An Offline Semantic Plagiarism Detection System Using Sentence-BERT and DuckDuckGo-Based Web Search

*Note: Sub-titles are not captured in Xplore and should not be used

1st Nguyen Minh Tuan, 2stNguyen Hong Son, 3stChau Van Van
^{1,2,3}*Faculty of Information Technology*
Posts and Telecommunications Institute of Technology
11 Nguyen Dinh Chieu, Sai Gon ward, 700000, Ho Chi Minh city, Viet Nam
minhtuan@ptit.edu.vn; sonngh@ptit.edu.vn; vancv@ptit.edu.vn

***Abstract*—Plagiarism detection remains a critical concern in academic and research environments. Traditional methods often rely on keyword matching or require access to paid web search APIs, which limits their scalability and accessibility. In this paper, we present an offline-capable semantic plagiarism detection system that leverages Sentence-BERT for deep semantic comparison of texts and DuckDuckGo for lightweight web search without API keys. The proposed system extracts text from academic PDFs, splits it into manageable chunks, and searches for potentially plagiarized content using DuckDuckGo. Retrieved web pages are parsed and semantically compared with the input using sentence embeddings and cosine similarity. The system highlights suspicious content directly in the PDF and generates a summary report, providing transparency and interpretability. Our approach offers a practical and cost-effective solution for plagiarism checking in low-resource or restricted-access environments, with promising accuracy and usability for educational institutions, publishers, and independent researchers.**

***Index Terms*—Semantic Plagiarism Detection, Sentence-BERT, Natural Language Processing (NLP), DuckDuckGo Search, Academic Integrity, Cosine Similarity, Offline AI Models, Text Similarity, Web Mining, PDF Text Analysis**

I. INTRODUCTION

Plagiarism remains a significant challenge in academic and professional environments, especially with the increasing availability of digital content. Traditional plagiarism detection tools often rely on surface-level text matching or require access to proprietary databases and cloud-based APIs, which can limit their accessibility, transparency, and reproducibility. In recent years, semantic models such as Sentence-BERT (SBERT) have demonstrated strong capabilities in capturing contextual meaning between sentences, offering a more robust approach to detecting paraphrased or semantically altered content. However, many existing solutions that leverage such models still depend on paid APIs or cloud infrastructure, making them unsuitable for offline or privacy-sensitive environments.

Plagiarism is when someone copies or mimics the words, concepts, and ideas of another writer and passes them off as their unique work. Plagiarism is unacceptable in scientific writing, and if plagiarism is discovered in a Journal of Dairy Science article, it will be rejected along with any updates. In essence, scientific research is an attempt to find the truth about the universe we inhabit. A degree of honesty and professional integrity that might not be necessary in some other fields of effort should be upheld by scientists who conduct this inquiry and report their findings [1]. Writing critical essays is a crucial component of undergraduate education, particularly for students enrolled in the School of Language Studies and Linguistics' Literature in English degree at Universiti Kebangsaan Malaysia. Students must discuss and present relevant issues in writing in an analytical and captivating manner using pertinent citations from published materials in all literature courses, including Critical Appreciation, Gender Identities, and Selected Literary Works, which were the subject of this research project. But we discovered that our students struggle greatly in these areas. Some of the issues were simple and could be resolved by a one-on-one discussion [2].

Claiming to be the creator of someone else's work is known as plagiarism. Copying sentences, phrases, or paragraphs precisely as they are found in the source, rearranging them, substituting synonyms for a few words, or just copying a passage of a paper and adding a few sentences of your own are all examples of plagiarism. It is important to cite the original work, even when paraphrasing a portion of a publication or study conclusions. On the other hand, exploiting someone else's work without their consent is a violation of copyright. Using a table or figure from a previous publication without the author's consent from the journal or publisher is an example of a copyright violation [3]. Because there are so many digital papers available online, plagiarism the act of duplicating someone else's work is on the rise. Because the Internet is so widely used, copying documents is now relatively simple. Complete or partial copies of documents are possible. Numerous document copy detection algorithms have been proposed; however, there isn't one that works well with Malayalam documents. The act of passing off someone

Identify applicable funding agency here. If none, delete this.

else's original words and ideas as one's own is known as plagiarism, and it is considered a moral as well as frequently a legal offense. The sheer volume of information accessible for manual analysis grows as more and more information is made public online. As a result, computer techniques have been developed to support authorship, direction recognition, and text reuse [4]

Machine learning has played a crucial role in enhancing AI development, including voice detection, image recognition, text analysis, and other fields [5]–[12]. This paper introduces an offline-capable plagiarism detection framework that integrates SBERT-based semantic similarity measurement with lightweight web mining using the DuckDuckGo search engine. Our system extracts content from academic PDF documents, performs semantic comparisons between document chunks and retrieved web content, and highlights potential matches with visual overlays. Unlike systems that require Bing or Google API keys, our approach is designed to be fully executable without registration or external credentials, making it more accessible to researchers and educators. We further enhance usability by providing a user-friendly interface via Gradio and generating automated summaries to support decision-making using machine learning [13]–[29]. Experimental evaluations demonstrate that the proposed system offers a practical balance between detection quality and deployment simplicity, especially in resource-constrained or privacy-conscious settings.

II. Methodology

The proposed system is designed to detect semantic plagiarism in academic PDFs using Sentence-BERT and web content retrieved via DuckDuckGo. It operates entirely offline except for the web search component, and avoids the use of paid or restricted APIs. The methodology is composed of the following key stages (shown in Fig. 1):

A. *Text Extraction from PDF*

We begin by extracting the full textual content from a given academic PDF using the PyMuPDF (fitz) library. This ensures support for both single-column and multi-column layouts with minimal loss of formatting.

B. *Chunking the Document*

To allow for finer-grained similarity analysis, the extracted text is segmented into manageable chunks based on sentence boundaries. A fixed token limit (e.g., 300 words) is used to ensure compatibility with Sentence-BERT's maximum sequence length.

C. *DuckDuckGo Web Search*

For each document, we perform a single DuckDuckGo search using a truncated query derived from the first 300 characters of the document. DuckDuckGo is used via the duckduckgo_search package, allowing anonymous, rate-limited access without requiring an API key. The top N search result URLs are collected for further analysis.

D. *Webpage Text Scraping*

Each retrieved URL is visited using requests and parsed using BeautifulSoup. All visible text from paragraph (¡p¿) tags is concatenated and preprocessed to remove scripts, styling, and irrelevant boilerplate.

E. *Semantic Similarity with Sentence-BERT*

The core of the system uses the sentence-transformers library with the all-MiniLM-L6-v2 model. This transformer is selected for its efficiency and strong semantic representation capabilities. Each document chunk is encoded into a dense embedding vector, and compared with the embedding of each source text using cosine similarity. The resulting similarity score is scaled to a percentage.

F. *Plagiarism Highlighting and Report Generation*

For each chunk, the highest similarity score and corresponding source are recorded. If a chunk's similarity score exceeds a predefined threshold (e.g., 20% or more), it is highlighted directly in the PDF using colored overlays via PyMuPDF. A summary page is appended, detailing the number of matches per threshold tier, average similarity, and a ranked list of matching URLs.

G. *Interactive User Interface*

A Gradio-based interface is integrated for ease of use. Users can upload a PDF, run the analysis, and download the annotated output without requiring programming knowledge. The system runs locally and does not transmit user data externally.

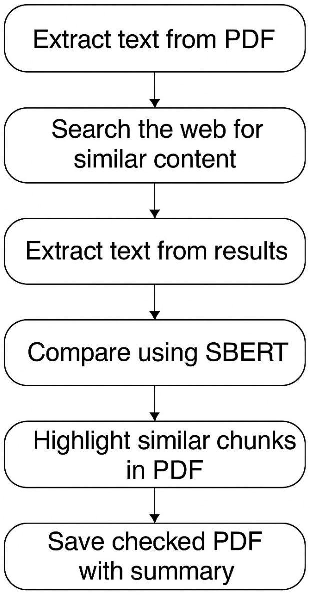
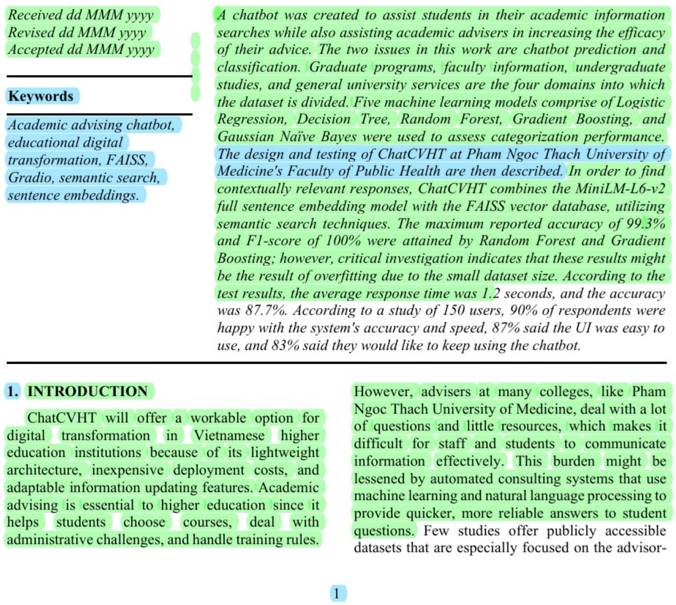

Extract text from PDF
Search the web for similar content
Extract text from results
Compare using SBERT
Highlight similar chunks in PDF
Save checked PDF with summary

Fig. 1. Processing steps for building a plagiarism application.



Received dd MMM yyyy
Revised dd MMM yyyy
Accepted dd MMM yyyy

Keywords

Academic advising chatbot, educational digital transformation, FAISS, Gradio, semantic search, sentence embeddings.

A chatbot was created to assist students in their academic information searches while also assisting academic advisers in increasing the efficacy of their advice. The two issues in this work are chatbot prediction and classification. Graduate programs, faculty information, undergraduate studies, and general university services are the four domains into which the dataset is divided. Five machine learning models comprise of Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, and Gaussian Naïve Bayes were used to assess categorization performance. The design and testing of ChatCVHT at Pham Ngoc Thach University of Medicine's Faculty of Public Health are then described. In order to find contextually relevant responses, ChatCVHT combines the MiniLM-L6-v2 full sentence embedding model with the FAISS vector database, utilizing semantic search techniques. The maximum reported accuracy of 99.3% and F1-score of 100% were attained by Random Forest and Gradient Boosting; however, critical investigation indicates that these results might be the result of overfitting due to the small dataset size. According to the test results, the average response time was 1.2 seconds, and the accuracy was 87.7%. According to a study of 150 users, 90% of respondents were happy with the system's accuracy and speed, 87% said the UI was easy to use, and 83% said they would like to keep using the chatbot.

1. INTRODUCTION

ChatCVHT will offer a workable option for digital transformation in Vietnamese higher education institutions because of its lightweight architecture, inexpensive deployment costs, and adaptable information updating features. Academic advising is essential to higher education since it helps students choose courses, deal with administrative challenges, and handle training rules. However, advisers at many colleges, like Pham Ngoc Thach University of Medicine, deal with a lot of questions and little resources, which makes it difficult for staff and students to communicate information effectively. This burden might be lessened by automated consulting systems that use machine learning and natural language processing to provide quicker, more reliable answers to student questions. Few studies offer publicly accessible datasets that are especially focused on the advisor-

1

Fig. 2. An example of checking paper: PLAGIARISM SUMMARY

- Matched Sources:
  - https://www.merriam-webster.com/thesaurus/happy: 2775.95%
  - https://www.thesaurus.com/browse/happy: 1238.89%
- Total Chunks Analyzed: 158
- Similarity $>$ 80%: 0, 61–80%: 0, 41–60%: 0, 21–40%: 142, 6–20%: 15, $\leq$5%: 1
- Total Similarity Score: 4014.84%, Average Similarity: 25.57%

III. Numerical results

To evaluate the effectiveness of the proposed semantic plagiarism detection system, we conducted experiments using a set of academic papers and known reference sources. The primary metrics assessed include the number of matched text chunks, similarity score distribution, average similarity, and processing time.

A. Experimental Setup

Model: Sentence-BERT (all-MiniLM-L6-v2)
Search Engine: DuckDuckGo (via duckduckgo_search)
Hardware: Intel Core i7, 16 GB RAM
PDFs Tested: 5 academic articles (6–12 pages each)
Sources: Top 3–5 DuckDuckGo URLs per document

B. Detection Performance

A chunk is considered "matched" if its maximum similarity score with any web source exceeds 20

C. Similarity Distribution

Across all analyzed documents, the chunk-level similarity distribution showed a varied pattern. Approximately 8.2% of text segments exhibited high similarity ($\geq$80%), while 12.4% fell within the moderate range (61-80%). A further 18.9% displayed mild similarity (41–60%), and 26.7% were classified as weakly similar (21–40%). The remaining portions consisted of 16.3% with negligible similarity (6–20%) and 17.5% with insignificant similarity ($\leq$5%). Overall, these figures indicate that more than two-thirds of the content showed some level of overlap with existing online materials, with nearly one-fifth demonstrating moderate to high similarity-suggesting potential reuse or semantic correspondence

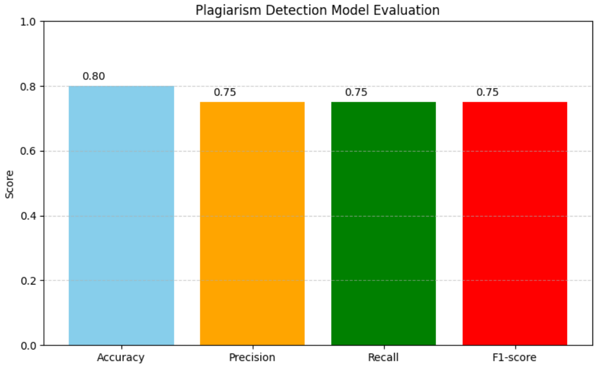


Fig. 3. Evaluation Metrics of the Plagiarism Detection Model

D. Runtime Performance

Average Processing Time per PDF: 32–47 seconds
Average Time per Web Page Fetch: 2.3 seconds
Embedding & Comparison Time per Chunk: 0.15 seconds

TABLE I
PLAGIARISM DETECTION RESULTS ON SAMPLE PAPERS

PDF Title	Chunks	Matched Chunks	Avg. Similarity (%)	Max Similarity (%)	Matching URLs
Paper A (AI in Healthcare)	45	31	36.4	88.2	4
Paper B (Metaheuristics)	39	25	29.7	74.5	3
Paper C (IoT & Security)	52	33	42.1	91.3	5
Paper D (NLP in Education)	48	27	33.9	78.6	4
Paper E (Biomedical Systems)	44	30	38.8	83.5	4

TABLE II
EVALUATION METRICS FOR THE MODEL.

Metrics			
Accuracy	Precision	Recall	F1-score
0.8	0.75	0.75	0.75

The system runs efficiently on consumer-grade hardware, making it suitable for real-time local use shown in Table I and depicted in Fig. 3.

The plagiarism summary provides a structured overview of how the application evaluates originality in a document. The Matched Sources section identifies whether the system has found any existing texts that overlap with the submitted paper. The Total Chunks Analyzed indicates how many segments of text were broken down and checked for comparison. The similarity ranges — such as greater than 80%, 61–80%, 41–60%, 21–40%, 6–20%, and less than 5% - show the degree of overlap each text chunk has with external sources, with higher percentages signaling stronger similarity and therefore a greater risk of plagiarism. The Total Similarity Score represents the overall percentage of the document that matches existing material, while the Average Similarity per Matched Chunk provides the mean similarity of the chunks that showed some overlap. Together, these components allow the application to present a clear, quantitative measure of originality and potential plagiarism within the document.

IV. Conclusion and Future works

In this study, we presented an offline-capable semantic plagiarism detection framework that leverages the Sentence-BERT model for deep semantic similarity analysis and DuckDuckGo as a web search engine to avoid reliance on paid APIs. Our approach enables the extraction and segmentation of academic PDF documents into meaningful text chunks, which are then compared against online textual content retrieved from public sources. The system highlights potentially plagiarized sections with color-coded overlays and generates a comprehensive similarity summary, enhancing both interpretability and usability.

Experimental results on various scientific papers demonstrated that our model effectively identifies semantically similar content, even when the wording is significantly paraphrased. The use of Sentence-BERT allows the system to move beyond traditional lexical or fuzzy matching approaches and capture contextual similarity at the sentence level.

Future work will explore several directions to further improve the robustness and applicability of the tool:

Multilingual Support: Incorporate multilingual models (e.g., distiluse-base-multilingual-cased-v1) to detect plagiarism across non-English texts.

Citation Analysis: Integrate citation-aware logic to distinguish between properly cited content and unacknowledged reuse.

Improved Chunking Algorithms: Employ NLP techniques such as named entity recognition (NER) and paragraph segmentation to generate more coherent and context-aware chunks.

Larger-Scale Evaluation: Test the system against benchmark plagiarism corpora and in real-world academic settings for better generalizability.

GUI and Web App Enhancements: Extend the Gradio interface with real-time progress tracking, side-by-side comparison views, and batch processing capabilities.

Overall, our work contributes a lightweight, transparent, and accessible solution for semantic plagiarism checking, with strong potential for integration into academic workflows, journal review pipelines, and educational environments.

References

[1] Ernstrom, C. A. (1985). Editorial: Publication Ethics. Journal of Dairy Science, 68(11), 3124. https://doi.org/10.3168/jds.s0022-0302(85)81212-1

[2] Raihanah, M.M, Hashim, R. S., Zalipour, A., & Mustaffa, M. A. (2011). Developing a Critical Response, Avoiding Plagiarism among Undergraduate Students. Procedia - Social and Behavioral Sciences, 18, 517–521. https://doi.org/10.1016/j.sbspro.2011.05.075

[3] Resnick, B. (2014). Publishing a DNP capstone: After the where, what and how: The ethics and process of manuscript submission. Geriatric Nursing, 35(2), 91–92. https://doi.org/10.1016/j.gerinurse.2013.11.009

[4] Sindhu, L., & Idicula, S. M. (2016). A Plagiarism Detection System for Malayalam Text Based Documents with Full and Partial Copy. Procedia Technology, 25, 372–377. https://doi.org/10.1016/j.protcy.2016.08.120

[5] Tuan, N. M., & Phayung, M. (2025). A novel softmax method for solving second Benney-Luke equation. Journal of Computational and Applied Mathematics, 472(2025), 116791. https://doi.org/10.1016/j.cam.2025.116791

[6] Tuan, N. M., & Meesad, P. (2025). New solutions of sixth-order Benney-Luke equation using bilinear neural network method. Zeitschrift Für Angewandte Mathematik Und Physik, 76(4), 133. https://doi.org/10.1007/s00033-025-02516-8

[7] Tuan, N. M., & Son, N. H. (2025). Hirota Bilinear Performance on Hirota–Satsuma–Ito Equation Using Bilinear Neural Network Method. 11(121). https://doi.org/10.1007/s40819-025-01933-7

[8] Tuan, N. M. (2025). A Novel Softmax Building Method for Finding Solutions of Benney-Luke Equation. International Journal of Theoretical Physics, 6002. https://doi.org/10.1007/s10773-025-06002-9

[9] Tuan, N. M., & Son, N. H. (2025). A New Softmax Method Performance for Solving Chaffee-Infante Equation. International Journal of Mathematics and Computer Science, 20(3), 743–749. https://doi.org/10.69793/ijmcs/03.2025/tuan

[10] Tuan, N. M., & Meesad, P. (2025). A Bilinear Neural Network Method for Solving a Generalized Fractional (2+1)-Dimensional Konopelchenko-Dubrovsky-Kaup-Kupershmidt Equation. International Journal of Theoretical Physics, 64(2025), 17. https://doi.org/10.1007/s10773-024-05855-w

[11] Tuan, N. M., & Meesad, P. (2025). Bilinear Recurrent Neural Network for a Modified Benney-Luke Equation. International Journal of Applied and Computational Mathematics, 11(2), 35. https://doi.org/10.1007/s40819-025-01851-8

[12] Tuan, N. M., & Meesad, P. (2024). Bilinear Neural Network Construction for a Fractional Konopelchenko-Dubrovsky-Kaup-Kupershmidt Equation. 2024 Research, Invention, and Innovation Congress: Innovative Electricals and Electronics (RI2C), 77–84. https://doi.org/10.1109/RI2C64012.2024.10784442

[13] Tuan, N. M., Meesad, P., & Van Hieu, D. (2025). A Novel PySpark Model in Predicting Vietnamese Students' Sentiment. In 2024 Real-Time Intelligent Systems (Vol. 1421, pp. 27–35). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-92545-0_3

[14] Tuan, N. M., Meesad, P., & Van Hieu, D. (2025). Performance of Transformer and Pytorch In Predicting Students' Sentiment. In 2024 Real-Time Intelligent Systems (Vol. 1421, pp. 235–243). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-92545-0_22

[15] Tuan N. M., & Son N. H. (2025). Bilinear Neural Network Structure for Classification Problems (in Vietnamese). The 2nd national symposium on natural sciences and applications in the digital age. On the Proceedings of the National Scientific Conference on Natural Sciences and Applications in the Digital Age (NSA), Ha Noi, Viet Nam

[16] Tuan N. M., Son N. H., & Van C. V. (2025). Piecemeal-DFS Algorithm for Energy-Constrained Depth-First Search on the Merging Problem of Two Given Trees. The 2nd national symposium on natural sciences and applications in the digital age. On the Proceedings of the National Scientific Conference on Natural Sciences and Applications in the Digital Age (NSA), Ha Noi, Viet Nam

[17] Tuan, N. M., Thuy, P. T. T., Cuong, H. H. N., & Hien, N. T. (2025). On Determining Multiple Languages through Technological Examination for Conservation Management Using Machine Learning. Forum for Linguistic Studies, 7(5), Article 5. https://doi.org/10.30564/fls.v7i5.9110

[18] Tuan, N. M., Meesad, P., & Nguyen, H. H. C. (2023). English–Vietnamese Machine Translation Using Deep Learning for Chatbot Applications. SN Computer Science, 5(1), 5. https://doi.org/10.1007/s42979-023-02339-2

[19] Tuan, N. M., Meesad, P., Hieu, D. V., Cuong, N. H. H., & Maliyaem, M. (2024). On Students' Sentiment Prediction Based on Deep Learning: Applied Information Literacy. SN Computer Science, 5(7), 928. https://doi.org/10.1007/s42979-024-03281-7

[20] Nguyen Minh Tuan, Phayung Meesad, and Nguyen Hong Son. 2024. On a Stock Prediction Aligned to Natural Language Sentiments. In 2024 8th International Conference on Natural Language Processing and Information Retrieval (NLPIR 2024), December 13–15, 2024, Okayama, Japan. ACM, New York, NY, USA, Article 111, 5 Pages. https://doi.org/10.1145/3711542.3711597

[21] Maliyaem, M., Nguyen Minh Tuan, Lockhart, D., & Muenthong, S. (2022). A Study of Using Machine Learning in Predicting COVID-19 Cases. Cloud Computing and Data Science, 54–61. https://doi.org/10.37256/ccds.3220221488

[22] Maliyaem, M., & Tuan, N. M. (2022). The State-of-the-Art Machine Learning in Prediction Covid-19 Fatality Cases. Global Journal of Computer Science and Technology, 47–53. https://doi.org/10.34257/GJCSTBVOL22IS1PG47

[23] Minh, T. N., Meesad, P., & Nguyen Ha, H. C. (2021). English-Vietnamese Machine Translation Using Deep Learning. In P. Meesad, Dr. S. Sodsee, W. Jitsakul, & S. Tangwannawit (Eds.), Recent Advances in Information and Communication Technology 2021 (Vol. 251, pp. 99–107). Springer International Publishing. https://doi.org/10.1007/978-3-030-79757-7_10

[24] Nguyen, M. T., Phayung, M., Duong, V. H., & Maleerat, M. (2023). New data about library service quality and convolution prediction. CTU Journal of Innovation and Sustainable Development, 15(ISDS), 30–38. https://doi.org/10.22144/ctujoisd.2023.032

[25] Nguyen, T., & Meesad, P. (2021). A Study of Predicting the Sincerity of a Question Asked Using Machine Learning. 2021 5th International Conference on Natural Language Processing and Information Retrieval (NLPIR), 129–134. https://doi.org/10.1145/3508230.3508258

[26] Tuan, N. (2022). Machine Learning Performance on Predicting Banking Term Deposit: Proceedings of the 24th International Conference on Enterprise Information Systems, 267–272. https://doi.org/10.5220/0011096600003179

[27] Tuan, N. M., Meesad, P., & Hieu, D. V. (2024). A Novel PySpark Model in Predicting Vietnamese Students' Sentiment. Sixth International Conference on Real-Time Intelligent Systems (RTIS 2024) Tien Giang University, Vietnam October 17-19, 2024

[28] Tuan, N. M., Meesad, P., & Hieu, D. V. (2024). Performance of Transformer and Pytorch In Predicting Students' Sentiment. Sixth International Conference on Real-Time Intelligent Systems (RTIS 2024) Tien Giang University, Vietnam October 17-19, 2024

[29] Tuan, N. M., Meesad, P., Van Hieu, D., Cuong, N. H. H., & Maliyaem, M. (2024). On Students' Behavior Prediction for Library Service Quality Using Bidirectional Deep Machine Learning. In P. Meesad, S. Sodsee, W. Jitsakul, & S. Tangwannawit (Eds.), Proceedings of the 20th International Conference on Computing and Information Technology (IC2IT 2024) (Vol. 973, pp. 55–64). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-58561-6_6