**Exploring Disease Similarity: A Comparative Analysis of NLP-based Similarity Metrics**

**Abstract**

This study highlights the importance of utilizing Natural Language Processing (NLP) and Machine Learning (ML) methodologies, particularly employing Jaccard similarity, Cosine similarity, text preprocessing, and sophisticated transformer models like BERT (Bidirectional Encoder Representations from Transformers) and RoBERTa (Robustly optimized BERT approach) to detect correlations among diseases. Through the examination of symptoms, treatments, risks, causes, and descriptions, the research aims to proactively prevent the onset of specific diseases. The integration of transformer models, such as BERT and RoBERTa, to grasp intricate semantic connections within the textual data, thereby fostering a more nuanced comprehension of disease interrelationships. This comprehensive strategy adds to our insight into disease associations and streamlines the creation of more efficient preventive measures and personalised healthcare strategies.

**Keywords:** Natural Language Processing(NLP), Machine Learning(ML), Disease Classification, Information Retrieval System, Jaccard Similarity, BERT, RoBERTa, Medical Information Extraction (IE), System Architecture.

1. **INTRODUCTION**

We present an innovative research initiative centered on the development of an NLP-driven system tailored for Disease Classification and Information Retrieval. This effort responds to the increasing demand for accessible and insightful healthcare information. Our project seeks to empower healthcare professionals and individuals in search of diverse health-related information. Situated at the crossroads of NLP and the dissemination of healthcare information, our solution acknowledges the pivotal role of precise and comprehensive disease knowledge in effective treatment and prevention. This is particularly crucial in the face of the escalating volume of medical literature and diverse information sources.

In contrast to traditional disease information repositories, our system excels in the classification of diseases and provides in-depth insights into causes, symptoms, risk factors, treatments, and preventive measures. This method enhances comprehension and addresses the challenges presented by the rapidly evolving landscape of healthcare information.

The recent strides in NLP have significantly influenced the healthcare sector, fostering innovative solutions to bridge gaps in the accessibility and availability of medical literature. Harnessing NLP's capacity to handle unstructured text data, our research project employs these techniques to categorize and retrieve information about various conditions. The ultimate objective is to establish a dynamic, user-friendly platform facilitating effortless access to healthcare insights.

1. **LITERATURE REVIEW**

This paper explores the recent shift from conventional Machine Learning to Deep Neural Networks in medical Information Extraction (IE) over the last three years. It emphasises the dominance of Deep Learning in Natural Language Processing for tasks like named entity identification and relation extraction in medical contexts. Adaptive learning approaches are recommended to address challenges such as limited data and the unique characteristics of medical language, with potential significant impacts on medical informatics [1].

A natural language processing (NLP) pipeline was created to extract clinical data from unstructured esophagogastroduodenoscopy (EGD) reports, focusing on 10 gastric diseases. Validation with 1000 reports showed high accuracy, with sensitivity, positive predictive value (PPV), and F1 score exceeding 0.96 for gastritis and 0.97 for ulcers and neoplastic diseases. Applying the pipeline to 248,966 reports spanning a decade revealed patient demographics, disease extent, and locations. This study demonstrates the NLP pipeline's potential for automated extraction of gastric disease information from EGD reports, enabling large-scale clinical research to enhance our understanding of gastric diseases [6].

The abstract emphasises the significance of precise prognostic staging in cancer prediction and therapy decisions. It emphasises the difficulty of unstructured medical records as well as the requirement for a standardised clinical decision stage technique. The literature supports the need for such a procedure, particularly in cases of breast cancer. This work used data from 465 patients in India to successfully extract critical prognostic features from various medical records using natural language processing, machine learning, and rule-based approaches. The study predicts prognostic phases with excellent accuracy in both rural and urban settings, demonstrating the possibility for improved breast cancer prediction. A generic staging system like this can considerably improve patient care and treatment decisions [3].

Because of the increased integration of electronic health records (EHRs), opportunities for automated healthcare systems and clinical research have emerged. Information Extraction (IE), which pulls clinical data from textual records, is a critical component enabling secondary EHR utilisation. This review of the literature looks at recent research on clinical IE applications. A thorough search yielded 263 articles published between January 2009 and September 2016. These publications are evaluated based on their publication sources, data origins, clinical IE tools, methodology, and applications in disease, drug research, and workflow advancements. Despite the numerous uses of clinical IE, there is a significant gap between EHR-based clinical studies and clinical IE research, showing that there is opportunity for development in bridging this gap and developing healthcare informatics [2].

This conference paper examines the critical function of Information Extraction (IE) in the medical and insurance industries when dealing with electronic medical record contents. It emphasises the ability of artificial intelligence to extract and use important information from these records, lowering labour costs and increasing efficiency. Currently, manual data extraction is the most used way. The report, however, highlights the growing interest in using Optical Character Recognition (OCR) and Natural Language Processing (NLP) technology to automate the procedure. To that end, the authors present the Medical OCR dataset (MedOCR) and host an evaluation competition as part of the eighth China Health Information Processing Conference (CHIP2022). The competition drew 18 teams, with OCR-based systems producing excellent results with a concentration on Acc as the assessment parameter. The report emphasises the importance of information extraction in the medical sector and serves as a useful resource for future research and development in this field [4].

This work addresses the difficult job of obtaining critical information from infectious illness cases in medical literature, which is critical for public health research. The study uses natural language processing (NLP) to mine clinical and social determinant data from published cases. The proposed framework combines data preparation, natural language processing for named entity recognition, and evaluation components, with an emphasis on COVID-19 case reports. When compared to benchmark approaches, the results show enhanced performance in named entity recognition and relation extraction. The work emphasises the utility of using transfer learning for future research in this domain and the potential for adapting this method to other infectious diseases [5].

This work investigates the use of natural language processing (NLP) to self-reported narratives of headache problem patients, with the goal of automatically classifying and extracting relevant information from clinical descriptions. The study identifies different word choices in narratives of migraine and cluster headache patients using lexical, semantic, and theme analysis, which aligns with expert knowledge of these conditions. In identifying headache attack descriptions, machine learning (ML) approaches such as logistic regression and support vector machines perform well. The work emphasises NLP's ability to detect key language elements in clinical narratives and the promise for future breakthroughs using larger datasets and neural NLP methods [7].

The recent call by the United Nations to address the issues faced by the 300 million people globally who live with rare diseases emphasises the crucial need for comprehensive epidemiological data. Existing methods for finding, extracting, and curating epidemiologic information (EI) for rare diseases are time-consuming and error-prone, limiting our understanding of these conditions. In response, this paper provides a fresh approach: the creation of EpiPipeline4RD, a deep learning-based pipeline for the extraction of rare disease epidemiology information. The study produces good precision, recall, and F1 scores using a curated corpus for Named Entity Recognition, illustrating the efficacy of an automated curation paradigm. The EpiPipeline4RD project has the potential to greatly increase public health support and research in rare illnesses [8].

The study's reported high accuracy in the validation phase with 1,000 reports raises concerns about its applicability to diverse datasets, medical institutions, and populations. There's a significant research gap in evaluating the NLP pipeline's performance on a more representative set of data to ensure its robustness. Additionally, the focus on a specific disease scope may limit understanding of its variability across a broader range of conditions. Neglecting the performance on rare diseases, potential data imbalances, temporal variability, and dependency on report quality further underscore the need for a comprehensive assessment. The study lacks explicit exploration of clinical relevance, interpretability, integration into existing systems, and ethical considerations, emphasizing the need to address practical and responsible application of the technology in healthcare.

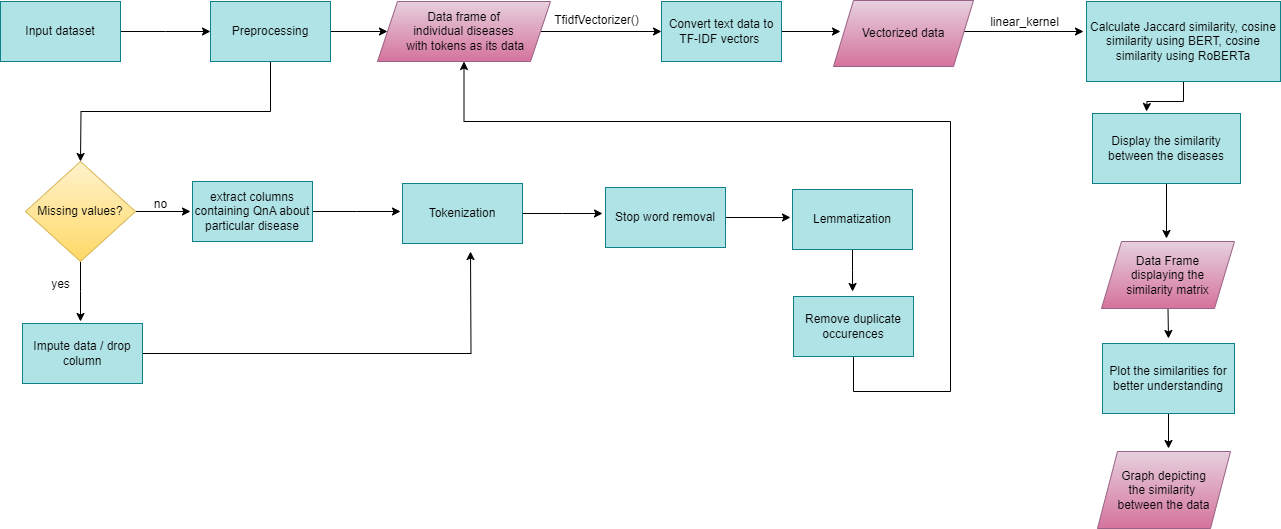
1. **RESEARCH METHODOLOGY**

**3.1 Dataset Description**

MedQuAD stands for “Medical Question Answering Dataset.” It contains 3 features: Question, Answer, and Source. It also has a target column which contains the focus area of every question-answer mentioned. This medical dataset has 47,457 instances which have been obtained from 12 NIH (National Institutes of Health) websites like GARD, niddk.nih.gov, cancer.gov, etc. This MedQuAD dataset answers a variety of questions that a user can have regarding the disease like the causes, the side-effects, treatments etc, and the necessary tests and medication required for the same [9].

**3.2 Methodology**

In this research, upon loading the dataset, our initial procedure involved implementing preprocessing techniques to identify and handle missing data. After successfully cleaning the dataset, the subsequent step focused on isolating information pertaining to questions and answers related to glaucoma. This extracted data was then utilised to construct a new DataFrame. Subsequently, a unified column was generated by combining the questions and answers, streamlining the tokenization process.

****

**Figure 1. Proposed system architecture**

To facilitate tokenization, the essential "punkt" tokenizer was downloaded. This tokenizer, designed for sentence tokenization, breaks down sentences into individual values. Following this, the tokenization process was applied, and a new column was created to store the resulting tokens. Further refinement involved the removal of stopwords from the obtained tokens. Subsequently, lemmatization was applied to reduce the tokens to their root forms, and duplicate entries were eliminated from the dataset.

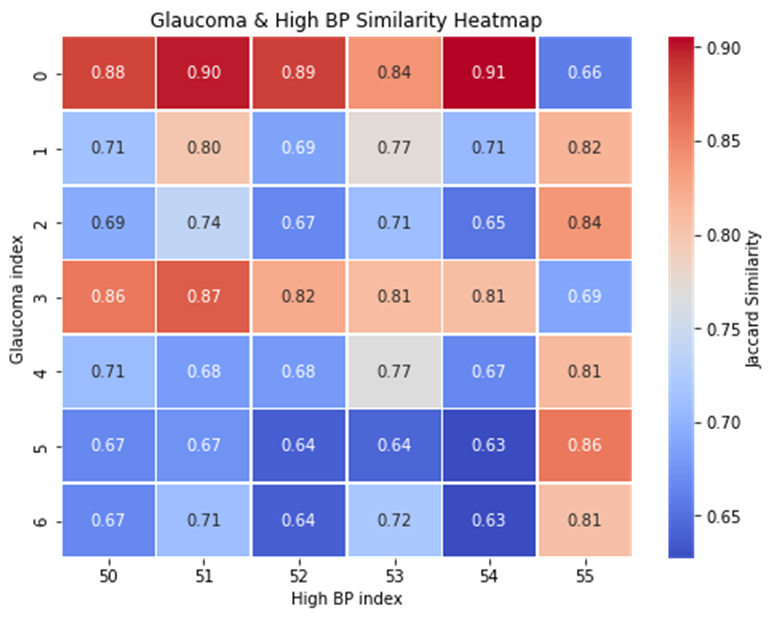
The lists of tokens were then converted into strings, and an additional column named "index" was introduced. The same sequence of steps was replicated for the "High BP" dataset, paving the way for the subsequent task of determining the similarity between glaucoma and High Blood Pressure (BP). Before conducting the similarity analysis, separate datasets were established, each containing lemmatized data for one of the two diseases.

In addition to employing traditional NLP techniques, advanced transformer models, namely BERT (Bidirectional Encoder Representations from Transformers) and RoBERTa (Robustly optimised BERT approach), were integrated into the tokenization process. These transformer models capture complex contextual information, providing a more sophisticated representation of the textual data. The contextual embeddings obtained from BERT and RoBERTa enriched the datasets, offering a semantic understanding of the underlying textual content.

The Jaccard similarity metric was employed to quantify the similarity between corresponding rows of the two disease-specific datasets, augmented by the advanced contextual features captured by BERT and RoBERTa. Following the computation of similarities, a heatmap visualisation was employed to provide a comprehensive depiction of the analogous rows, facilitating a clearer understanding of their relationships. The identical process was then repeated to ascertain the similarity between glaucoma and diabetes, leveraging the semantic insights provided by BERT and RoBERTa models in the analysis.

1. **RESULTS**

**4.1** **Similarity between glaucoma and High Blood Pressure**

****

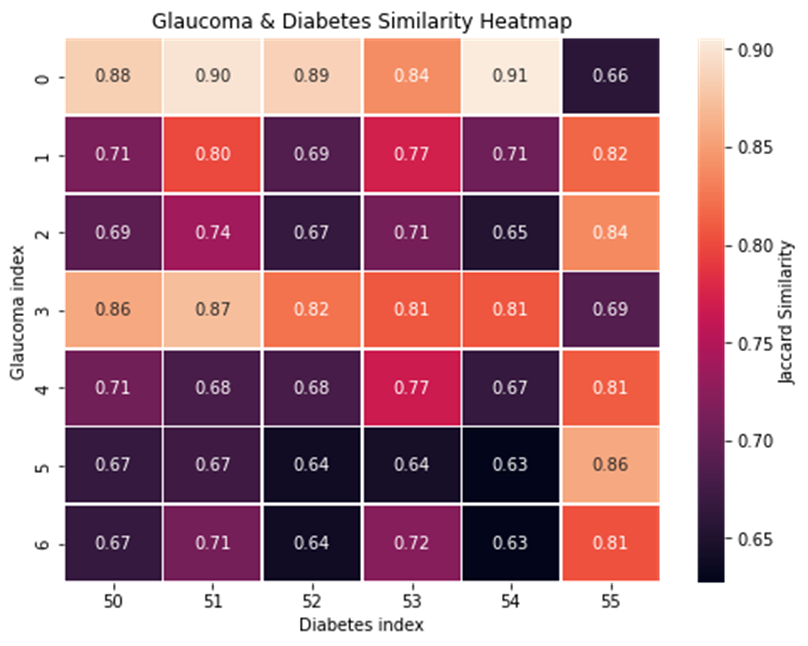
**Figure 2. Glaucoma & High BP similarity**

From Figure 2 we can see that the highly correlated rows are:

* row 0 (glaucoma) and row 13 (high bp)
* row 0 (glaucoma) and row 8 (high bp)
* row 0 (glaucoma) and row 10 (high bp)
* row 0 (glaucoma) and row 7 (high bp)
* row 0 (glaucoma) and row 9 (high bp)
* row 0 (glaucoma) and row 12 (high bp)
* row 3 (glaucoma) and row 13 (high bp)

Hence, we can say that High BP and glaucoma are highly correlated and similar to each other. A patient having high BP has a high chance of developing glaucoma.

**4.2** **Similarity between Glaucoma and Diabetes**

****

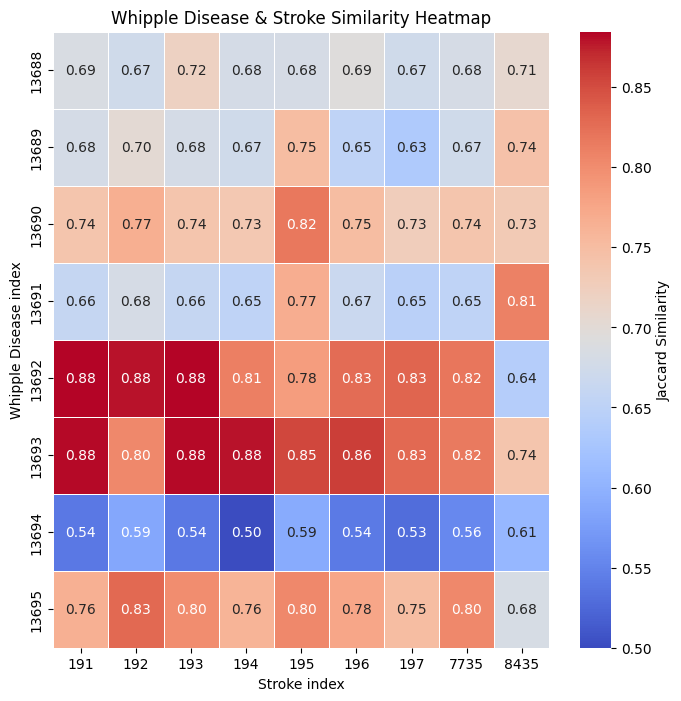
**Figure 3. Glaucoma & Diabetes similarity**

From **Figure 3** we can see that the highly correlated rows are:

* row 0 (glaucoma) and row 54 (diabetes)
* row 0 (glaucoma) and row 51 (diabetes)
* row 0 (glaucoma) and row 52 (diabetes)
* row 0 (glaucoma) and row 50 (diabetes)
* row 1 (glaucoma) and row 55 (diabetes)
* row 2 (glaucoma) and row 55 (diabetes)
* row 3 (glaucoma) and row 51 (diabetes)
* · row 5 (glaucoma) and row 55 (diabetes)

Hence, we can say that diabetes and glaucoma are highly correlated and similar to each other. A patient having diabetes can increase his/ her chance of developing glaucoma.

**4.3** **Similarity between Whipple’s Disease and Stroke**

****

**Figure 4. Whipple’s Disease & Stroke similarity**

From **Figure 4** we can see that the highly correlated rows are:

* row 13693 (whipple) and row 191 (stroke)
* row 13692 (whipple) and row 192 (stroke)
* row 13692 (whipple) and row 193 (stroke)
* row 13692 (whipple) and row 191 (stroke)
* row 13693 (whipple) and row 194 (stroke)

Hence, we can say that Stroke and Whipple's Disease are highly correlated and similar to each other. A patient having Whipple's disease has a high chance of getting a stroke.

| Similarity Method | Glaucoma\_High BP | Glaucoma\_Diabetes | Whipple\_Stroke |
| --- | --- | --- | --- |
| Jaccard Similarity | 73% | 75% | 72.5% |
| BERT | 94% | 89% | 96% |
| RoBERTa | 99% | 98% | 99% |

**Table 1. Similarity table**

1. **CONCLUSION**

Based on the findings of the study, it becomes evident that there exist correlations between different diseases. The presence of one particular ailment in a patient may increase the likelihood of developing another condition that shares a relationship with it. This investigation contributes valuable insights into the interconnected nature of diseases, allowing for a better comprehension of which ailments are interrelated. Such knowledge can prove instrumental in aiding patients to proactively prevent the onset of additional health conditions, thereby supporting efforts to maintain overall well-being.

1. **FUTURE SCOPE**

The future scope of this study involves expanding our analysis to encompass a broader spectrum of medical conditions. Incorporating additional diseases into the comparative framework could unveil intricate connections and correlations, providing a more comprehensive understanding of health-related textual data.

Furthermore, exploring advanced natural language processing (NLP) techniques and incorporating machine learning models may enhance the accuracy and depth of our similarity assessments. This could contribute to the development of more robust tools for medical data analysis and assist in the early detection or prediction of diseases based on textual patterns.

Additionally, extending the study to include larger and more diverse datasets could yield more representative insights, allowing for the identification of commonalities and differences across various demographic groups. Collaboration with medical experts and practitioners may also enhance the clinical relevance of our findings, paving the way for practical applications in healthcare decision-making and patient care.

**References**

[1] Hahn, Udo, and Michel Oleynik. "Medical information extraction in the age of deep learning." Yearbook of medical informatics 29.01 (2020): 208-220.

[2] Wang, Yanshan, et al. "Clinical information extraction applications: a literature review." Journal of biomedical informatics 77 (2018): 34-49.

[3] Deshmukh, P.R., Phalnikar, R. Information extraction for prognostic stage prediction from breast cancer medical records using NLP and ML. Med Biol Eng Comput 59, 1751–1772 (2021).<https://doi.org/10.1007/s11517-021-02399-7>

[4] Liu, L., Chang, D., Zhao, X., Guo, L., Chen, M., Tang, B. (2023). Information Extraction of Medical Materials: An Overview of the Track of Medical Materials MedOCR. In: Tang, B., et al. Health Information Processing. Evaluation Track Papers. CHIP 2022. Communications in Computer and Information Science, vol 1773. Springer, Singapore.<https://doi.org/10.1007/978-981-99-4826-0_13>

[5] Raza, S., Schwartz, B. Entity and relation extraction from clinical case reports of COVID-19: a natural language processing approach. BMC Med Inform Decis Mak 23, 20 (2023).<https://doi.org/10.1186/s12911-023-02117-3>

[6] Song, G.; Chung, S.J.; Seo, J.Y.; Yang, S.Y.; Jin, E.H.; Chung, G.E.; Shim, S.R.; Sa, S.; Hong, M.S.; Kim, K.H.; et al. Natural Language Processing for Information Extraction of Gastric Diseases and Its Application in Large-Scale Clinical Research. J. Clin. Med. 2022, 11, 2967.<https://doi.org/10.3390/jcm11112967>

[7] Vandenbussche, N., Van Hee, C., Hoste, V. et al. Using natural language processing to automatically classify written self-reported narratives by patients with migraine or cluster headache. J Headache Pain 23, 129 (2022).<https://doi.org/10.1186/s10194-022-01490-0>

[8] Kariampuzha, W.Z., Alyea, G., Qu, S. et al. Precision information extraction for rare disease epidemiology at scale. J Transl Med 21, 157 (2023).<https://doi.org/10.1186/s12967-023-04011-y>

[9]Soufyane, A., Abdelhakim, B. A., & Ahmed, M. B. (2021, January). An intelligent chatbot using NLP and TF-IDF algorithm for text understanding applied to the medical field. In *Emerging Trends in ICT for Sustainable Development: The Proceedings of NICE2020 International Conference* (pp. 3-10). Cham: Springer International Publishing.

[10]Szarvas, György, Richárd Farkas, and Róbert Busa-Fekete. "State-of-the-art anonymization of medical records using an iterative machine learning framework." *Journal of the American Medical Informatics Association* 14.5 (2007): 574-580.

[11]D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., ... & Sculley, D. (2022). Underspecification presents challenges for credibility in modern machine learning. *The Journal of Machine Learning Research*, *23*(1), 10237-10297.

[12]Harrison, C. J., & Sidey-Gibbons, C. J. (2021). Machine learning in medicine: a practical introduction to natural language processing. *BMC medical research methodology*, *21*(1), 1-11.

[13]Hashimoto, Daniel A., Guy Rosman, Daniela Rus, and Ozanan R. Meireles. "Artificial intelligence in surgery: promises and perils." *Annals of surgery* 268, no. 1 (2018): 70.

[14]Lund, B. D., Wang, T., Mannuru, N. R., Nie, B., Shimray, S., & Wang, Z. (2023). ChatGPT and a new academic reality: Artificial Intelligence‐written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology*, *74*(5), 570-581

[15]Pandey, S., & Pandey, S. K. (2019). Applying natural language processing capabilities in computerized textual analysis to measure organizational culture. *Organizational Research Methods*, *22*(3), 765-797.

[16]Bauer, M., Glenn, T., Monteith, S., Bauer, R., Whybrow, P. C., & Geddes, J. (2017). Ethical perspectives on recommending digital technology for patients with mental illness. *International journal of bipolar disorders*, *5*(1), 1-14.

[17]Reier Forradellas, Ricardo Francisco, and Luis Miguel Garay Gallastegui. "Digital transformation and artificial intelligence applied to business: Legal regulations, economic impact and perspective." *Laws* 10, no. 3 (2021): 70.

[18]Gurcan, Fatih, and Nergiz Ercil Cagiltay. "Research trends on distance learning: A text mining-based literature review from 2008 to 2018." *Interactive Learning Environments* 31.2 (2023): 1007-1028.

[19]Carrell, David S., et al. "Challenges in adapting existing clinical natural language processing systems to multiple, diverse health care settings." *Journal of the American Medical Informatics Association* 24.5 (2017): 986-991.

[20]Liu, Hui, Yaohua Tian, Yan Xu, Zhe Huang, Chao Huang, Yonghua Hu, and Jun Zhang. "Association between ambient air pollution and hospitalization for ischemic and hemorrhagic stroke in China: a multicity case-crossover study." *Environmental Pollution* 230 (2017): 234-241.