," 2023. [Online]. Available: https://arxiv.org/pdf/2305.11747.pdf

17. N. Dziri, A. Madotto, O. Zaïane, and A. J. Bose, "Neural path hunter: Reducing hallucination in dialogue systems via path grounding," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 2197–2214. [Online]. Available: https://aclanthology.org/2021.emnlp-main.168
18. W. Sun, Z. Shi, S. Gao, P. Ren, M. de Rijke, and Z. Ren, "Contrastive learning reduces hallucination in conversations," 2022. [Online]. Available: https://arxiv.org/pdf/2212.10400.pdf
19. Z. Ji, Z. Liu, N. Lee, T. Yu, B. Wilie, M. Zeng, and P. Fung, "Rho (ρ): Reducing hallucination in open-domain dialogues with knowledge grounding," 2023. [Online]. Available: https://arxiv.org/pdf/2212.01588.pdf
20. P. K. Choubey, A. Fabbri, J. Vig, C.-S. Wu, W. Liu, and N. Rajani, "CaPE: Contrastive parameter ensembling for reducing hallucination in abstractive summarization," in *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 10 755–10 773. [Online]. Available: https://aclanthology.org/2023.findings-acl.685
21. E. Jones, H. Palangi, C. Simões, V. Chandrasekaran, S. Mukherjee, A. Mitra, A. Awadallah, and E. Kamar, "Teaching language models to hallucinate less with synthetic tasks," 2023. [Online]. Available: https://arxiv.org/pdf/2310.06827.pdf
22. A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019. [Online]. Available: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
23. S. Moon, P. Shah, A. Kumar, and R. Subba, "Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, July 2019. [Online]. Available: https://github.com/facebookresearch/opendialkg
24. T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ser. ICWSM '17, 2017, pp. 512–515. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14955/14805
25. M. Samory, "The 'call me sexist but' dataset (cmsb)," GESIS - Leibniz-Institute for the Social Sciences. Data File Version 1.0.0, https://doi.org/10.7802/2251, 2021. [Online]. Available: https://search.gesis.org/research_data/SDN-10.7802-2251?doi=10.7802/2251

A Appendix: Commands for Fine-Tuning GPT-2 Model

A.1 Prepare Dataset: convert CSV to JSON

```
python convert_opendialkg.py --input_file
↪ /home/dice/leyuan/nph/nph/data/opendialkg.csv --out_file
↪ /home/dice/leyuan/nph/nph/data/opendialkg.json
```

A.2 Train Dialogue Generation Model

```
python -m dialkg.mask_refine --model_name_or_path gpt2
↪ --train_dataset_path /home/dice/leyuan/nph/nph/data/opendialkg.json
↪ --eval_dataset_path /home/dice/leyuan/nph/nph/data/opendialkg.json
↪ --do_train --output_dir /home/dice/leyuan/nph/nph/trained_ly
↪ --num_train_epochs 3 --train_batch_size 16 --eval_batch_size 16 --kge
↪ /home/dice/leyuan/nph/nph/graph_embeddings --graph_file
↪ /home/dice/leyuan/nph/nph/opendialog --pad_to_multiple_of 8
↪ --max_adjacents 100 --patience 10 --max_history 3
```