Chapter 9
Information Retrieval Systems in Healthcare: Understanding Medical Data Through Text Analysis

Srijit Panja
Saras Analytics, India

ABSTRACT

Healthcare systems generate an immense volume of data, which presents a unique opportunity to make a significant impact. This chapter examines the role of information retrieval systems in healthcare, specifically focusing on how text analysis can be utilized to enhance the understanding of medical data. By employing advanced text mining tools, this chapter demonstrates how we can extract valuable insights from these complex documents. It also presents text analytics as a solution-oriented approach, particularly beneficial in managing crises within healthcare systems and in making informed decisions based on accurate data analysis. The technical foundation of the study is rooted in the fields of natural language processing and artificial intelligence, with a focus on methodologies related to the semantics of words and text (e.g., text corpus, dictionaries, and text embeddings). Through this exploration, the chapter aims to highlight the transformative potential of text analysis and information retrieval systems in revolutionizing healthcare data understanding.

INTRODUCTION

The recent COVID-19 pandemic has catalyzed extensive scientific efforts to develop vaccines and effective management strategies (Bansal et al., 2020; Vaishya et al., 2020; Wang et al., 2021). This urgency has led to an exponential increase in both scientific and non-scientific literature over the past year, encompassing a broad range of research papers, documents, news articles, and fact-checks (Arayata et al., 2022; Cortez et al., 2022; Garcia et al., 2022; Parel et al., 2022). These constitute a substantial corpus of textual data accumulated during the crisis. To navigate this information deluge, information retrieval systems (IRS),

DOI: 10.4018/979-8-3693-3661-8.ch009

particularly search and query-processing engines, play a crucial role. They employ inverted indexing schemes and various weighting approaches to extract the most relevant data in response to specific queries. The retrieved documents then undergo further analysis to extract insights. Sentiment analyzer tools, for instance, identify major and minor opinions related to a query, revealing subjects closely aligned with it. Additionally, document summarizers are employed to condense extensive scientific findings into concise, critical points. This summarization process can also integrate multiple documents, allowing for the logical inference of potentially unrelated or independent research, thereby uncovering novel solutions. Additionally, keyword analysis facilitates topic modeling, where the frequency of certain words within a research-oriented corpus helps identify the most prevalent areas of study. The popularity of research topics and related areas can also be assessed by the IRS based on their usage by a broader user base. Creating a parallel database that tracks document popularity based on search engine appearances further aids in understanding research trends. Moreover, the less explored areas of research can offer untapped perspectives, which can be revealed through text analysis tools. Belief Index Generators, for instance, play a pivotal role in discerning the veracity of information by quantifying the trustworthiness of various articles. This level of text analysis not only contributes to resolving the primary crisis but also aids in comprehending the available information, deriving hidden insights, and mitigating panic by verifying the accuracy of disseminated information.

MAIN FOCUS OF THE CHAPTER

This chapter aims to illuminate the potential of advanced text retrieval and analysis mechanisms in healthcare. A key challenge in this domain is ensuring easy and streamlined access to various healthcare-related text resources, such as insurance documents, hospital terms and conditions, health reports, and electronic medical records. By employing text mining tools, not only can we extract insights from these complex documents, but we can also simplify their comprehension. This chapter delves into tools like document summarizers, word maps, and search engines, elaborating on their suitability for these purposes. Traditionally, text analytic tools have been viewed as a component of broader data analysis efforts. This chapter, however, seeks to emphasize the direct application of these tools in healthcare, highlighting their practical utility in this specific context. It presents text analytics as a solution-oriented approach, particularly useful in managing crises within healthcare. For instance, while the use of text search tools to determine popularity trends is common, this chapter introduces their application in identifying under-researched articles to uncover fresh perspectives or discover new solutions. The technical foundation of this study lies within the fields of natural language processing (NLP) in artificial intelligence (AI). It predominantly focuses on methodologies related to the semantics of words and text, encompassing aspects such as text corpus and dictionaries. A significant concept discussed throughout the chapter is text embeddings, which involve encoding text in a machine-understandable format. These embeddings vary, with some based on word frequency, while others capture meanings and contexts. In the context of healthcare data, both approaches are shown to be effective.

The implementation of the studies discussed in this chapter has been exclusively conducted using the Python programming language. The chapter explores both machine learning and deep learning approaches as integral parts of the study (Kazi, 2024; Patibandla et al., 2024). However, it is important to note that the chapter's scope is confined to the possibilities offered by text-based technologies in healthcare. It does not aim to provide a comprehensive guide to the digitization of healthcare but rather focuses on the