NLP-Based Early Disease Diagnosis

Abdul Aziz Essam
Masters of Data Analytics
Western University
London, Canada
aessam@uwo.ca

Bassam Syed
Masters of Data Analytics
Western University
London, Canada
bsyed5@uwo.ca

Snehashis Padhi
Masters of Data Analytics
Western University
London, Canada
spadhi2@uwo.ca

Abstract

Advances in medicine have improved healthcare outcomes, yet early diagnosis remains a challenge for many, leading to severe complications. This research explores the use of Natural Language Processing (NLP) to enable accessible, data-driven, and automated early disease diagnosis tools. We preprocess data from extensive medical texts, apply topic modeling algorithms, and compare their effectiveness in identifying disease-related patterns. The findings highlight the potential for enhanced diagnostic accuracy and propose a framework for future development.

Our work involves meticulous preprocessing of unstructured medical data, including cleaning, tokenization, and normalization, to ensure high-quality input for the NLP models. We implement and evaluate multiple topic modeling approaches, such as Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF), to uncover latent patterns in disease-related texts. Comparative analysis reveals key differences in model performance, helping identify the most effective methods for capturing diagnostic cues. Furthermore, we delve into the interpretability of these models, enabling healthcare practitioners to better understand the generated outputs and make informed decisions. By combining insights from domain experts with model outputs, we create a synergistic approach to early diagnosis that bridges the gap between raw data and actionable medical insights.

Future work will focus on expanding datasets to include more diverse and multilingual medical texts, ensuring broader applicability of the models. Additionally, efforts will aim at validating the framework in clinical settings to refine its practical utility. Investigating the integration of real-time patient data and exploring transfer learning techniques to adapt models to rare diseases are also promising directions for future exploration.

I. Introduction

Early diagnosis of diseases is critical for improving patient outcomes and reducing healthcare costs. Despite technological advancements, many individuals fail to detect diseases early, leading to complications and increased mortality rates. This research addresses the gap in accessible tools for early disease detection by leveraging NLP techniques to analyze medical literature and identify patterns indicative of diseases.

Motivation: To create accessible, data-driven tools for early disease detection, empowering individuals to take proactive steps toward better health outcomes.

Objective: To explore NLP techniques and topic modeling algorithms to detect disease-related patterns from medical texts.

II. Methodology

This section provides a comprehensive, step-by-step explanation of the methodology employed in the study to achieve its objectives. The process is divided into three main phases: data extraction, preprocessing, and topic modeling. Each phase is elaborated in detail below.

Data Extraction

The dataset utilized in this study was derived from the Professional Guide to Diseases by Laura Willis[1]. This medical reference book spans 537 pages and covers a broad spectrum of medical conditions, including cardiovascular, neurological, gastrointestinal, renal, and urological disorders. To enable computational analysis, key sections of the text were converted into a machine-readable format using the PyPDF2 library.

Steps Involved:

A. Parsing PDF Files:

The PyPDF2 library was employed to extract text from the PDF version of the book.

Special care was taken to handle formatting inconsistencies and ensure accurate text retrieval.

B. Organizing Data:

Extracted data was segmented into disease-specific categories to streamline subsequent analyses.

Disease chapters and subheadings were used as guiding structures for categorization.

C. Annotating Key Elements:

Symptoms and critical terminologies were annotated to facilitate targeted exploration.

Manual verification was conducted to ensure annotations were accurate and clinically meaningful.

This phase established a well-organized and structured dataset that served as the foundation for the analytical processes that followed.

Data Preprocessing

To prepare the raw text data for analysis, a comprehensive preprocessing pipeline was implemented. This pipeline addressed common challenges in textual data processing and ensured the dataset's quality and usability. Below, we present a detailed breakdown of the techniques applied:

Techniques Applied:

Tokenization:

Text was divided into smaller units, or tokens (words or phrases), to allow for granular analysis.

Libraries such as NLTK and SpaCy were used to ensure efficient and accurate tokenization [2].

Lemmatization:

Words were reduced to their root or base forms, enhancing standardization across the dataset.

Examples: "running" was reduced to "run," and "better" was standardized to "good." [2]

Stopword Removal:

Common stopwords (e.g., "and," "the," "is") were removed to reduce noise. [2]

Domain-specific stopwords irrelevant to disease analysis (e.g., "patient," "hospital") were also excluded.

Dimensionality Reduction:

A Term Frequency-Inverse Document Frequency (TF-IDF) matrix was constructed to quantify the importance of terms in the dataset.

Singular Value Decomposition (SVD) was applied to reduce the matrix's dimensionality, retaining only the most significant features for analysis. [3]

Visualization:

Word frequency analysis was performed using WordCloud and Matplotlib. Two visualizations were created prior to preprocessing:

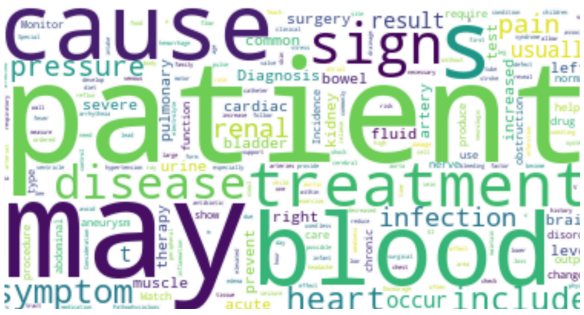

Figure 1: Word Cloud before pre-processing

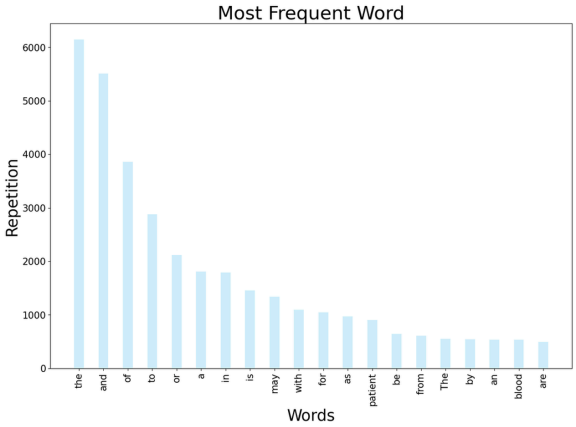

Figure 2: Frequency of top words before pre-processing

After preprocessing, updated visualizations were generated:

WordCloud After Preprocessing: Showcasing the dominant terms after applying tokenization, lemmatization, and stopword removal.

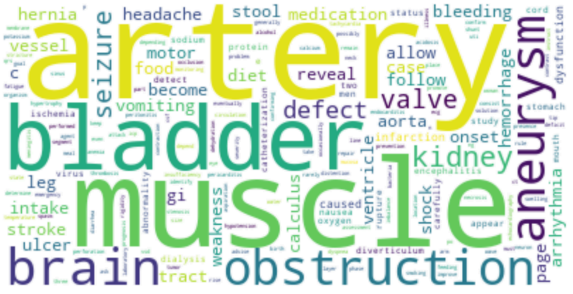

Figure 3: Word Cloud after pre-processing

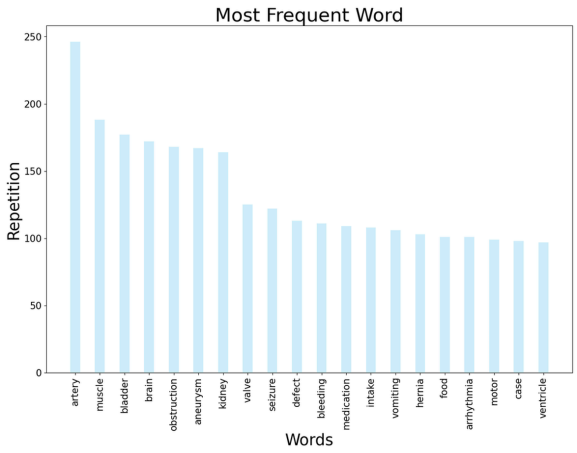

Figure 4: Frequency of top words before pre-processing

Most Frequent Words Plot After Preprocessing: Reflecting a cleaner and more relevant set of terms for analysis.

Topic Modeling

The topic modeling phase aimed to uncover latent patterns in the text data and group diseases based on shared symptomatology. Three advanced algorithms were employed: Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), and Correlation Explanation (CorEx). Each method was selected for its unique capabilities and contributions to text analysis. The methodology and results of each technique are detailed below.

1) Latent Semantic Analysis (LSA)

LSA is a dimensionality reduction technique that applies Singular Value Decomposition (SVD) to a term-document matrix to identify latent structures in the data. It is particularly effective in capturing synonymy (words with similar meanings) and polysemy (words with multiple meanings).[3]

Steps Involved:

1. Constructed a term-document matrix using Term Frequency-Inverse Document Frequency (TF-IDF).
2. Applied Truncated SVD to reduce dimensionality while retaining the most critical features.
3. Identified latent topics and mapped them to corresponding disease categories based on dominant terms.

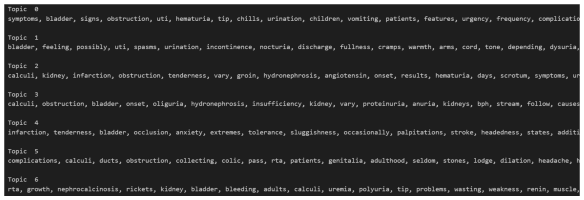

Figure 5: Top words for each topic using LSA

2) Latent Dirichlet Allocation (LDA)

LDA is a probabilistic topic modeling algorithm where each document is represented as a mixture of topics, and each topic is represented as a distribution over words. LDA excels in uncovering complex relationships between terms and assigning words to multiple topics when appropriate. [4]

Steps Involved:

1. Initialized the number of topics and set hyperparameters, such as alpha and beta, to control topic sparsity and word distribution.
2. Performed iterative updates using Gibbs sampling to refine topic distributions.
3. Evaluated the clinical interpretability and coherence of the extracted topics.

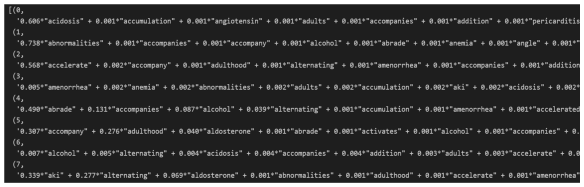

Figure 6: The word "acidosis" makes up 60.6% of the content in this topic, indicating it is the dominant term.

3) Correlation Explanation (CorEx)

CorEx is a deterministic clustering algorithm based on information theory, which identifies groups of terms that maximize total correlation (mutual information). Unlike LSA and LDA, CorEx does not rely on probabilistic assumptions, making it effective for smaller datasets and clear, interpretable results.[5]

Steps Involved:

1. Constructed a correlation graph of terms to quantify relationships.
2. Grouped terms into clusters using total correlation to maximize interpretability.
3. Validated clusters by comparing them to established clinical groupings of symptoms and diseases.

```
0: acidosis,confusion,atn,coma,purpura,overload,irritability,muscle,critically,pericarditis
1: apsgn,addition,anxiety,tolerance,stroke,states,sluggishness,extremes,headedness,palpitations
2: rta,accelerate,glomerulonephritis,nephrocalcinosis,accelerated,ankles,ascites,thromboembolic,atherosclerosis,malnutrition
3: scrotum,groin,walking,attempt,swelling,erythema,hydrocele,waddle,protect,urination
4: onset,kidneys,leukocytosis,speed,hypoalbuminemia,hyperlipidemia,thrombosis,enlarge,epigastric,proteinuria
5: changes,starting,span,signs,palpation,orifices,menses,membranes,mechanisms,males
6: atrophy,bradycardia,brain,chin,completely,hematuria,macrophage,potassium,rather,sclerotherapy
7: alcohol,trigger,temperature,stone,smelling,renin,rapidly,range,phosphorus,pallor
8: abdomen,tip,size,potter,parotitis,nocturia,midline,lines,incontinence,transilluminated
9: abrade,smell,pruritus,micturition,lowset,ileus,hydronephrosis,hair,gnawing,tactile
10: calcification,starts,spasms,rise,patterns,nose,motor,lobes,structures,lethargy
```

Figure 7: Results showing top words from the CorEx model

III. Results and Discussion

A. Overview of Results

The selection of the most effective topic modeling algorithm was guided by both qualitative assessments of model interpretability and a quantitative measure, cosine similarity. Cosine similarity evaluates the alignment between vectors in a high-dimensional space, quantifying the similarity between textual representations of disease symptoms and their respective topics [6]. Latent Semantic Analysis (LSA) was selected as the best-performing model due to its ability to handle large datasets and identify latent relationships effectively. All subsequent testing focused on LSA to validate its practical utility in disease diagnosis. Three test cases were employed to illustrate LSA's capabilities:

Test Case 1: Urinary Tract Infections

Symptoms such as frequent urination, painful urination (dysuria), and bladder tenderness were accurately grouped by LSA. The model effectively captured latent relationships between terms, enabling precise identification of relevant topics.

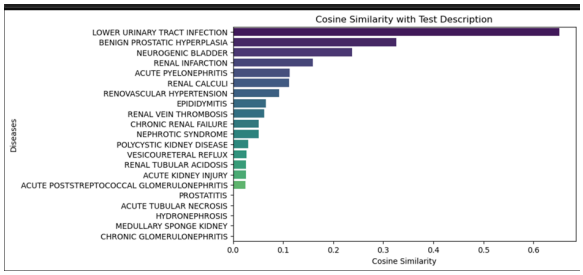

Figure 8: Cosine similarity for test case 1.

Test Case 2: Cardiovascular Disorders

Conditions such as ischemia, cyanosis, and hypotension were analyzed. LSA demonstrated strong interpretability by grouping related symptoms into cohesive topics. This highlighted its capacity to manage complex relationships within clinical data.

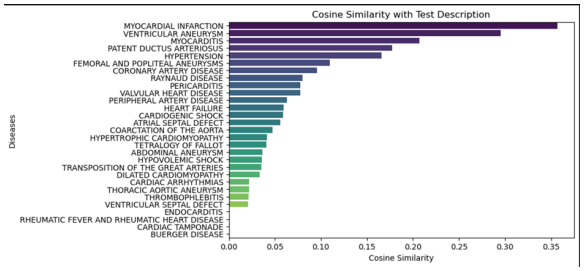

Figure 8: Cosine similarity for test case 2.

Test Case 3: Neurological Disorders

LSA successfully identified key terms associated with neurological conditions, such as spasms, lesions, and metabolic changes. The model's ability to uncover hidden patterns facilitated a deeper understanding of symptom clusters.

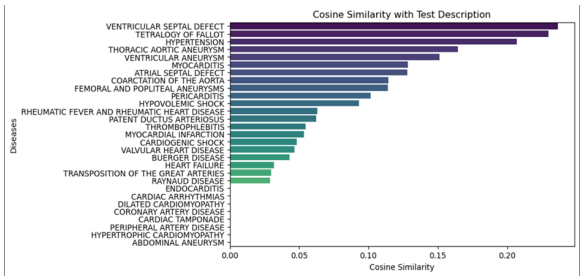

Figure 8: Cosine similarity for test case 3.

B. Comparative Analysis

The comparative performance of the three models across the test cases is summarized as follows:

Algorithm	Strengths	Weaknesses	Performance
CorEx	Captures mutual information between terms.	Lacks focus; poor grouping of symptoms.	Low
LDA	Probabilistic topic modeling; structured output.	Overlap of general terms; dominance of a few words.	Moderate
LSA	Captures latent semantics; coherent topics.	Some term overlap.	High

Table 1: Strengths and weaknesses of the three models

C. Discussion of Findings

The table summarizes the performance, strengths, and weaknesses of three topic modeling algorithms—Correlation Explanation (CorEx), Latent Dirichlet Allocation (LDA), and Latent Semantic

Analysis (LSA). A detailed comparison is presented below:

1) Correlation Explanation (CorEx)

- Strengths: CorEx captures mutual information between terms, making it effective at identifying meaningful relationships.
- Weaknesses: It lacks focus and struggles with grouping symptoms effectively, leading to poor clustering results.
- Performance: Due to its limitations in handling complex datasets, CorEx delivers a Low performance rating.

2) Latent Dirichlet Allocation (LDA)

- Strengths: LDA excels in probabilistic topic modeling, providing a structured and interpretable output. Its ability to represent documents as mixtures of topics makes it highly effective for nuanced text analysis.
- Weaknesses: LDA often exhibits an overlap of general terms, and its outputs can be dominated by a few frequently occurring words. Fine-tuning hyperparameters is essential to address these limitations.
- Performance: LDA achieves a Moderate performance rating, balancing interpretability and complexity.

3) Latent Semantic Analysis (LSA)

- Strengths: LSA effectively captures latent semantics and produces coherent topics by leveraging Singular Value Decomposition (SVD). It excels in identifying broad patterns and term relationships.
- Weaknesses: Some term overlap is observed, which can reduce the clarity of extracted topics. Additionally, LSA assumes linear relationships, limiting its performance in non-linear contexts.
- Performance: LSA outperforms the other algorithms with a High performance rating, making it a strong choice for discovering latent structures in large datasets.

IV. Future Directions

Further research will focus on the following directions:

1. Model Performance Optimization: Fine-tuning hyperparameters and exploring hybrid models that combine strengths of LSA, LDA, and CorEx to improve accuracy.
2. Automated Disease Classification: Building a classifier to map detected topics directly to disease categories for real-time diagnosis.
3. Expanding the Dataset: Incorporating data from additional medical literature, including journals, research papers, and clinical notes, to improve model robustness.
4. Cross-Domain Application: Exploring the use of topic modeling for diseases across different medical domains such as infectious diseases, oncology, and autoimmune disorders.
5. Visualization Enhancements: Developing interactive visualization tools to assist medical professionals in exploring and interpreting the results.
6. Validation in Clinical Settings: Conducting pilot studies with healthcare professionals to evaluate the system's performance and usability in real-world scenarios.

References

1. L. Willis, Professional Guide to Diseases. Wolters Kluwer, 2020.
2. T. Hofmann, "Probabilistic Latent Semantic Analysis," in Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, 1999.
3. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," in Journal of Machine Learning Research, vol. 3, pp. 993-1022, 2003.
4. G. AlSumait, D. Barbará, and C. Domeniconi, "On-line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking," in IEEE ICDM, 2008.
5. G. Ver Steeg and A. Galstyan, "Discovering Structure in High-Dimensional Data Through Correlation Explanation," in Advances in Neural Information Processing Systems, 2014.
6. A. R. Lahitani, A. E. Permanasari, and N. A. Setiawan, "Cosine Similarity to Determine Similarity Measure: Study Case in Online Essay Assessment," in *IEEE*, 2016.