Journal of Information Systems Engineering and Management
2025, 10(29s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Assessing Textual Conformity and Originality in AI-Generated Content: An Analytical Approach

Marián Chrobák
PhD Student, University of Economics in Bratislava, Slovakia. marian.chrobak@student.euba.sk

ARTICLE INFO

Received: 28 Dec 2024

Revised: 18 Feb 2025

Accepted: 26 Feb 2025

ABSTRACT

The rapid increase in the use of artificial intelligence (AI) has also been seen in the generation of texts in various fields, from academic papers to marketing content. At the same time, there is a growing need for efficient and accurate comparison between AI-generated texts in terms of form and content. Using a screening design of experiment, this contribution compares the influence of selected factors on the similarity and originality of texts generated by artificial intelligence. The response is the result of a formal comparison of the agreement of the texts generated by two anti-plagiarism programs commonly used to check the originality of diploma theses. The used method, with a certain degree of reliability, points to the influence of various factors on the originality of texts generated in this way and provides a new insight into the dynamics of textual creativity and innovation in the era of artificial intelligence. It is a helpful tool for researchers, educators and professionals from various fields who want to ensure that the texts produced by the user interface are original and do not violate copyright.

Keywords: Artificial Intelligence, Text Generation, ChatGPT, Originality Measurement, Plagiarism Detection

INTRODUCTION

Artificial Intelligence (AI) is a phenomenon that has recently occupied many scientists, educators, but is also of interest to practitioners and the general public. However, their views on AI show a degree of polarisation, ranging from fear of the apocalyptic consequences of a takeover to an unbounded confidence in the infallibility to genius of this technical invention. There has also been a qualitative leap in AI's ability to generate meaningful text. This phenomenon brings new challenges, but also questions about textual congruence and originality in the content generated by AI. The use of AI for text generation also has its detractors, but also proponents and admirers [1]–[3]. On the one hand, many attribute too much expertise to AI-generated texts. On the other hand, opponents talk about empty sentences not containing content to copyright issues. Students have a special approach to AI. On the one hand, there is great enthusiasm because they are excited by the tempting idea of having assignments worked out for them by AI. On the other hand, this brings some concern because with the development of AI, the various tools available to educators to reveal the authorship of written term papers or theses have become more stringent [4]–[6]. Issues related to textual matching and originality create a fundamental problem in this context. This is because they have a direct impact on copyright and intellectual property protection. Moreover, the credibility of content is nowadays a key factor and therefore it is important to have tools to assess whether the generated content is authentic and original [7], [8]. There are several methods and tools to assess textual consistency and originality in AI-generated content. These include the use of statistical models, comparison with existing databases, and the use of plagiarism detection tools [9], [10]. Even less attention has been paid to the various factors influencing the degree of originality of the text generated by AI. What impact does the language in which the text is generated, the repeatability of the assignment, the reproducibility, or the focus of the text itself have on originality? It is the aforementioned problem that motivated the writing of this article.

METHODOLOGY AND METHODS

We used the Design of Experiment method, namely Full Factorial Design [11], [12], to assess the influence of the selected factors. The latter involves systematically testing all possible combinations of the levels of all factors involved in the experiment. Thus, we examine each combination of factors to determine their effect on the outcome variable, the response.

The implementation of the experiment consisted of simulating the writing of a thesis for different combinations of levels of the selected factors and asking the AI to write an introduction of approximately 1200 words. Three different Submitters were asked to generate the texts, of which two Submitters used the ChatGPT-3.5 model and the third Submitter used the ChatGPT-4 model to generate the texts. By varying the submitters, i.e., the senders of the text generation request from different IP addresses (different computers), the impact of reproducibility on the originality of the generated text was assessed.

To assess the originality of the generated texts on the basis of very similar but formally different assignments, four papers from the field of economics (A1, A2, A3, and A4) were used. The first two assignments (A1, and A2), although different in title, are practically almost identical in terms of research area. The requirement was for them to be scientific and sophisticated, and therefore for some precision in the definition of concepts and the formulation of relationships. The theory of the economic health of a company uses the same indicators as the theory of bankruptcy. This means that their definitions should be very similar, if not identical. The last two topics have a looser assignment in terms of precision and scientificity, but they also overlap in content. Thus, in total, there are 4 levels of the same factor called "Assignment". The assignments and annotations of all 4 papers are given in Table 1.

Table 1 Assignment for AI

Assignment A1	**Assignment:** Indicators of the company's economic health.
	Annotation: Propose a sophisticated model for evaluating economic indicators that characterize the financial health of a manufacturing enterprise. Estimate its performance. Scientifically justify its contribution to theory and practice. Suggest a way of implementing the model in the practical conditions of the company. Define the importance of using such models as decision support for senior managers.
Assignment A2	**Assignment:** Indicators of the bankruptcy model in the enterprise
	Annotation: Propose a sophisticated model for evaluating the economic indicators that characterize the bankruptcy theory of a manufacturing firm. Estimate its performance. Scientifically justify its contribution to theory and practice. Suggest a way of implementing the model in the practical conditions of the company. Define the importance of using such models as decision support for senior managers.
Assignment A3	**Assignment:** Social aspects of potential bankruptcy in an enterprise
	Annotation: Using a concrete example from practice to describe the social aspects of a potential bankruptcy in a company, outline the basic characteristics of a manufacturing company that is potentially at risk of bankruptcy. Describe the method of averting bankruptcy in the practical conditions of a company. Define the importance of these measures as part of the decision-making process for top managers.
Assignment A4	**Assignment:** Social aspects of assessing the economic health of the enterprise
	Annotation: Using a concrete example from practice to describe the social aspects of assessing the economic health of a company, outline the basic characteristics of a production company in relation to its economic health. Describe the importance of assessing economic health in the practical company conditions. Define the importance of follow-up measures as part of the decision-making process for top managers

We note that all four assignments are quite similar in content, which was actually the intention.

Another factor that was examined was the "Language" of the generated text. All assignments were first entered into the AI in the Slovak language, and then, after translation of the assignments, they were entered into the AI in the English language. This means that this factor had two levels namely Slovak and English language. The aim of such an approach was to assess the difference between the originality of texts written in English or Slovak language. The Slovak language in the experiment represented the language with marginal representation in the database on which the AI (specifically ChatGPT) relied when generating the texts. In contrast, the English language is dominant in this context and thus appears to generate more meaningful combinations when generating texts. Another factor within the experiment was "Repeatability". Each relevant assignment, from each assignor, in both languages was entered into the AI repeatedly after a certain amount of time. This factor thus had two levels. We did not consider time itself as a factor. We reduced its possible negative effect by randomizing the order of trials of the experiment.

To determine the response rate, we evaluated the generated texts using two selected anti-plagiarism systems commonly used in the originality checking of theses at universities. We selected such anti-plagiarism systems which, in addition to the English language, had a non-negligible database in the Slovak language, in order to eliminate at least a little possible disproportion between the English text and the Slovak text.

Thus, the response was the arithmetic mean of two numerical variables representing the percentage of agreement with other texts in the available database of the two anti-plagiarism systems mentioned above. We did not distinguish between the theses being compared that we had entered into the system and the theses that were previously in the system. The goal was simply to determine the originality of the generated texts, which simulated the introduction to a thesis of at least 1200 words. It should be noted that ChatGPT did not provide such a range of text on the first try and therefore it was necessary to repeatedly ask for further continuation. Each text generated was approximately 1200 words in length. These texts, i.e. the introductions of the final papers, were fed into both anti-plagiarism detection systems and their percentage scores were used as the basis for calculating the response in the planned experiment.

The response was described as “Originality”. It should be noted that the temporal sequence of individual text generation and insertion into the anti-plagiarism systems was strictly controlled based on the randomization of the experimental trials. When the experiment was carried out, it was found that the influence of texts other than those inserted by us on each other's originality was minimal. Judgments within the experiment of the entered texts with texts outside the experiment showed minimal agreement. In order to avoid confounding the measurement of previously entered texts in the system, for each text the agreement with other texts entered after it was taken into account. In fact, both anti-plagiarism systems used provided a breakdown of partial hits with other texts, in addition to an overall numerical expression of the hits. By comparing and then summing the individual partial hits, the total hit of the text under consideration was estimated. Specifically, when assessing the first text in each language, almost zero agreement was recorded, so it was waited until a given text showed agreement with the other texts under consideration. These were then counted and matched to the text in question. This avoided some contamination of the data, which would have resulted in the text inserted first being judged as original, while the text inserted later would have been considered plagiarized.

For two factors, i. e. “Repeatability and Reproducibility” (R&R) [13] the inspiration was the Measurement System Analysis (MSA) method [14]. In the context of MSA, R&R are the key components of statistical evaluation of a measurement system, which are aimed at assessing the reliability and accuracy of the measurement. If the MSA specifies that the test cannot be repeated (which is our example), the R&R study may be limited in its ability to fully characterize the variability of the measurement system. In such cases:

- Repeatability – repeatability may be limited to one-off measurements, which means that the results may be less reliable and represent only an instantaneous picture of variability.
- Reproducibility – Inter-operator repeatability can be investigated by having different operators make measurements under their specific conditions, but without the possibility of repeating the measurement on the same sample.

It was originally intended to use R&R to assess the reliability of the two anti-plagiarism systems used. However, when the experiment was carried out, it was found that the differences in determining the hit rate of the two anti-plagiarism programs used, for all the texts generated by us, did not exceed 1%. Therefore, the assessment in question R&R has not been implemented. However, something else is the assessment of R&R when generating the texts. These were considered as separate factors of the experiment implemented. The Analysis of Variance method was used to assess the influence of the factors on the response (ANOVA) [15].

EVALUATION OF THE ORIGINALITY OF THE TEXTS

A Full Factorial Design with 4 factors and 1 response was used to construct the experimental design. The individual factors had different numbers of levels of each other (see Table 2).

Table 2 Description of the levels of the factors of the experiment

Factor	Levels	Values
Submitter	3	S1; S2; S3
Assignment	4	A1; A2; A3; A4
Repeatability	2	R1; R2
Language	2	English; Slovak

In total, texts were generated based on the factor "Assignment A1" to "Assignment A4" (see Table 1) in two versions of the factor "Repeatability R1" and "Repeatability R2", the factor "Submitters S1", "Submitters S2" and "Submitters S3" in two language versions, i.e., the factor "Language English" and "Language Slovak".

It should be noted that the English versions of the texts were created by translating the original Slovak language assignment into English and then using ChatGPT to generate the text from the English version of the assignment. Thus, using AI via ChatGPT, 2 x 3 x 4 x 2 = 48 simulated different texts (introductions to theses) were produced.

All of these texts were then subjected to the classically used system for detecting the correspondence of the theses with other theses. The Minitab software product was used to create the design of the experiment. The generation of the texts was done in a randomized order, i.e. the order was determined randomly by the computer.

The actual results of the experiment were analyzed using the aforementioned Minitab software.

EXPERIMENTAL RESULTS

Based on the results obtained by implementing the Full Factorial Design, namely generating texts using "ChatGBT" for all combinations of each factor level and then assessing their agreement using the two anti-plagiarism systems, the response for all 48 trials of the experiment was determined using ANOVA. The results of this analysis are presented in Table 3.

Table 3 Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-value
Submitter	2	389.01	194.504	1.35	0.272
Assignment	3	1 578.17	526.055	3.64	0.021
Repeatability	1	7.44	7.442	0.05	0.822
Language	1	12.71	12.710	0.09	0.768
Error	40	5 781.82	144.545		
Total	47	7 769.14			

ANOVA shows that the factors ("Submitter", "Repeatability" and "Language") did not show statistically significant effects ("P-value > 0.05") indicating that these characteristics do not have a significant effect on the originality measures in this dataset. However, there is a statistically significant difference in originality between the different assignments. In fact, the factor "Assignment" showed a statistically significant effect on originality. This suggests that differences in assignments are important and affect the originality of papers. The aforementioned result is confirmed by the Main Effects Plot (see Fig. 1), where significant differences in effects are only observed at the levels of the factor "Assignment". Some non-significant changes can also be observed for the factor "Submitter", but the ANOVA clearly showed that these are statistically insignificant changes.

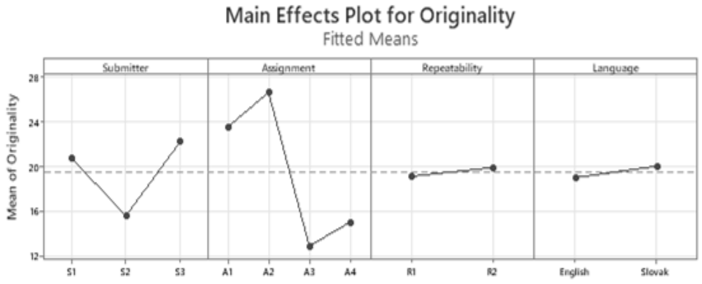

Fig. 1 Main Effects Plot for Originality

Pairwise comparisons of the interactions of each factor are presented in Fig. 2. The interaction plot allows to quickly and intuitively identify if and how the factors interact. The graph provides information for deciding how to adjust the factors to achieve a desired response or for identifying configurations that should be subject to optimization. For example, consider a graph that shows the differences in originality levels for different "Assignment A1" to "Assignment A4" and their "Repeatability R1" and "Repeatability R2". The non-significant interaction suggests that the effect of repetition is consistent across different assignments.

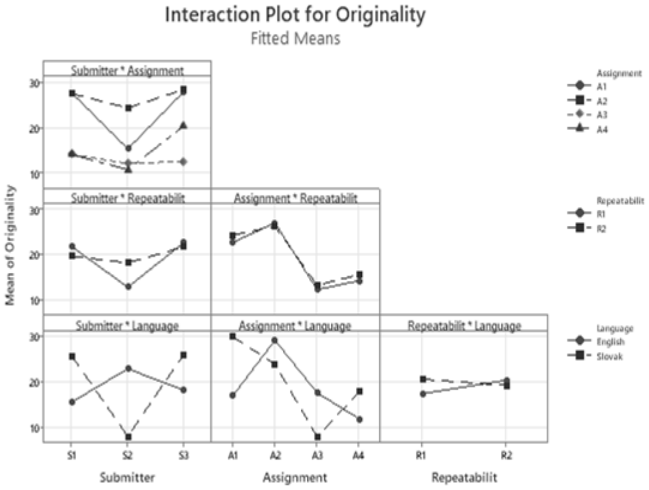

Fig. 2 Interaction Plot for Originality

The negligible variability in originality between repetitions for individual tasks may indicate that text originality is robust or the AI model is insensitive to random fluctuations in the quality of the generated content. A similar procedure was used to interpret the other interactions. For the interaction graph, it is important to observe whether the lines for different levels of factor pairs do not intersect or cross.

Crossing lines indicate an interaction between factors, meaning that the effect of one factor on the response varies depending on the level of the other factor. Since, based on the ANOVA conducted, the factor "Assignment" has a significant effect on the response, we will use the Bonferroni post hoc test (Table 4) to analyze in more detail the effect of its levels on the originality of the generated texts (Table 4).

Table 4 Analysis of Variance

group 1	group 2	meandiff	p-adj	lower	upper	reject
A1	A2	3.1000	0.9184	-9.8297	16.0297	False
A1	A3	-10.6750	0.1379	-23.6047	2.2547	False
A1	A4	-8.5333	0.3050	-21.4631	4.3964	False
A2	A3	-13.7750	0.0328	-26.7047	-0.8453	True
A2	A4	-11.6333	0.0915	-24.5631	1.2964	False
A3	A4	2.1417	0.9708	-10.7881	15.0714	False

In the Bonferroni post hoc test results table, each column has a specific meaning. The columns "group 1" and "group 2", identify the groups that are being compared. The column "meandiff" indicates the difference in means between the two groups. The "p-adj" column represents the p-value after adjusting for multiple comparisons (a p-value of less than 0.05 indicates that the difference in means is statistically significant). The "lower" and "upper" columns indicate the lower and upper bounds of the 95% confidence interval for the difference in means. If this interval includes zero, it usually means that the difference in means is not statistically significant. The "reject" column indicates whether the null hypothesis of equality of means between the two groups has been rejected. A single comparison of "Assignment A2" versus "Assignment A3" showed a statistically significant difference with a p-value of 0.0328, with the null hypothesis rejected (reject = True). This result shows that "Assignment A2", i.e. indicators of the bankruptcy model in the enterprise and "Assignment A3", i.e. social aspects of potential bankruptcy in an enterprise have significantly different levels of originality, with "Assignment A2" having a higher degree of agreement than "Assignment A3". Subsequently, the results of the experiment were visualized using Box plot graph.

Fig. 3 presents the variability of the measured hits for each task in terms of the two languages used.

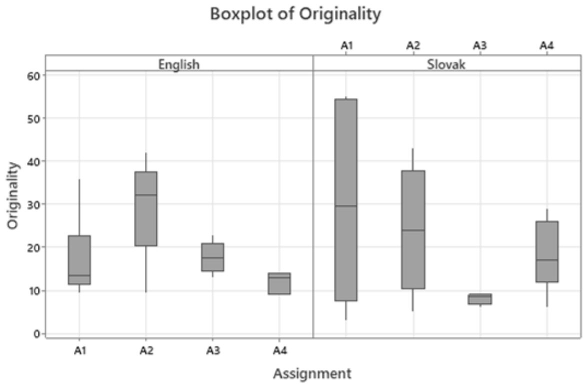

Fig. 3 Boxplot of Originality by Language

The visualization shows some, albeit statistically insignificant, need for attention for languages other than English. In fact, when generating the texts, the originality between the English and Slovak texts differed. These differences suggest that linguistic nuances may affect the evaluation or performance of originality. Fig. 4 shows how the variability of the degree of agreement varies between the different Repeated Assignments (Repeatability) for each relevant Assignment.

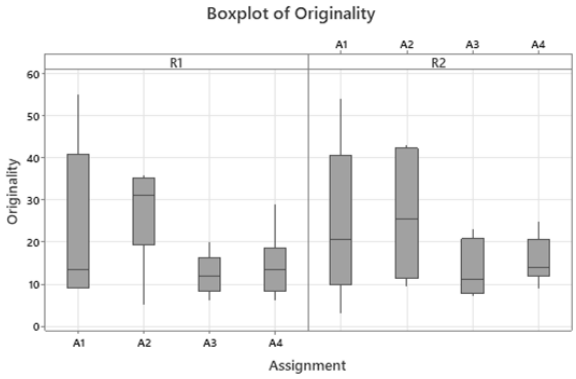

Fig. 4 Boxplot of Originality by Repeatability

The box plot allows to visually assess the variance and the means of the originality values within the aforementioned categories. Fig. 5 shows how the variability of agreement varies between different “Assignment A1” to “Assignment A4” and between different “Submitter S1” to “Submitters S3”.

This graph visualizes the variability of hits in the relationship between specific Assignments and different Senders. It visualizes the dynamics by which relevant factors influence the degree of agreement of the generated texts in the experimental design.

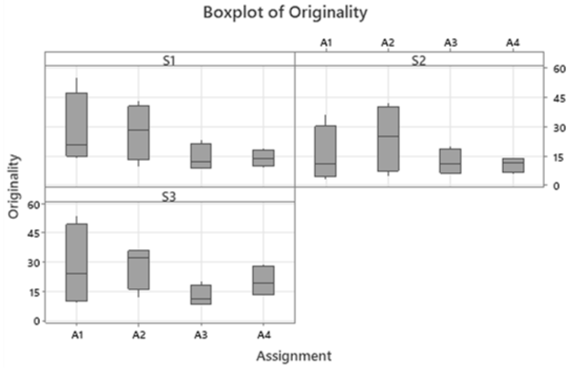

Fig. 5 Boxplot of Originality by Submitter

DISCUSSION

The results of the experiment showed that the factor “Assignment” has a statistically significant effect on the originality of the texts. This points to the fact that differences in assignments can have a direct impact on how the AI generates original content. This may be of interest to AI developers and researchers working to improve generative models for specific applications such as academic papers, marketing content, or even literary production. Bonferroni's post hoc test subsequently showed that the most significant difference in text originality is between the assignment (A2) and the assignment (A3). This suggests that the difference in the requirement for sophistication and complexity of topics may have a significant effect on the originality of texts. Specifically, the results highlighted the need to investigate how different types of tasks affect the creativity or originality of final papers. The factors “Submitter," “Repeatability," and “Language" did not show statistically significant effects on originality, indicating that these characteristics do not have a significant effect on measures of originality, at least in this dataset.

In discussing the results of the experiment to assess the originality of AI-generated texts, it can be concluded that the analytical findings should be further verified and expanded in broader and deeper studies to confirm their consistency and applicability in real-world settings. This is because larger studies create the conditions for a deeper penetration into the implications of the findings. They will also allow for a more accurate contextualisation of them in comparison to the existing literature and thus the possibility to suggest possible directions for future research and practical applications.

According to the literature, AI models such as GPT (Generative Pre-trained Transformer) are known to be highly dependent on the quality and diversity of the training data. The results of this experiment are consistent with this theory and show that different tasks can stimulate AI to produce more or less original content. This is confirmed by recently published studies [6]-[8] that investigate how context and question wording affect AI outputs. A comparison of the findings from the experiment with the existing literature shows that the design and results are in line with theoretical and empirical knowledge on AI performance in text generation. In particular, the differences in assignments and their impact on text originality engage the discussion of how AI models respond to different types of input. Reference [16] shows on AI text generating models (such as GPT models) are dependent on the diversity and quality of the training data.

The experiment confirmed that different “Assignments” had a significant impact on the originality of the texts, indicating how the models respond to specific inputs based on their prior “learning”. This is important for understanding the possible limitations of AI in generating original content depending on the similarity of the training data. References [17], [18] shows that the AI's response to different formulations of a task can significantly affect the quality and originality of the generated text. Experimental findings where different tasks produced different degrees of originality illustrate how the AI interprets and responds to different types of instructions, which is crucial for the development of more efficient text generation systems. These sources highlight the need for a deeper understanding of the impact of training data and task on AI performance in text generation. The experimental results provide practical evidence that supports these theoretical assumptions and suggests the need for further research in this area, particularly in the context of education and the ethical use of AI.

These findings can serve as a basis for further studies that could investigate how different AI settings, such as different model versions (e. g. ChatGPT-3.5 versus ChatGPT-4) or language versions, affect the originality and usefulness of the generated text. In addition, it would be useful to investigate how other characteristics of the task, such as length and complexity, affect the AI's ability to generate original content.

CONCLUSION

The results of the experiment conducted suggest that Assignment Type has an impact on the originality values of the papers. This is useful in deciding the importance of individual assignments in the context of their impact on originality. This finding may be important for planning educational activities or for studies related to creativity in different contexts. Further research in this area could help to better understand how different factors influence AI's ability to produce original and authentic content, which is crucial for its ethical use in education, publishing, and commercial applications.

REFERENCES

[1] H. El Mostafa, and F. Benabbou, “A deep learning based technique for plagiarism detection: a comparative study,” IAES Int. J. Artif. Intell., vol. 9, no. 1, pp. 81-90, 2020.

[2] W. Binder, “Technology as (dis-) enchantment. AlphaGo and the meaning-making of artificial intelligence,” Cult. Sociol., vol. 18, no. 1, pp. 24-47, 2024.

[3] J. Rudolph, S. Tan, and S. Tan, “ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?” J. Appl. Learn. Teach., vol. 6, no. 1, pp. 342-363, 2023.

[4] K. I. Roumeliotis, and N. D. Tselikas, N. D., “ChatGPT and Open-AI Models: A Preliminary Review,” Future Internet, vol. 15, no. 192, pp. 1-24, 2023.

[5] T. Foltýnek, N. Meuschke, and B. Gipp, “Academic plagiarism detection: a systematic literature review,” ACM Comput. Surv., vol. 52, no. 6, pp. 1-42, 2019.

[6] M. A. B. Tobar, M. G. van den Brand, and A. Serebrenik, “Cross-Language Plagiarism Detection: Methods, Tools, and Challenges: A Systematic Review,” Int. J. Adv. Sci. Eng. Inf. Technol., vol. 12, no. 2, pp. 589-599, 2022.

[7] M. Jiffriya, M. A. Jahan, and R. Ragel, “Plagiarism detection tools and techniques: A comprehensive survey,” Journal of Science-FAS-SEUSL, vol. 2, no. 02, pp. 47-64, 2021.

[8] J. Wang, and Y. Dong, “Measurement of text similarity: a survey,” Information, vol. 11, no. 9, 421, pp. 1-17, 2020.

[9] D. Gupta, “Study on Extrinsic Text Plagiarism Detection Techniques and Tools,” Journal of Engineering Science & Technology Review, vol. 9, no. 5, 2016.

[10] D. W. Prakoso, A. Abdi, and C. Amrit, “Short text similarity measurement methods: a review. Soft Comput., vol. 25, pp. 4699-4723, 2021.

[11] S. Greenhill, S. Rana, S. Gupta, P. Vellanki, and S. Venkatesh, “Bayesian optimization for adaptive experimental design: A review,” IEEE access, vol. 8, pp. 13937-13948, 2020.

[12] J. Antony, “Design of experiments for engineers and scientists”, Elsevier. 2023.

[13] F. Melchor, R. Rodriguez-Echeverria, J. M. Conejero, A. E. Prieto, and J. D. Gutiérrez, “A model-driven approach for systematic reproducibility and replicability of data science projects,” In International Conference on Advanced Information Systems Engineering, pp. 147-163, Cham: Springer International Publishing, 2022.

[14] U. Lorenzo-Seva, and P. J. Ferrando, “MSA: The forgotten index for identifying inappropriate items before computing exploratory item factor analysis,” Methodology, vol. 17, no. 4, pp. 296-306, 2021.

[15] W. Zhang, and Y. Qi “ANOVA-nSTAT: ANOVA methodology and computational tools in the paradigm of new statistics,“ Comput. Ecol. Softw., vol. 14, no. 1, pp. 48-67, 2024.

[16] N. Schopow, G. Osterhoff, and D. Baur, “Applications of the Natural Language Processing Tool ChatGPT in Clinical Practice: Comparative Study and Augmented Systematic Review,” JMIR Medical Informatics, vol. 11, e48933, 2023.

[17] H. M. Alhaidry, B. Fatani, J. O. Alrayes, A. M. Almana, N. K. Alfhaed, H. Alhaidry, ... and N. K. Alfhaed Sr, “ChatGPT in dentistry: a comprehensive review,” Cureus, vol. 15, no. 4, 2023.

[18] L. Mindner, T. Schlippe, and K. Schaaff, “Classification of human-and ai-generated texts: Investigating features for chatgpt,” In International Conference on Artificial Intelligence in Education Technology, Singapore: Springer Nature Singapore, pp. 152-170, 2023.