**Disease Classification and Information Retrieval System**

**Abstract**

This research underscores the significance of leveraging Natural Language Processing (NLP) and Machine Learning (ML) techniques, specifically Jaccard similarity, text preprocessing, as well as advanced transformer models like BERT (Bidirectional Encoder Representations from Transformers) and RoBERTa (Robustly optimised BERT approach), to identify similarities between diseases. By analysing symptoms, treatments, risks, causes, and descriptions, the study aims to proactively prevent the development of specific diseases. The incorporation of transformer models such as BERT and RoBERTa enhances the ability to capture intricate semantic relationships within the textual data, thereby providing a more nuanced understanding of disease interrelations. This comprehensive approach not only contributes to our understanding of disease connections but also facilitates the development of more effective preventive measures and personalised healthcare strategies.

**Project Objective**

Develop an NLP-driven system that classifies and provides detailed information about different diseases mentioned in the text data.

**Keywords**

Natural Language Processing(NLP),MachineLearning(ML),Disease Classification,Information Retrieval System, Jaccard Similarity, BERT, RoBERTa, Medical Information Extraction (IE), System Architecture

**1.Introduction**

We introduce a novel research initiative focused on developing an NLP-driven system for Disease Classification and Information Retrieval, addressing the growing need for accessible and insightful healthcare information. This project aims to empower healthcare professionals and individuals seeking information on diverse health issues. Positioned at the intersection of NLP and healthcare information distribution, our solution recognizes the crucial role of accurate and comprehensive disease knowledge in effective treatment and prevention, especially amid the escalating volume of medical literature and diverse information sources.

Unlike conventional disease information repositories, our system excels in classifying diseases and delivering detailed insights into causes, symptoms, risk factors, treatments, and preventive measures. This approach enhances understanding and addresses the challenges posed by the rapidly evolving landscape of healthcare information.

Recent advancements in NLP have significantly impacted the healthcare sector, facilitating innovative solutions to bridge gaps in medical literature accessibility and availability. Leveraging the capability of NLP to handle unstructured text data, our research project employs these techniques to categorise and retrieve information about various conditions. The ultimate goal is to create a dynamic, user-friendly platform that facilitates easy access to healthcare insights.

**2.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].

Gap Areas -1.**Limited Generalizability:**

* + The high accuracy reported in the validation phase with 1,000 reports might not necessarily generalize well to different datasets, medical institutions, or diverse populations. The performance of the NLP pipeline should be tested on a more diverse and representative set of data to ensure its robustness.

1. **Disease Scope and Variability:**
   * The performance of the NLP pipeline may vary when applied to a broader range of conditions or diseases not initially considered. Consideration of a wider disease spectrum would provide a more comprehensive evaluation of the pipeline's utility.
2. **Performance on Rare Conditions:**
   * The summary does not specify the performance of the NLP pipeline on rare or less common diseases. Assessing the pipeline's effectiveness in identifying and extracting information related to rare conditions is crucial for its clinical utility.
3. **Data Imbalance:**
   * The summary provides performance metrics. It's essential to examine the performance on less prevalent diseases to address potential imbalances in the dataset and ensure equitable representation of all diseases.
4. **Temporal Variability:**
   * The application of the pipeline to reports spanning a decade suggests temporal variability in reporting styles, medical practices, and diagnostic criteria. This variability could impact the pipeline's performance over time and needs to be carefully considered in the interpretation of results.
5. **Dependency on Report Quality:**
   * The accuracy of the NLP pipeline relies heavily on the quality and consistency of the input reports. In real-world scenarios, variations in report styles, terminology, and completeness may affect the pipeline's performance.
6. **Clinical Relevance and Decision Support:**
   * While the study highlights the potential for large-scale clinical research, the clinical relevance and impact of the extracted information on decision-making and patient outcomes are not explicitly addressed. Understanding how the extracted data can be practically used in a clinical setting is essential.
7. **Interpretability and Explainability:**
   * The summary does not discuss the interpretability and explainability of the NLP pipeline's decisions. Providing insights into how the pipeline arrives at its conclusions is crucial for gaining trust from healthcare professionals and ensuring the responsible use of the technology.
8. **Integration with Existing Systems:**
   * The practical integration of the NLP pipeline into existing healthcare information systems is not discussed. Considerations related to interoperability, user interface, and seamless integration with clinical workflows are crucial for real-world application.
9. **Ethical and Privacy Concerns:**
   * The extraction of patient demographics raises ethical and privacy concerns. The study should address how patient privacy is protected and how the extracted data can be used responsibly in compliance with healthcare regulations.

**3.Material and 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.

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.

**Jaccard similarity**:

Jaccard similarity, named after the French mathematician Paul Jaccard, is a measure of similarity between two sets. It is defined as the size of the intersection of the sets divided by the size of the union of the sets. The Jaccard similarity coefficient, often denoted as J(A, B), is calculated using the following formula:

J (A, B) = Size of Intersection of sets A and B (1)

Size of Union of sets A and B

In terms of set notation, if A and B are two sets, the Jaccard similarity can be expressed as:

J (A, B) = |A Ս B| (2)

|A Ո B|

The resulting coefficient ranges from 0 to 1, where:

- 0 indicates no similarity (no common elements between sets),

- 1 indicates complete similarity (sets are identical), and

- Values in between represent partial similarity.

Jaccard similarity is often used in various fields, including data mining, text analysis, and bioinformatics, to measure the similarity between two sets of data or documents. In text analysis, for example, it can be applied to compare the similarity of two documents based on the presence or absence of words or terms.

**BERT (Bidirectional Encoder Representations from Transformers) Using Cosine Similarity:**

BERT is a state-of-the-art natural language processing model that pre-trains on large amounts of unlabeled text and can be fine-tuned for specific tasks.Cosine Similarity is used to measure the similarity between two vectors.

BERT embeddings for text are generated, and these embeddings are treated as vectors in a high-dimensional space.

Cosine Similarity between the vectors of two pieces of text is calculated. The higher the cosine similarity, the more similar the texts.

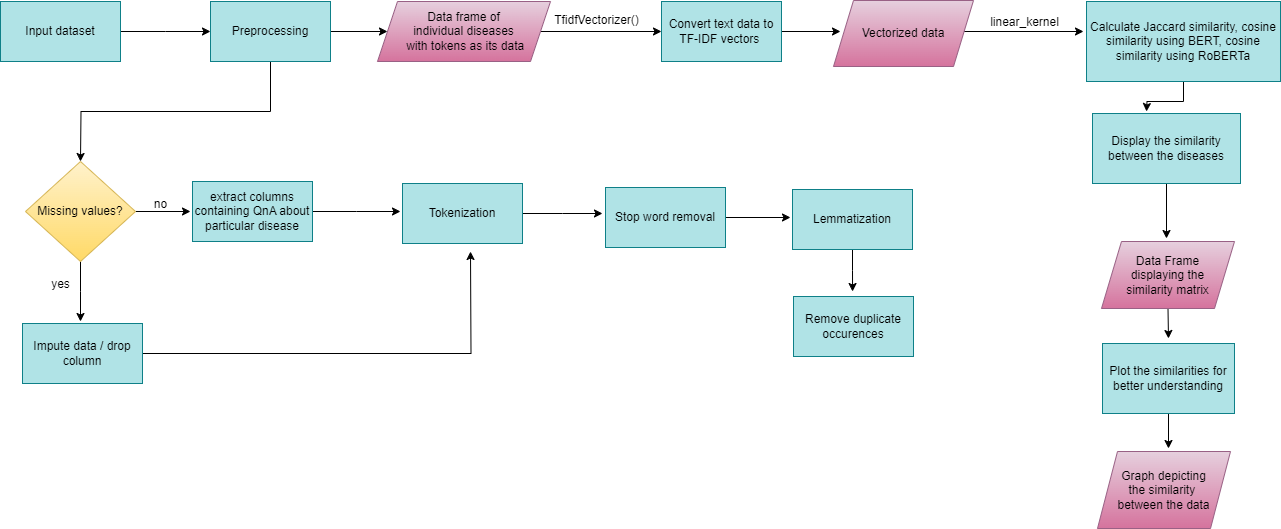
BERT with cosine similarity is commonly used for tasks such as semantic similarity, duplicate detection, and information retrieval.

**RoBERTa (Robustly optimized BERT approach) Using Cosine Similarity:**

RoBERTa is an improvement over BERT, designed to optimize pre-training and remove certain limitations. Similar to BERT, RoBERTa utilizes cosine similarity for measuring vector similarity. Roberta generates embeddings for text and represents them as vectors.

Cosine Similarity is applied to these vectors to quantify the similarity between two pieces of text. RoBERTa with cosine similarity is applied in various NLP tasks, including document retrieval, paraphrase detection, and sentiment analysis.

**4. System Architecture**

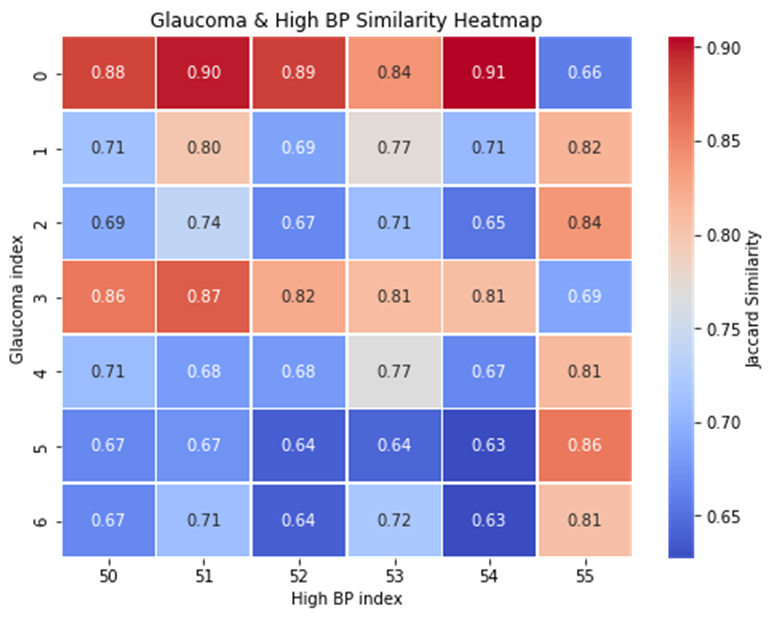


**Figure 1.** Proposed system architecture

**Figure 1** represents the working of our model in a diagrammatic format.

**5.Results**

5.1 **Similarity between glaucoma and High BP**

****

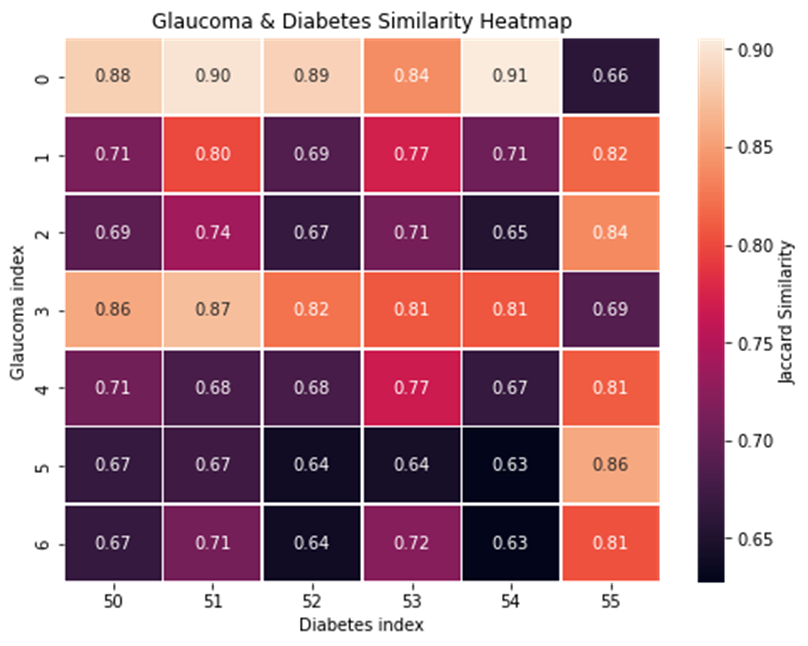
**Figure 2.** Glaucoma & High BP similarity

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.

5.2 **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.

**Table 1.** Disease Similarity Analysis

| Similarity Method | Glaucoma and High BP | Glaucoma and Diabetes |
| --- | --- | --- |
| Jaccard Similarity | 73% | 75% |
| BERT | 94% | 89% |
| RoBERTa | 99% | 98% |

**Table 1** suggests that the diseases Glaucoma and High BP, and Glaucoma and Diabetes are similar to each other. The Jaccard similarity, BERT and RoBERTa metrics are giving the same result i.e., similarity exists. It also shows that BERT and RoBERTa are giving higher similarity than Jaccard index.

**6.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.

**7.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.

**8.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]<https://www.kaggle.com/datasets/jpmiller/layoutlm>