Natural Language Processing Algorithms for Academic Content Generation

Alfredo Paredes-Vargas
Information Systems Engineering
Universidad Peruana de ciencias Aplicadas UPC
Lima, Peru
u201911812@upc.edu.pe

Piero Salinas-Llorca
Information Systems Engineering
Universidad Peruana de ciencias Aplicadas UPC
Lima, Peru
u20201C202@upc.edu.pe

José Santisteban
Information Systems Engineering
Universidad Peruana de Ciencias Aplicadas
Lima, Perú
PCSILSAN@upc.edu.pe

2024 IEEE 4th International Conference on Advanced Learning Technologies on Education & Research (ICALTER) | 979-8-3315-1024-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICALTER65499.2024.10819214

***Abstract*—This study has conducted several activities to analyze and compare the performance of five Natural Language Processing (NLP) models: GPT-3, T5, ERNIE, BERT, and XLNet for generating academic content in universities located in Lima, Peru. Some NLP models, including GPT and BERT, were assessed based on accuracy, quality, speed, and customization capabilities. The research examined the use of various algorithms for academic tasks like essay and summary creation. The findings show that all models perform adequately; however, some excel in generating coherent content tailored to different academic styles, thereby optimizing the time of both students and teachers. In conclusion, the ERNIE model emerges as one of the best, as it enables the generation of diverse and high-quality content efficiently.**

Keywords—algorithm, SCORM, NLP, academic content, higher education.

I. Introduction

The creation of educational content encounters various challenges [1]. There is an increasing demand for updated and adaptable materials that address the rapid changes in knowledge and technology [2]. Universities must create content that is accessible, relevant, and promotes student engagement [3]. However, developing this content demands significant time, skilled personnel, and investment in technology and authoring tools. These requirements can hinder institutions' ability to adapt to emerging educational needs. The increasing demand for online learning and the use of Learning Management Systems (LMS) necessitate content that is both high-quality and compliant with international standards like the Sharable Content Object Reference Model (SCORM). This compliance promotes the adoption of global standards in content creation, fostering a more inclusive education that meets the needs of today's digital environment. Additionally, it allows for the reuse and interoperability of educational materials [4].

Generative Artificial Intelligence (GAI) has transformed academic content generation by automatically creating texts, images, simulations, and other educational resources [5]. By utilizing advanced models like Generative Pre-trained Transformer (GPT), GAI can create high-quality, adaptable educational materials, greatly decreasing the time and costs involved in content development [6]. Additionally, it enables educators to focus on personalizing and continuously improving learning, providing more dynamic and interactive experiences for students.

The study aims to evaluate and compare various algorithms to identify which ones provide greater efficiency and quality in developing educational materials. This analysis aims to find suitable solutions for enhancing academic productivity, reducing work hours, and optimizing coherence and accuracy in text generation.

The remainder of the paper is organized as follows: Section 2 discusses methodology, and Section 3 provides the results and analysis. Finally, in Section 4, the conclusion of the presented work is provided.

II. Methodology

To achieve our goal in this research, we used a dataset specifically for this comparison and pre-trained the models with relevant information to assist in generating academic content. The overall framework of the research method is shown in Figure 1.

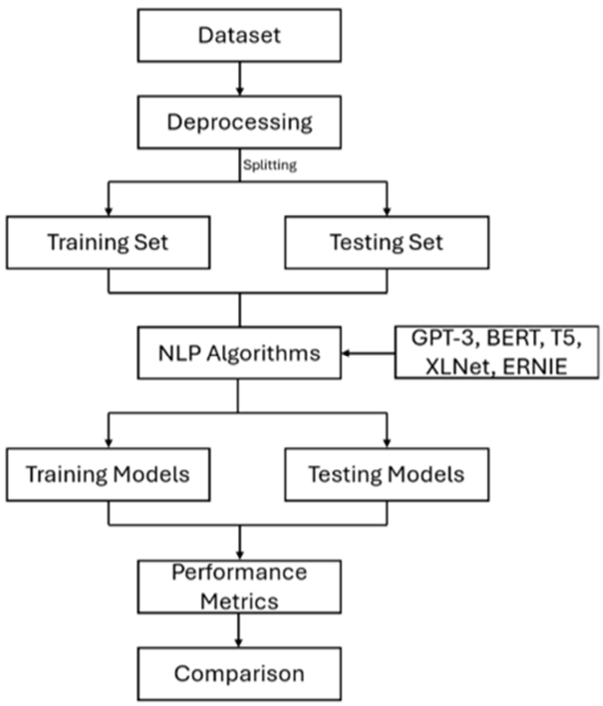

Fig. 1. Steps of the research process

A. Dataset

Stratified sampling was used to ensure thematic diversity and representativeness of the corpus, considering the distribution of academic disciplines in repositories such as arXiv and Google Scholar. This dataset contains a wide range of academic texts, including abstracts and full articles across various disciplines such as engineering, social sciences, and humanities. These resources were chosen for their relevance in producing high-quality academic content in both Spanish and English. Having a broader range of languages facilitates the examination of how algorithms interact with linguistic

979-8-3315-1024-4/24/$31.00 ©2024 IEEE

and thematic structures, which are crucial for the adaptability of generated content [7].

The selection of sources is focused on the accessibility of academic articles that offer representative examples of conducted research. This approach ensures that the NLP algorithms are evaluated not only on the quality of the content they generate but also on their ability to adapt to various academic contexts. In addition, external validation of the results is planned through collaboration with researchers who replicate the study with similar data sets. According to Yin W. et al. [8], datasets that combine content from various sources and fields of study offer a more comprehensive view of the models' ability to generalize. This ability is crucial for this type of analysis. Below in Table 1, we outline the algorithms that will be compared, using metrics such as accuracy, linguistic quality, and content generation time measured in milliseconds “ms”.

TABLE I. DATASET ON THE ALGORITHMS AND THEIR MULTIPLE PERFORMANCE VARIABLES

Algorithm	Input (Prompt)	Coherence	Precision	Linguistic Quality	Generation Time (s)	Length (Words)	Academic Discipline	Expert Evaluation
GPT-3	"Explain the theory of..."	4.5	4.8	4.7	3.5	500	Social Sciences	4.6
BERT	"Describe the process..."	3.8	4.0	3.9	4.0	450	Computer Science	4.1
T5	"List the factors..."	4.0	4.2	4.1	3.8	480	Social Sciences	4.3
XLNet	"Discuss the importance..."	4.2	4.5	4.3	3.7	490	Social Sciences	4.4
ERNIE	"Analyze the evolution..."	4.1	4.3	4.0	4.2	470	Computer Science	4.2

When creating academic content, it is crucial to focus on coherence, linguistic quality, and generation time. These factors ensure that the produced material is relevant, accurate, and easy to understand. Coherence allows for a logical and smooth development of ideas, while accuracy ensures that the information is correct and well-supported. A scoring system ranging from 1 to 5 was chosen to offer more rating options for comparing the activities conducted by each algorithm.

B. Preprocessing

Data preprocessing included cleaning of special characters and labels, text normalization, tokenization, and stopword removal. Model selection focused on cutting-edge NLP technologies such as GPT-3, BERT, XLNet, T5, and ERNIE, excluding simpler algorithms (n-grams, Markov Chains) and less effective approaches for academic text (LSTM, RNN). Recent models such as LaMDA were not considered due to access restrictions. GPT-3, developed by OpenAI, is notable for its ability to produce coherent content across various contexts. This capability stems from its broad recognition and extensive feedback, which facilitate improvements and optimization in the way information is presented [9]. BERT, developed by Google, emphasizes understanding word context bidirectionally, making it especially useful for interpretation tasks like text classification [10].

T5 transforms any activity or task into a text-based problem, facilitating tasks such as summary, translation, or generation. XLNet, which was also developed by Google, employs a pretraining method that effectively captures long-term relationships between words [11]. Finally, we have ERNIE from Baidu, which focuses on integrating structured data to enhance accuracy in content generation [12].

III. RESULTS AND ANALYSIS

In this section, we analyzed the performance of the models in generating academic content. The metrics evaluated include Precision, Recall, F1 Score, and Processing Time (ms). A comparative graph was created to analyze each metric.

A. Precision

Precision is a metric that evaluates the predictions of an NLP model, measuring the ratio of correct predictions to the total predictions made [13]. In academic content generation, precision is essential because a highly precise model reduces the likelihood of producing irrelevant or incorrect information. This aligns with the findings in [16], which suggest that precision is particularly important when the goal is to minimize false positives, thereby resulting in fewer incorrect predictions.

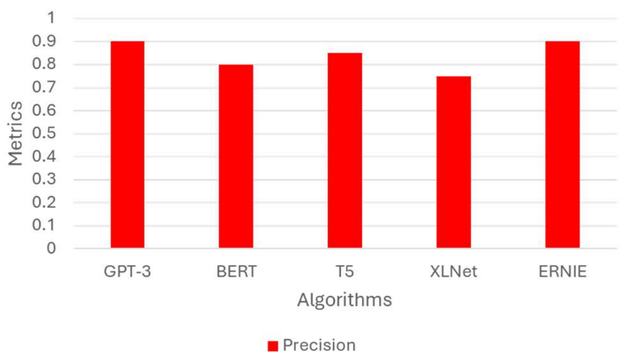

Fig. 2. Comparative analysis of NLP algorithms regarding precision

Figure 2 shows that GPT-3 and ERNIE exhibit outstanding precision levels, differing from the other models by only a few decimal points. This reflects their ability to generate highly relevant academic content while minimizing the production of unrelated information. These results align with the characteristics of GPT-3, which effectively handles broad contexts, and ERNIE, which excels at integrating structured knowledge to enhance content relevance.

B. Recall

Sensitivity indicates a model's capacity to correctly identify all relevant instances within a dataset [14]. Liu et al. [9] emphasizes that sensitivity is essential in NLP tasks aimed at capturing as many relevant instances as possible, particularly in generating comprehensive academic content. A high level of sensitivity indicates that the model or algorithm effectively identifies the most relevant content.

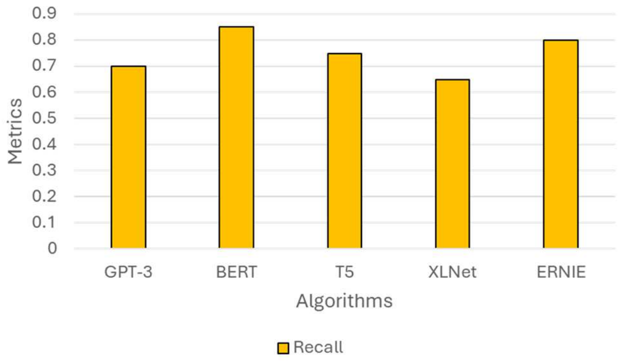

Fig. 3. Comparative analysis of NLP algorithms regarding sensitivity.

According to Figure 3, BERT and ERNIE demonstrate a higher recall compared to the other models. This indicates their superior ability to identify relevant instances within the evaluated data, a critical factor for generating comprehensive and representative academic content. This recall advantage may be linked to BERT's bidirectional architecture and ERNIE's contextual capabilities.

C. F1 Score

It is a metric that integrates precision and sensitivity into a harmonic measure, achieving a balance between the two [16]. Devlin et al. [10] indicate that the F1 Score is especially valuable when a trade-off is necessary to minimize errors or avoid potential false positives by omitting relevant instances. In content generation, a high F1 Score indicates that the model produces not only accurate text but also complete and relevant content.

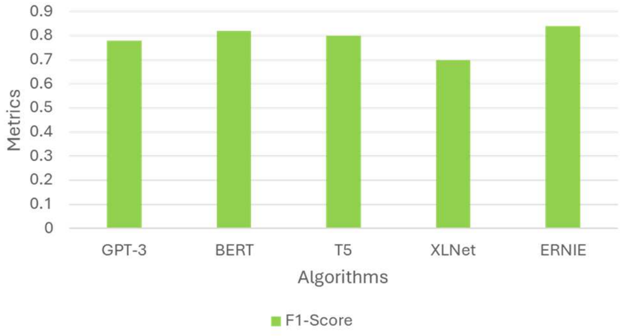

Fig. 4. Comparative analysis of NLP algorithms regarding the F1 Score.

Figure 4 illustrates that ERNIE outperforms the other models in F1-Score, establishing itself as the most balanced in terms of precision and recall. This strong performance reflects its ability to generate academic content that combines accuracy with adequate coverage, minimizing errors while producing complete information.

D. Processing Time

Radford et al. [11] indicate that processing time is a crucial factor in generating academic content, as it affects the model's viability for real-time or large-scale applications.

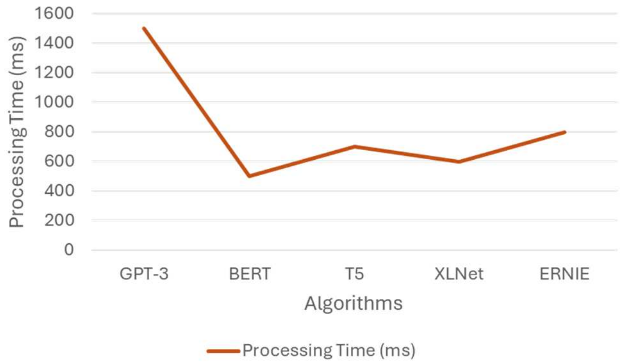

Fig 5. Comparative analysis of NLP algorithms regarding processing time.

Figure 5 highlights that while GPT-3 achieves the highest precision, it also has the longest processing time, which could limit its applicability for tasks requiring real-time results. Conversely, BERT and XLNet achieve significantly shorter processing times, making them ideal for applications where speed is critical, without compromising acceptable performance in other metrics.

E. *Comparative analysis of NLP algorithms*

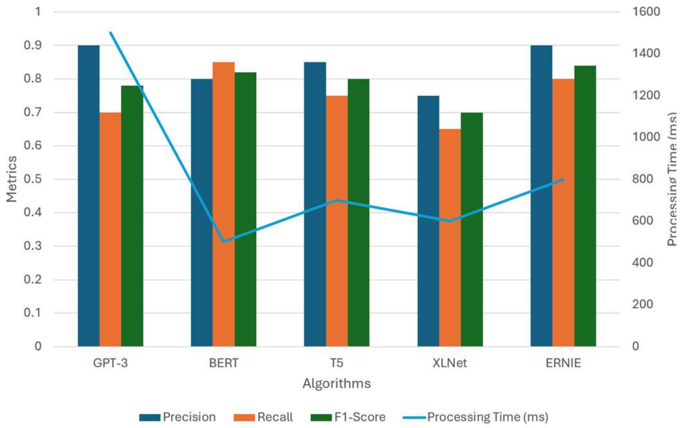

Fig. 6. Comparative analysis of NLP algorithms

Figure 6 illustrates that GPT-3 achieved higher precision, while ERNIE distinguished itself with balanced performance across all metrics, featuring a high F1 score and relatively efficient processing times. Although more complex models like GPT-3 and ERNIE exhibit longer processing times, BERT and XLNet provide a good balance between performance and efficiency.

IV. CONCLUSION

This study was conducted to evaluate and compare five natural language processing (NLP) algorithms like GPT-3, BERT, T5, XLNet, and ERNIE in the generation of academic content, a critical task to address the needs for quality and adaptability in higher education. These models were analyzed in terms of precision, recall, F1-Score, processing time, and fundamental metrics to ensure that the generated texts are coherent, relevant, and effective.

The results revealed that GPT-3 excelled in terms of high precision and linguistic quality, making it ideal for tasks prioritizing the fidelity of the generated content. BERT and XLNet demonstrated a balance between efficiency and processing time, making them suitable for applications where speed is critical. T5, on the other hand, showed solid performance in general content generation tasks, while ERNIE stood out as the most balanced model, thanks to its ability to integrate structured data and effectively handle complex contexts.

However, it is important to consider that NLP algorithms may generate incorrect or inconsistent content, especially when dealing with ambiguous contexts or if the information they base their responses on contains errors or biases. This risk arises from how models interpret context or prioritize certain keywords over a full understanding of the content. Therefore, users must validate the generated information before using it, particularly in academic applications where accuracy and reliability are essential.

ACKNOWLEDGMENT

The authors express their gratitude to the Dirección de Investigación de la Universidad Peruana de Ciencias Aplicadas for the support provided for this research work.

REFERENCES

[1] M. Moreira, "Los materiales educativos: origen y futuro," 2007.

[2] F. Gareis, F. Kendziur, & C. Machiavello, "Los desafíos de la comunicación y la producción de materiales educativos para Atamá, el ambiente virtual de aprendizaje de Entre Ríos," 2017.

[3] C. Gamrat, L. Lenze, J. Bardzell, & E. Glantz, "Improved Student Engagement in Higher Education's Next Normal," 2021.

[4] P. M. Barbetta, "Tecnologías de escritura correctivas y compensatorias para estudiantes de secundaria con problemas de aprendizaje," 2023.

[5] J. Franganillo, "La inteligencia artificial generativa y su impacto en la creación de contenidos mediáticos," 2023.

[6] T. Farrely & N. Baker, "Generative Artificial Intelligence: Implications and Considerations for Higher Education Practice," 2023.

[7] Y. Wenpeng, K. Katharina, & Y. Mo, "Comparative Study of CNN and RNN for Natural Language Processing," 2024.

[8] A. Jain & A. Goel, "for natural language processing models*," 2020.

[9] Y. Liu, M. Ott, N. Goyal, et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," 2019.

[10] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019.

[11] A. Radford et al., "GPT-2: Better Language Models and Their Implications," 2019.

[12] M. Araujo, P. Gonçalves, & F. Benevenuto, "Towards understanding the impact of machine learning datasets on NLP tasks," 2020.

[13] T. Brown et al., "Language models are few-shot learners," 2020.

[14] J. Devlin, M.-W. Chang, K. Lee, & K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2019.

[15] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, & Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," 2019.

[16] Y. Sun, S. Wang, Y. Li, S. Feng, & H. Tian, "ERNIE: Enhanced Representation through Knowledge Integration," 2020.

[17] J. M. Swales & C. B. Feak, "Academic Writing for Graduate Students: Essential Tasks and Skills," 2004.

[18] M. McDonald, "Time Efficiency in Automated Text Generation Systems," 2018