ISSN: 2755-0109

Journal of Media & Management

SCIENTIFIC Research and Community

Review Article

Open Access

Multipage Medical Document Classification of Printed and Scanned Pages Using Machine Learning Algorithms

Shreedhar Deshmukh

Assist Professor, NSB Academy, Bangalore, India

ABSTRACT

The rapid rise of Artificial Intelligence (AI) and Machine Learning (ML) is redefining how healthcare organizations and BPO (Business Process Outsourcing) service providers operate. With increasing pressure to manage large volumes of patient documentation quickly and accurately—particularly for insurance and legal claims—many companies are turning to AI to streamline operations and enhance service delivery. A significant challenge in this space involves handling diverse medical records, which often include a combination of scanned handwritten notes, typed physician reports, lab test results, and radiology findings. These documents must be properly sorted, categorized, and chronologically arranged to reconstruct a clear patient history. Traditionally, this has been a time-consuming and error-prone manual process. Today, AI-driven solutions are stepping in to automate document classification using advanced natural language processing techniques such as TF-IDF, Count Vectorization, and word embeddings. These methods help machine learning models learn from historical data and accurately assign documents to relevant categories like diagnoses, prescriptions, and test results. We have tried to develop an application called auto index system, that classifies a medical document by analysing its content and categorizing it under predefined class (eg. Consultation, Anesthesia, Lab Report-Culture test, Radiology, office visit) topics and creating an index of pages falling under the category. We have plenty of classes to classify but fixed the scope to four-five predefined topics namely, Progress notes, consultation, CT, Lab reports, radiology. We use term frequency technology to convert and count number of words in the medical text documents and classify them based on weightage calculation.

***Corresponding author**

Shreedhar Deshmukh, Assist Professor, NSB Academy, Bangalore, India.

Received: June 03, 2025; **Accepted:** June 12, 2025; **Published:** June 26, 2025

Keywords: Text Mining, Medical Documents, Term Frequency, OCR, TFIDF Score, Multiclass

Introduction

Text mining, also referred to as text data mining, which is said to be text analytics, refers to the process of deriving high-quality information from text or collection of words. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. In the medical field, now a days in multi-speciality hospitals or the health care companies will have medical documents which are usually bunch of printed or hand written or scanned copies of patient history or consultation, diagnosed and prescribed documents. Used for different purposes like visit to next hospital or for claiming patient's insurance. The health care companies collect documents from patient as well as from hospital for processing the insurance. In the countries like US it is a regular process and the same context is followed now in India. The major problem for handling these documents is the order or the chronology of patient disease diagnosed and treated subsequently and indexing the documents digitally date wise. The patient may have visited the hospital many times for treatment and claiming for the treatment. In such a case classifying the document categorically and chronologically will be difficult manually. The use of Artificial intelligence and machine learning would help in text mining of medical documents which involves the process of converting the documents into optical character reader format and structuring the input text to find the related words from the document (OCR, derive features and also the removal of other unwanted words), and derive patterns within the structured data to match with the document and finally the evaluation, interpretation and classifying the document. The overall goal is to turn text into data for analysis, via application of natural language processing (NLP), machine learning and different analytical methods.

Literature Review Medical Text Classification

Medical data is heterogeneous and have different types and forms. The medical data contains the patient's sensitive data through the medical treatment process, biological data, genetic sequence data, image reports, pathology, cure plans, drug reports, and many other data types. These data are about the records in presence, but they can be used for the future by analyzing them. medical data can be classified into medical images, clinical notes, and some other types of data. Natural languages contain words, sentences, paragraphs, and each contains elements with explicit and implicit meanings. Understanding these is part of text classification. Moreover, this a legal problem because the formation of natural languages is different. The text's situation in the medical text is much complicated than the natural languages. X-rays, computed tomography (CT), magnetic resonance imaging (MRI), optical coherence tomography (OCT), microscopy, and positron emission tomography are medical image data (PET). Laboratory test results, physician diagnosis, medications, and therapies are included in the clinical text. Physiological measurements (laboratory outcomes,

vital signs), demographic details, payment, and insurance information are also included in other medical data [1].

Related work

Abdullah Muhammad Alghoson from Claremont Graduate University has proposed a method to classify the medical documents [2]. The classification was done using predefined terms from Medical Subject Headings (MESH). It used a corpus of 50 full-text journal articles (N=50) from MEDLINE, which were already indexed by experts based on MESH. Using natural language processing (NLP), the algorithm classifies the collected articles under MESH subject headings. The algorithm's outcome was evaluated by measuring its precision and recall of resulting subject headings from the algorithm, comparing results to the actual documents' subject headings. The algorithm classified the articles correctly under 45% to 60% of the actual subject headings and got 40% to 53% of the total subject headings correct. This holds promising solutions for the global health arena to index and classify medical documents expeditiously.

Problem Statement

In today's rapidly evolving business environment, Artificial Intelligence (AI) is revolutionizing industries across the board — and healthcare Business Process Outsourcing (BPO) is no exception. As healthcare organizations and legal firms increasingly rely on BPO partners for administrative and support services, the need for greater efficiency, speed, and accuracy has never been higher. One of the most pressing challenges faced by healthcare BPOs, especially those serving law firms involved in insurance claims, is the classification and indexing of vast volumes of complex medical documents. These documents often include scanned handwritten prescriptions, typed physician notes, lab reports, radiology images, and comprehensive medical histories all varying in format and quality. Traditionally, processing this information required significant manual effort, leading to delays, inconsistencies, and higher operational costs. However, the integration of AIpowered software solutions is now transforming this landscape. AI enables automated data extraction, intelligent document categorization, and rapid indexing — significantly reducing manual workload and human error. More importantly, it enhances accuracy in identifying critical medical information relevant to insurance cases, improving compliance and ensuring faster turnaround times [3-7].

Traditionally Healthcare BPO (Business Process Outsourcing) firms servicing for insurance companies and law attorneys relies heavily on manual tasks, including data entry, claims processing, medical coding, patient record management, and billing. These processes are time-consuming, error-prone, and require significant human effort.

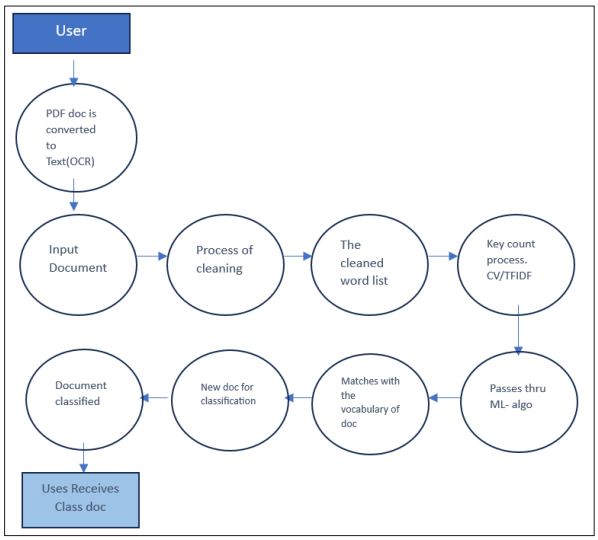
With the integration of Artificial Intelligence (AI), companies can automate and optimize these operations, leading to faster processing, reduced costs, and improved accuracy. These BPOs receive medical documents in PDF format, which must be manually reviewed and classified based on treatment details and indexed in a date-wise chronological order. This process is highly time-consuming, especially for employees who are doctors or healthcare professionals. Managing and organizing large patient records, often spanning 1,000 to 10,000 pages, becomes a significant challenge. AI can streamline this process by automatically classifying medical documents into predefined categories essential for insurance claims (e.g., Consultation, Lab Reports, Medical Bills, etc.). However, a key challenge arises when certain documents contain information about a patient's medical condition and require an explanatory note about what could be the cause of the decease or what he is suffering from.

Our Experiment

Model Design for Medical Document Classification

A Design flow diagram is used to convey the basic working of the application. Data flow diagrams can be divided into different levels in order to show the flow of the system at a low level, such as the inputs and the output at a higher level with most accurate results, with the underlying functions.

Model of Medical Document Classification



Steps

Step1: A medical document is input from the source PDF format converted to text document. This document is passed to the system to read. The program reads the document and creates a list of phrases to be sent to next step.

Step 2: process of cleaning, the document contains words, phrases and apart from this symbols, punctuations, date etc which may lead junk words collection and the result will be a wrong classification, so we need only those words which are related to the document which can make sure the document classified correctly.

Step 3: the cleaning process creates a list of words removing all unwanted characters from document and creates a list of complete words.

Step 4: The list of words is passed through another set of operation called Count vectorizer (CV) to count the number of words in the list of words and TF-IDF which is used to count the frequency of words and showing the weightage of word for the document. CV transfers the document into 1 and 0 bits and TFIDF converts this into frequency values which computer understands further.

Step 5: further this list is passed through the ML – algorithm, in our case we have used Logistic Regression and Decision Tree. It can be checked with many related algorithms. The algorithm uses CV or TF-IDF vectors and matches with the vocabulary values of previously created unique words.

Step 5: The ML algorithm gives weightage to the document's text so that it can be classified into any of the class given. It is a probabilistic method which gives percentage based on the document matching the words of vocabulary of given class and the higher the percentage is the chance of document classified in that category. This is assured by the metrics Precision and recall of ML algorithm. We always try to check for false positives than true positive because the medical document is a critical one which is if misclassified may lead to a critical problem.

Step 6: The user receives the class of the document he passed.

Result:

	Precision	Recall	F1-score	Support
Affidavit	1.00	0.96	0.98	46
Anesthesia_ Record	0.99	0.79	0.88	106
CT	0.94	0.99	0.96	423
Consent	0.93	0.96	0.95	269
Culture_and_ Sensitivity_Test	1.00	0.94	0.97	94
Accuracy	0.95	938		
Macro avg	0.97	0.93	0.95	938
Weighted avg	0.95	0.95	0.95	938

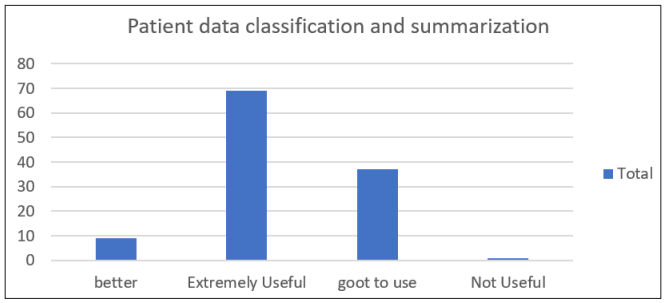
The above result shows the documents classified and the accuracy of classification can be seen in every class with precision and recall values defining the more accurate classification of the documents.

Creating Vocabulary

Before starting the above process of document classification, another process is followed with all the classes. The folders containing all the text files of different classes separated manually by checking each document manually. These documents are passed to the program which reads all the files from each folder and cleans the documents and creates a list of words in every class. This class is further sent to machine learning (ML) algorithm to get the features of every class which gives a collection of prominent words which can assure the class for any document. Here we have used Chi-Square test to get the features of the class and create a vocabulary. This vocabulary is further used in the process of document classification after the document is passed through the ML – Algorithm [8-11].

Further the Auto index system is created by reading the pdf pages and extracting the date from every heading page of the document and the first and last page of every document is identified by program code. Further, with all these elements (Date, first page, last page, page numbers, classified document) an auto index excel file is created, which gives a complete detail of "Date, first page, last page, page numbers, classified document".

Table-graph showing results of feedback of AI software used for patient data classification and summarization.



Conclusion

To accelerate the processing of medical documents, including screening, classification, and chronological arrangement, healthcare BPO service providers are leveraging technology to reduce workload and improve efficiency and accuracy. This adoption not only enhances their operational speed but also strengthens their market position, driving business growth, revenue generation, and overall economic development. Additionally, integrating AI and machine learning enables these companies to deliver more reliable results, stay competitive, and ensure long-term sustainability in the industry.

This study is to help healthcare BPOs to adapt new technology which would help the operation people to process fast and give a more accurate results to the stakeholders thereby generating more revenue which would add to their economy.

Our model of classifying documents and process of automating and creating an index helps the healthcare providers to go through on which date the patient admitted and what treatment and procedure was done and also the healthcare insurance industries will get the accurate data for their further process for claim of health insurance.

References

1. Guo F, Wu T, Jin X (2020) An Efficient Method Based on Region-adjacent Embedding for Text Classification of Chinese Electronic Medical Records. Proceedings - 2020 5th International Conference on Computational Intelligence and Applications, ICCIA 183-187.
2. Lee CH, Wang C, Fan X, Li F, Chen CH (2023) Artificial intelligence-enabled digital transformation in elderly healthcare field: Scoping review Adv Eng Inform 55: 1474-0346.
3. Obaido G, Mienye ID, Egbelowo OF, Emmanuel ID, Ogunleye A, et al. (2024) Supervised machine learning in drug discovery and development: Algorithms, applications, challenges, and prospects Mach Learn Appl 17: 2666-8270,
4. Lesley U, Kuratomi Hernández A (2024) Improving XAI explanations for clinical decision-making – Physicians' perspective on local explanations in healthcare Lecture notes in computer science, 0302-9743, 9783031665349, Springer Nature Switzerland 296-312.
5. Khalifa M, Albadawy M (2024) AI in diagnostic imaging: Revolutionising accuracy and efficiency Comput Methods Programs Biomed Update 5: 2666-9900
6. Lu SC, Swisher CL, Chung C, Jaffray D, Sidey Gibbons C (2023) On the importance of interpretable machine learning predictions to inform clinical decision making in oncology Front Oncol 10.3389/fonc.2023.1129380Google Scholar.
7. Hulsen T (2023) Explainable artificial intelligence (XAI): Concepts and challenges in healthcare AI 4: 652-666.
8. Heath Goodrum, Kirk Roberts, Elmer V (2020) Bernstam Automatic classification of scanned electronic health record documents https://www.sciencedirect.com/science/article/abs/pii/S1386505620309977, https://www.researchgate.net/publication/387290255_Medical_Document_Classification_using_NLP_Techniques.
9. Al Doulat A, Obaidat I, Lee M (2019) Unstructured medical text classification using linguistic analysis: A supervised deep learning approach. Proceedings of IEEE/ACS International Conference on Computer Systems and Applications, AICCSA, 2019-Novem 1-7.
10. Aldhoayan M, Zhou L (2016) An accurate and customizable text classification algorithm: Two applications in healthcare. 2016 IEEE 6th International Conference on Computational Advances in Bio and Medical Sciences, ICCABS 2016. https://ieeexplore.ieee.org/document/7802778/metrics#metrics.
11. Hughes M, Li I, Kotoulas S, Suzumura T (2017) Medical Text Classification Using Convolutional Neural Networks. Studies in Health Technology and Informatics 235: 246-250.