Detecting AI-enhanced Opinion Spambots: a study on LLM-generated Hotel Reviews

Vijini Liyanage*, **Davide Buscaldi***, **Pénelope Forcioli**†
*University Sorbonne Paris Nord
93430 Villetaneuse, France
{liyanage, buscaldi}@lipn.univ-paris13.fr
†Ecole Polytechnique
91120 Palaiseau, France
penelope.forcioli@polytechnique.edu

Abstract

Opinion spamming is the posting of fake opinions or reviews to promote or discredit target products, services, or individuals. The concern surrounding this activity has grown steadily especially because of the development of automated bots for this purpose ("spambots"). Nowadays, Large Language Models (LLMs) have proved their ability to generate text that is almost indistinguishable from human-written text. Therefore, there is a growing concern regarding the use of these models for malicious purposes, among them opinion spamming. In this paper, we carry out a study on LLM-generated reviews, in particular hotel reviews as we chose the well-known Opinion Spam corpus by Myle Ott as the seed for our dataset. We generated a set of fake reviews with various models and applied different classification algorithms to verify how difficult is it to detect this kind of generated content. The results show that by providing enough training data, it is not difficult to detect the fake reviews generated by such models, as they tend to associate the aspects in the reviews with the same attributes.

Keywords: artificially generated text, detection, classification, opinion spam

1. Introduction

Opinion spamming encompasses the dissemination of counterfeit reviews or opinions for the purpose of promoting or discrediting products, services, or individuals (Liu, 2012). Fake reviews can be generated either by humans or by using text-generating algorithms to automate the process. Historically, the detection of automated spamming has been relatively straightforward, largely due to the mechanical and less expressive nature of machine-generated text compared to human-authored content. Nonetheless, the advent of Large Language Models allowed for a paradigm shift, as these models have demonstrated a remarkable capability to produce text that closely mimics human writing. Consequently, there is a growing unease surrounding the potential misuse of these advanced models, one of which involves their deployment in opinion spamming.

Numerous studies have leveraged Natural Language Processing (NLP) techniques to detect fake reviews. Researchers have explored sentiment analysis, textual patterns, and linguistic features to distinguish between genuine and artificially generated content. (Martinez-Torres and Toral, 2019) successfully employed machine learning methods for sentiment analysis to identify deceptive hotel reviews. (Elmogy et al., 2021) utilized supervised machine learning classifiers, including Random Forest and Support Vector Machines (SVM), to classify fake hotel reviews. They demonstrated the effectiveness of these algorithms in achieving high precision and recall rates.

Recent advancements in synthetic text generation models, in particular the Generative Pretrained Transformer (GPT) family of language models (Radford et al., 2019), have introduced new challenges in fake review detection. Researchers have begun to adapt their detection methods to identify reviews generated by these sophisticated models. One of the first works to address this problem is the one by (Salminen et al., 2022). In their work, they found out that human accuracy in detecting fake reviews is only slightly higher than random chance, and that when applying text-based fake-review detection, the more words a review has, the higher the chance of detecting its true label (fake or real).

In this paper, we tackle this problem from an NLP perspective to understand what are the linguistic features that allow text-based classification models to distinguish between generated and original text. We explain our approach in building corpora composed of artificially generated hotel reviews leveraging smaller Large Language Models (LLMs), such as GPT-2, GPT-3, and TinyLLama. In our opinion, this would match the choice of malicious spammers, as these models do not require demanding hardware and produce tokens at a fast rate. After some preliminary tests, we discarded ChatGPT because on one hand it refused to follow the instruction and on the other one, when it did, it produced the same review over and over. Besides, we evaluate the detectability of the generated contents by employing

The Seventh Workshop on e-Commerce and NLP (ECNLP 7), pages 74–78
21 May, 2024.

statistical as well as deep learning-based classification models.

2. Datasets

As a set of original “seed" reviews to reproduce with the LLMs we used 400 truthful positive reviews from TripAdvisor and 400 truthful negative reviews from Expedia, Hotels.com, Orbitz, Priceline, TripAdvisor and Yelp which compose the “truthful" subset of the well-known Myle Ott’s Opinion Spam corpus (Ott et al., 2011). These reviews were written for 20 Chicago hotels: each hotel has 20 positive and 20 negative reviews.

Initially, we used the GPT-3 DaVinci model to generate for each of the original reviews a faked equivalent one. We fed to the model the prompt “Generate a {%polarity} review completing the following text:", where {%polarity} is replaced either by “negative" or “positive", and the sequence composed by the first 10 words of the genuine review and asked to complete and generate text up to the same length of the original one. We used temperature 0.9 to ensure a certain creativity by the model. Similarly, we created a set of reviews using the TinyLlama model (Zhang et al., 2024), one of the latest and smallest available LLMs.

Subsequently, we created another set with the objective of creating reviews that were more similar to the human-written ones. A way to do this is to fine-tune the language model with the set of genuine reviews. In this way, the model will tend to generate text that fits better the probability distribution of the tokens in the genuine set. Since we didn’t have the resources to fine-tune GPT-3 we switched to GPT-2 with 355M parameters. Again we fed a sequence of 10 words as seed text, including the hotel name.

Finally, following the counter-detection strategy applied by (Sadasivan et al., 2023), we built a final set by using paraphrasing. The paraphrasing model chosen was a T5 model fine-tuned on chatGPT-generated paraphrases available on Huggingface [1]. For each original review, we generate 5 different paraphrases and we randomly pick one of them. With this strategy, we paraphrased the reviews from the GPT-3 model and obtained another set of 800 fake reviews.

Below are examples of negative reviews referring to the same example, the original one and the generated ones:

- **Original review:** This hotel must have originally been an ordinary downtown hotel in a not very nice neighborhood. The owners have painted it over and fixed up the lobby to make it an "upscale" overpriced boutique hotel. But it’s still just an ordinary hotel with a paint job. Stay here if you want to experience staying in a flophouse, but with all the sleaze removed. Definitely stay at the chain’s other location, the Burnham, for a much, much nicer hotel.

- **Fine-tuned GPT-2 generated review:** We stayed in hotel Allegro and it was bad. Although the rooms are decent, they are not called “The Best in a Hilliard" and are not suited with insole. The bathroom was very small and has anearth and mineral salts. The pipes and sinks were old and the shower and tub are in need of a major overhaul. The hotel charges for internet and/or rental for 5 internet sessions a day. If you sign up for wireless internet, they do not fund wireless internet so slow and unfavorable wireless speeds.

- **Pre-trained GPT-3 generated review:** This hotel must have originally been an ordinary downtown hotel but now it is hardly more than a shabby motel. The walls are dirty and the furniture is worn, with some broken pieces. My room smelled musty and the carpets had stains. The bed was uncomfortable and the linen was old and threadbare. The staff were not particularly helpful and I felt like I was an inconvenience to them.

- **TinyLlama generated review:** This hotel must have originally been an ordinary downtown hotel. The rooms are a little too tiny and they lack basic amenities, including a pool. However, the location is very convenient as it is just a few minutes walk from Central Park.

- **Paraphrased GPT-3 review:** I’m sure the hotel was once a regular downtown hangout, but it’s now resembles ruins. The walls are unclean, the furniture is old and scratchy, there are stains in every room, an uncomfortable bed, and lifeless linens.

The datasets are available at the following repository: https://github.com/dbuscaldi/op_spam_AI.

In Table 1 we show the average perplexities for each of the datasets. Perplexity is defined as:

$$PPL(X) = \exp\{-\frac{1}{t}\sum_{i}^{t}\log p_\theta(x_i|x_{<i})\}$$

where $X = (x_0, \dots, x_t)$ is the input sentence and p_θ is the estimated probability by a LLM (in our case, we chose the GPT-2 probabilities, as in the GLTR paper (Gehrmann et al., 2019)). Perplexity is used

[1] https://huggingface.co/humarin/chatgpt_paraphraser_on_T5_base

by some commercial detectors, notably GPTZero[2], as a feature to detect whether a text is generated or not. The principle of perplexity is that it measures the “randomness" of a text sequence. For instance, a perplexity of 3 means that after every word in the text, there are on average 3 choices to continue the sentence. The perplexity values obtained on our datasets show that GPT-3 pre-trained has the lowest variability among all generated ones. On the other hand, we can see that the TinyLlama and the paraphrase one have higher perplexity, confirming the hypothesis that paraphrasing and changing the generating model may help confuse detectors that are based on a specific model.

3. Experiments and Results

To evaluate the detectability of the fake reviews, we employed a variety of classification models. First of all, we started with a basic Multinomial Naïve Bayes with tf.idf weights, without lemmatization and stopwords removal. We evaluated the model on a random split of 80:20 for training and testing. The results with this model were already very good, obtaining a F-1 score of 96%, indicating that the task could be solved just by looking at the vocabulary. Therefore, we compared the log-probabilities of the words in the generated and non-generated class, calculating their difference. We show in Table 2 the 10 most discriminating words for both classes.

It can be observed how the most discriminating words for the generated category tend to be attributes (“unhelpful", “terrible", “delicious", “outdated", ...) while the ones for the non-generated category seem more related to objects or places (“door", “floor", “coffee", “michigan", “ave", ...) and personal pronouns (“she", “he", “your", ...). Similar results are obtained with TinyLlama, with some variations on the attributes (“stunning" is more prevalent among generated texts). We tried similar experiments with bi-grams and tri-grams as features instead of words and this difference in style is even more clear: in the most important trigrams for the generated class we find tri-grams that match a pattern “X was/were Y" where X is usually a service or an aspect of the hotel and Y an adjective. In the non-generated most representative tri-grams we find tri-grams such as “in the room", “in the bathroom", “the first night". The difference in style was expected as other works about generated text detection (Antoun et al., 2023) have noticed the tendency of LLMs to produce recurrent patterns in the output.

To verify the importance of vocabulary overlap for detection, we carried out an experiment in which we varied the proportion of training and test data. Note that in a realistic scenario training data would not be balanced, as annotated corpora have sizes that are only a small fraction of the total number of reviews on platforms. We carried out 10 experiments for each proportion of test and training data. The results, separating recall and precision for each class, are presented in Figure 1.

As can be seen, the precision for non-generated texts tends to be lower than for generated texts, while the recall shows the inverse. This indicates the presence of many false negative examples, i.e. the model is prone to classify machine-generated text as human-written, especially when training data are few. This phenomenon has also been observed by (Wang et al., 2023) in experiments on cross-domain classification. Scores for TinyLlama are lower than for GPT-3, indicating that these generated texts are more difficult to detect.

If we take into account the paraphrased corpus, both precision and recall are high, but the accuracy of the detection is more sensible to the availability of training data, as can be seen in Figure 2. Therefore paraphrasing makes it slightly more difficult to detect fake reviews when there is not enough training data.

Finally, we made some experiment with state-of-the-art classification algorithms based on transformers models, specifically BERT$_{base}$, SciBERT, XLNet$_{large}$, and ELECTRA$_{small}$. These models were imported as pre-trained models from Hugging Face (Wolf et al., 2020) and fine-tuned using Simple Transformers [3]. We employed the BERT tokenizer across all models. The fine-tuning process included 10 epochs, a batch size of 16, and a maximum sequence length of 128. For standalone models, we used unprocessed text as input.

The dataset was split into an 80:20 ratio for training and testing. Each model underwent three experimental iterations, and the average F1 scores resulting from these experiments are provided in Table 3. Among all detection models, BERT$_{base}$ seems to be the most effective in detecting the generated content. GPT-2 reviews are the least predictable, given the fine-tuning process that made them more similar to the human-written ones. The pre-trained models seem rather easy to detect with any of the Transformer-based models.

4. Conclusions

In this work, we created various collections of automatically generated reviews for automated detection, based on the Opinion Spam corpus by (Ott et al., 2011). The results show that the vocabulary and style of generated reviews is very different from the one used in the authentic ones, making it relatively easy to detect the fake ones, provided

[2] https://gptzero.me

[3] https://simpletransformers.ai

original	GPT-2 fine-tuned	GPT-3 pretrained	paraphrased	TinyLlama
60.433	26.401	20.102	32.828	43.820

Table 1: Average perplexity for each collection.

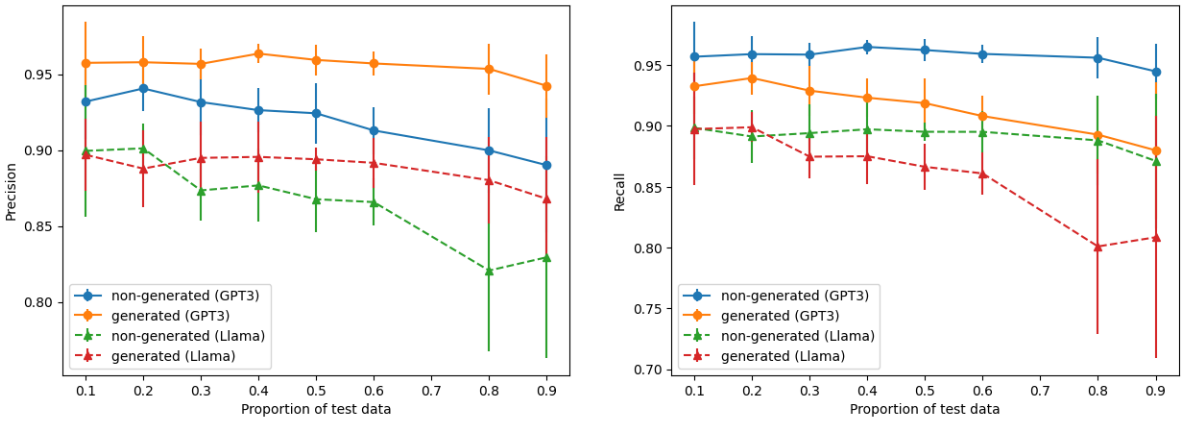

Figure 1: Precision and recall for each class on the GPT3 dataset vs. original reviews, and TinyLlama vs. original reviews, varying the proportion of test and training data. The error bar indicates the standard deviation calculated over 10 experiments.

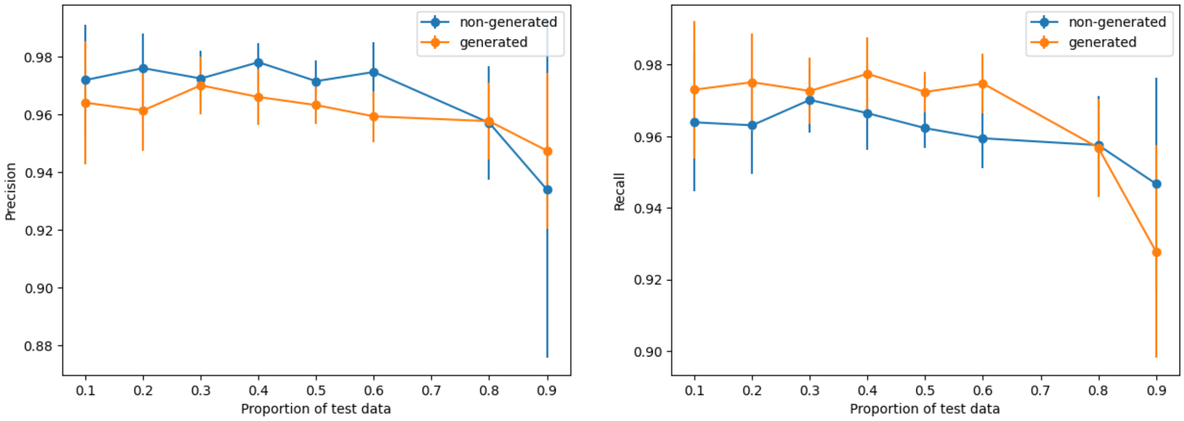

Figure 2: Precision and recall for each class on the paraphrased dataset vs. the original reviews, varying the proportion of test and training data.

Generated		Authentic	
Word	**delta**	**Word**	**delta**
unhelpful	2.889	door	-1.694
incredibly	2.620	floor	-1.680
delicious	2.609	coffee	-1.656
outdated	2.402	next	-1.636
terrible	2.306	your	-1.557
accommodating	2.257	concierge	-1.513
anyone	2.229	she	-1.478
uncomfortable	2.129	ave	-1.468
amenities	1.907	mile	-1.420
musty	1.873	etc	-1.402

Table 2: The 10 most discriminating words for each category (GPT-3 dataset) sorted by their log-probability difference (delta).

Model	GPT-2	GPT-3	para	Llama
BERT$_{base}$	97.83	99.38	98.29	99.74
SciBERT	93.66	93.75	97.62	99.35
XLNet$_{large}$	87.87	92.70	95.32	98.57
ELECTRA	92.49	93.49	95.37	99.34

Table 3: F1 Scores obtained by Transformer-based Classification Models. “para" indicates the paraphrased corpus.

that a large and varied enough training data set is available.

These results are partially comforting as it looks like it is not possible to use LLMs to automatically produce undetectable fake reviews without the intervention of a human, lowering the harm potential of these models. We didn't test how often neural hallucinations occur in these reviews, but after inspection we could observe that some reviews mention features that are not present in the targeted hotel. For future works, we plan to extend our tests with models with higher temperatures and to measure the hallucination phenomenon.

Acknowledgements

This work is supported/ partially supported by a public grant overseen by the IdEx Université Paris Cité (ANR-18-IDEX-0001) as part of the Labex Empirical Foundations of Linguistics - EFL

Bibliographical References

Wissam Antoun, Virginie Mouilleron, Benoît Sagot, and Djamé Seddah. 2023. Towards a robust detection of language model generated text: Is chatgpt that easy to detect?

Ahmed M Elmogy, Usman Tariq, Mohammed Ammar, and Atef Ibrahim. 2021. Fake reviews detection using supervised machine learning. *International Journal of Advanced Computer Science and Applications*, 12(1).

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.

Nitin Jindal and Bing Liu. 2008. Opinion spam and analysis. In *Proceedings of the 2008 international conference on web search and data mining*, pages 219–230.

Bing Liu. 2012. Opinion spam detection. In *Sentiment Analysis and Opinion Mining*, pages 113–125. Springer.

Maria del Rocío Martinez-Torres and Sergio L Toral. 2019. A machine learning approach for the identification of the deceptive reviews in the hospitality sector using unique attributes and sentiment orientation. *Tourism Management*, 75:393–403.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can ai-generated text be reliably detected?

Joni Salminen, Chandrashekhar Kandpal, Ahmed Mohamed Kamel, Soon-gyo Jung, and Bernard J Jansen. 2022. Creating and detecting fake reviews of online products. *Journal of Retailing and Consumer Services*, 64:102771.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Alham Fikri Aji, and Preslav Nakov. 2023. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model.