CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17820

Evaluating coherence in AI-generated text

María Olmedilla[1], José Carlos Romero[2], Rocío Martínez-Torres[3], Nicolas R. Galvan[4], Sergio Toral[4]

[1]SKEMA Business School, Université Côte d'Azur, France, [2]Applied Computational Social Sciences Data-Intensive Governance-Institute, Université Paris Dauphine-PSL, France, [3]Facultad de Ciencias Económicas y Empresariales, University of Seville, Spain, [4]E. T. S. Ingeniería, University of Seville, Spain.

How to cite: Olmedilla, M.; Romero, J. C.; Martínez-Torres, R.; Galvan, N.; Toral, S. 2024. Evaluating coherence in AI-generated text. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17820

Abstract

This study examines the role of coherence in AI-generated online reviews and its effect on perceived authenticity and consumer trust. By applying advanced metrics like BERT Score, BART Score, and Disco Score, the research analyzes the coherence of AI-generated text using Generative AI models, specifically Llama-2, on Amazon beauty product reviews. Results indicate that AI-generated reviews exhibit higher coherence compared to human-generated content, suggesting that Generative AI can produce seemingly authentic content. This finding challenges the ability to distinguish between human and AI-generated reviews, raising important questions about consumer trust in digital marketplaces. The study underscores the importance of coherence in online content's credibility and opens avenues for further research on Generative AI's role in e-commerce.

Keywords: *Generative-AI; Online reviews; Llama-2, BERT, Coherence.*

1. Introduction

In the digital era, the authenticity and integrity of online content have become paramount, especially with the proliferation of user-generated and AI-generated texts in digital marketplaces. The concept of coherence in online reviews, which is a significant component of texts, not only signifies authenticity but also impacts consumer trust and decision-making. This paper aims to delve into the intricate relationship between coherence and perceived authenticity in AI-generated text, highlighting the important role of advanced computational and linguistic tools in this evaluation process.

There has been some efforts among researchers to underscored the critical role of advanced computational tools and linguistic theories to explore the coherence in AI-generated text. For

instance, Ai et al. (2019) discuss the integration of text coherence in text generation, emphasizing the significance of coherence metrics like semantics and syntax-based coherence in enhancing text generation. Likewise, Elkhatat et al. (2023) investigate the distinction between human and AI-authored content. Their work reveals the capabilities and limitations of AI content detection tools, highlighting the ongoing challenge in accurately identifying AI-generated text.

Furthermore, the advent of sophisticated computational linguistic tools, such as BERT Score, BART Score, and Disco Score, has revolutionized our ability to assess coherence in AI-generated text, their effectiveness in maintaining the integrity of online reviews, and the implications for consumer trust in an era increasingly dominated by generative AI.

2. Research Background

2.1. Generative AI and online reviews

The exploration of Generative Artificial Intelligence (AI) in the context of online reviews represents a growing field that holds promise for reshaping e-commerce and user interaction on online platforms. Generative AI, leveraging models such as Transformers, has extended its utility beyond traditional applications like image processing and natural language processing to the generation of online reviews (Bandyopadhyay et al., 2023). The exploration of generative AI in online reviews seeks to understand its impact on credibility, authenticity, and usefulness (Hu et al., 2012). Besides, the effectiveness of generative AI in producing online reviews that are perceived as helpful and authentic by users is crucial for its application in e-commerce platforms, a critical factor for consumer trust and decision-making

Nevertheless, the integration of generative AI into the creation of online reviews is not without its challenges. Studies contrasting prior laboratory evidence have shown that the use of generative AI in text generation can result in a decline in the quality of online reviews. This effect is particularly pronounced among non-expert reviewers, though it also leads to an increase in the quantity of content produced per reviewer, highlighting a trade-off between quality and volume (Knight & Bart, 2023). In the case of the recommender systems, generative AI has been employed to enhance the informativeness and efficiency of user-generated reviews. By combining traditional collaborative filtering methods with advanced deep learning architectures for text processing, researchers have achieved superior performance in recommendation accuracy compared to baseline systems. This demonstrates generative AI's potential to significantly improve the personalization and relevance of online recommendations (Shalom et al., 2019). Another critical application of generative AI in online reviews is in spam detection. Innovative methods based on aspect-level analysis of reviews have shown promise in identifying and mitigating spam content, thereby protecting the integrity of online review

platforms (Wang et al., 2022). Generative AI's capability for content generation has also been explored in review generation tasks, in thir regard the research by Zang and Wan (2017) focused on long review generation within the encoder-decoder neural network framework. This line of work addressed the challenges of automating review generation to produce helpful and persuasive online content. Furthermore, online reviews are central in influencing consumer buying choices, although only a small number of users dedicate effort to crafting constructive reviews. Fortunately, the latest advances in deep neural networks presents a promising avenue for generating content that closely resembles authentic reviews (Kaghazgaran et al., 2020).

In summary, the intersection of generative AI with online reviews encompasses a diverse range of applications, from enhancing the personalization of recommendations to improving the integrity and usefulness of user-generated content. Despite the challenges related to content quality, the ongoing research and development in this area (Ooi et al., 2023) underscore generative AI's transformative potential in e-commerce and beyond.

2.2. Measuring the coherence between true and fake reviews

The foundational principle that coherence is indispensable for effective written discourse, as argued by Bamberg (1983), sets the basis for understanding its relevance in evaluating online reviews. Giora (1985) extends this by emphasizing the role of "aboutness" or the discourse topic in achieving text coherence, thus highlighting the intrinsic connection between coherence and the pragmatic formation of text. This theoretical framework can help assessing the credibility of online reviews, where coherence may indicate authenticity.

Furthermore, the methodology developed by Foltz et a.l (1998) for measuring textual coherence with latent semantic analysis introduces a quantitative approach to this qualitative attribute, suggesting that coherence can be systematically analyzed and assessed. Likewise, Cui et al. (2017) contribute to this domain by employing deep learning models to evaluate text coherence, illustrating the potential of advanced computational techniques in discerning well-organized from poorly structured texts, a distinction crucial for identifying fake reviews.

Additionally, the role of prior knowledge in text comprehension introduces a nuanced perspective on how individual differences in knowledge coherence could influence the perception and evaluation of online reviews (McCarthy & McNamara, 2021). This aspect is particularly relevant in the context of spam detection, such as the approach proposed by Yang (2015), which uses coherence metrics to distinguish between genuine and spam reviews.

In this regard, Singh et al. (2020) empirically demonstrate that fake news articles show lower textual coherence compared to legitimate counterparts, an insight that can be extended to online reviews, suggesting that incoherence may be a sign of fraudulent content. This is validated by Liu et al. (2024), who propose a coherence-based ranking system for online reviews,

highlighting the influence of coherence on consumer decision-making efficiency and its significance in enhancing marketplace integrity.

BERT and BART Scores have emerged as key in understanding the structural, stylistic, and semantic coherence of online reviews. Authors such as Koh (2011) have developed methodologies that quantify the sentiment in online reviews, acknowledging the profound impact of linguistic coherence on product sales and the importance of sentiments beyond mere numerical ratings. This has significant implications, as demonstrated by Purnawirawan et al. (2014), where coherence in review content alongside source credibility was shown to significantly influence readers' perceptions and intentions in online review scenarios. In evaluating Polish texts' coherence, Telenyk et al. (2021) employed neural networks and a pre-trained BERT model, showcasing the evolving methodologies in coherence assessment.

The Disco Score, although less prominent in the literature, also has potential in defining the fine line between authentic and fraudulent content. Research by Bhāle and Tongare (2018) focused on profiling online hotel reviews to distinguish between genuine and fake content, comparing different travel websites to understand the variances brought about by online review structures.

Consequently, all these research controbutions show the importance of coherence in online reviews, not only as a symbol of authenticity but also as a determinant of consumer trust and decision-making. By leveraging advanced computational tools and linguistic theories, researchers and practitioners alike can better navigate the complexities of online consumer feedback, ensuring the reliability and integrity of user-generated content in digital marketplaces.

3. Methodology

The datasets employed in this paper consist of online reviews across various Amazon product categories, which were obtained from the work of Ni et al. (2019). These datasets contain millions of reviews classified into 29 different product categories. To test our methodology we have focused only on the product category "*Beauty Products"*.

Limiting the analysis to one product category ensures more coherence among the reviews, as they concern the same topic. Our dataset has been preprocessed to keep only the following useful information: "review ID, text of the review and the *verified purchase* label. Such label allows us to verify the authenticity of the authorship of the review, so we can asure that the review has been written by a human verified by Amazon.

To develop the generation of artificial reviews (AI-generated text) we use a Large Language Model (LLM), particularly Llama-2, which is ranked currently as one of the state-of-the-art for open-source models and widely used in research (Touvron et al. 2023). We need to fine-tune the model to generate more accurate text regarding the topic analyzed.

Figure 1 shows an overview of the methodology. In *step #1* we fine-tune a Llama-2 model to generate artificial reviews. We use two different models, 7-billion-parameter and 13-billion-parameter, which differ in the size (number of parameters) of the model. Due to limitations in computational resources, it was not feasible to employ a larger model, such as the 70-billion-parameter one. As an input to the fine-tuning we give to both models 30,000 reviews from the category "*Beauty Products"* that have the *verified purchase* label. Subsequently, in *step #2*, we generate 10,000 artificial reviews using each model, aiming to compare these with verified real reviews. In the last step, step #3, we apply the three metrics most commonly used in the literature for measuring text coherence: *BERTScore*, *BARTScore*, and *DiscoScore*. These three metrics are applied to three distinct datasets: a chunk of real reviews obtained from the *verified purchase* input data, reviews generated by the Llama-2-7b model, and reviews generated using the Llama-2-13b model.

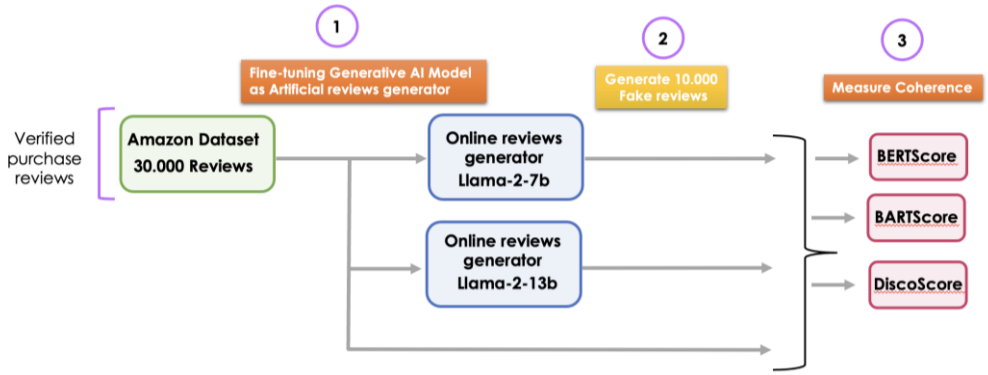

Figure 1. Outline of the methodology

BERTScore, *BARTScore*, and *DiscoScore* are three distinct metrics designed to evaluate text from various angles, emphasizing semantic similarity, text generation quality, and discourse coherence, respectively. *BERTScore* assesses semantic similarity by comparing the embeddings of tokens generated by BERT between a reference text and a candidate text. It calculates recall, precision, and F1 scores based on the maximum cosine similarities of token pairs, emphasizing the semantic overlap and accuracy of the information conveyed. *BARTScore*, on the other hand, leverages BART, a pre-trained text generation model, to predict words or sentences in the reference text, thus evaluating the generated text's ability to capture the intended meaning and information. It uses a log probability score, aiming for a score closer to zero, to gauge the effectiveness of text generation in terms of fluency and coherence. *DiscoScore* specifically targets discourse coherence using BERT to analyze how well sentences and ideas flow together in a text. It focuses on nominal and semantic entity focus, tracking important nouns and broader semantic entities to understand how ideas relate. This metric examines focus continuity and

coherence relationships, measuring the discourse coherence beyond mere semantic similarity or text generation quality.

4. Results

For fine-tuning the generative AI models, generating artificial reviews, and applying the targeted metrics, we used Google Colab's free membership, which offers a T4 GPU with 16GB of VRAM. We have systematically applied this metrics to 100 reviews on each dataset. We systematically applied these metrics to 100 reviews in each dataset. However, an extension of the analysis to a larger set of reviews was constrained by limitations in our computational resources.

Table 1 presents the preliminary results of our study, demonstrating consistency across the three metrics. *BERTScore* and *DiscoScore* report more coherence as the score approaches to 1, while *DiscoScore* reports more coherence with the score closer to 0. The results show that, contratiwise, AI-generated reviews show more coherence in all metrics: up to 10% more in *DiscoScore*, 30% in *BARTScore* and 3% in *BERTScore*. This can be explained as generative AI models are designed to ensure the coherence of the text they generate, in contrast to human users who might not prioritize coherence when writing reviews. Surprisingly, Llama-2-7b demonstrates greater coherence than Llama-2-13b, despite being a smaller model with apparently lower performance. This phenomenon could be attributed to the bigger LLM model requiring more refined fine-tuning, possibly involving a larger dataset of reviews or more optimized hyperparameters.

Table 1. Results of the three metrics and the 3 datasets analyzed

Metric Applied	***Verified Review***	***Llama-2-7b***	***Llama-2-13b***
BERTScore	0.842	0.867	0.862
BARTScore	-4.434	-3.131	-3.235
DiscoScore	0.780	0.860	0.849

5. Conclusions

In this work we have developed a methodology to analyze the coherence in AI-generated text. Through a process of fine-tuning we ensure that our generative AI model generates more accurate text regarding the topic analyzed, allowing a more objective comparising respect to the verified reviews. We use 30.000 verified amazon reviews to fine-tune two different generative AI models: Llama-2-7b and Llama-2-13b, which we use afterwards to generate 10,000 artificial reviews with each one. We apply systematically 3 state-of-the-art metrics (*BERT Score*, *BART Score*, and *Disco Score*) to measure the coherence through a subset of 100 reviews from the three datasets. Results show that AI-generated text shows more coherence, up to 30% more. These results highlight the challenge to distinguish between human and AI-generated reviews,

and the impact on consumer trust in digital marketplaces. In our current pipeline of work we plan to extend this analysis to a broader dataset of reviews, and to add more metrics to measure the quality of the text generated, as the readability or qualitative characteristics of the review (length, keywords, topic modelling, etc).

References

Ai, L., Gao, B., Zheng, J., & Gao, M. (2019, December). On Improving Text Generation Via Integrating Text Coherence. In *2019 IEEE 6th International Conference on Cloud Computing and Intelligence Systems* (CCIS) (pp. 6-10). IEEE.

Bamberg, B. (1983). What makes a text coherent?. *College Composition and Communication*, 34(4), 417-429.

Bandyopadhyay T., Saha S., Pal D., (2023). Beyond Imitation: Exploring Novelty in Generative AI, *International Journal of Advanced Research in Science Communication and Technology*, 3 (2).

Bhāle, S., & Tongare, K. (2018). An empirical investigation of gist helpfulness in online reviews. *Journal of Business and Retail Management Research*, 13(02).

Cui, B., Li, Y., Zhang, Y., & Zhang, Z. (2017, November). Text coherence analysis based on deep neural network. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 2027-2030).

Elkhatat, A. M., Elsaid, K., & Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity*, 19(1), 17.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25(2-3), 285-307.

Giora, R. (1985). Notes towards a theory of text coherence. *Poetics Today*, 6(4), 699-715.

Hu, N., Bose, I., Koh, N. S., & Liu, L. (2012). Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decision Support Systems*, 52(3), 674-684.

Kaghazgaran, P., Wang, J., Huang, R., & Caverlee, J. (2020, July). Adore: Aspect dependent online review labeling for review generation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1021-1030).

Knight, S., & Bart, Y. (2023). Generative AI and User-Generated Content: Evidence from Online Reviews. Available at *SSRN 4621982*.

Koh, N. S. (2011). The valuation of user-generated content: a structural, stylistic and semantic analysis of online reviews. *Singapore Management University*.

Liu, Y., Qiao, D., & Li, X. (2024). In Coherence We Trust: Analyzing Effects of Discourse Coherence in Online Reviews. Available at *SSRN* 4714241.

McCarthy, K. S., & McNamara, D. S. (2021). The multidimensional knowledge in text comprehension framework. *Educational Psychologist*, 56(3), 196-214.

Ni, J., Li, J., & McAuley, J. (2019, November). Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In Proceedings of the 2019 Conference on

Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 188-197).

Ooi, K. B., Tan, G. W. H., Al-Emran, M., Al-Sharafi, M. A., Capatina, A., Chakraborty, A., & Wong, L. W. (2023). The potential of generative artificial intelligence across disciplines: Perspectives and future directions. *Journal of Computer Information Systems*, 1-32.

Purnawirawan, N., Dens, N., & De Pelsmacker, P. (2014). Expert reviewers beware! The effects of review set balance, review source and review content on consumer responses to online reviews. *Journal of Electronic Commerce Research*, 15(3), 162-178.

Shalom, O. S., Uziel, G., & Kantor, A. (2019, September). A generative model for review-based recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems* (pp. 353-357).

Singh, I., Deepak, P., & Anoop, K. (2020). On the coherence of fake news articles. In *ECML PKDD 2020 Workshops: Workshops of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2020): Ghent*, Belgium, September 14–18, 2020, Proceedings (pp. 591-607). Springer International Publishing.

Telenyk, S., Pogorilyy, S., & Kramov, A. (2021). Evaluation of the coherence of Polish texts using neural network models. *Applied Sciences*, 11(7), 3210.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Wang, S., Jiang, W., & Chen, S. (2022, December). An Aspect-Based Semi-supervised Generative Model for Online Review Spam Detection. In *International Conference on Ubiquitous Security* (pp. 207-219). Singapore: Springer Nature Singapore.

Yang, X. (2015, January). One methodology for spam review detection based on review coherence metrics. In *Proceedings of 2015 international conference on intelligent computing and internet of things* (pp. 99-102). IEEE.

Zang, H., & Wan, X. (2017, September). Towards automatic generation of product reviews from aspect-sentiment scores. In *Proceedings of the 10th International Conference on Natural Language Generation* (pp. 168-177).