Journal of Basic Physical Research ISSN: 2141-8403 Prints, 2141-8411 Online
Vol.14, Issue 2, October, 2025, pp. 1 – 9, Available online at https://journals.unizik.edu.ng/jbpr/

DEVELOPMENT OF A TEXT MINING SYSTEM FOR SUMMARISING SICKLE CELL DISEASE RESEARCH IN NIGERIA

Joshua Arome Ayegba[1], Olumide Owolabi[2], Hashim Ibrahim Bisallah[3], Muhammad Sanusi[4]
[1&4]Department of Computer Science, University of Abuja, FCT-Abuja, Nigeria.
Corresponding Email: joshua.ayegba2019@uniabuja.edu.ng

Abstract
Information about diseased and healthy individuals is readily available online. Mining such biomedical content provides valuable insights into patients' conditions and research trends. However, this content is scattered, and manually gathering relevant data using traditional search engines is both laborious and incomplete. This paper reviews an ongoing study that aims to develop a text mining system for summarising and predicting research findings on Sickle Cell Disease (SCD) in Nigeria. The system includes a focused web crawler equipped with natural language processing (NLP) tools to extract and summarise abstracts from relevant biomedical databases like PubMed and BMJ Journals. Designed with Python, the crawler effectively mimics search functions to systematically gather relevant data. Results show that the crawler successfully scraped structured data, including article titles and abstracts, using targeted keywords such as "sickcell Nigeria" and "sick cell Nigeria." The next step involves integrating NLP-based summarisation techniques to forecast research trends. This study advances biomedical text mining and provides a scalable solution for automating knowledge extraction in SCD research across Nigeria. Ongoing improvements aim to enhance the crawler's robustness, enabling bulk page crawling and expansion to additional databases.
***Keywords**: Text Mining, Natural Language Processing (NLP), Sickle Cell Disease (SCD), Biomedical Literature, Web Crawler, Machine Learning*

Introduction

The enormous volume of biomedical textual information that is being generated on a daily basis has been a challenge that has prompted researchers to develop domain-specific text processing systems. Extensive research has been carried out on automatic text summarisation methods in recent times (Moradi and Ghadiri, 2017). These studies have been conducted in order to equip physicians, researchers, and biomedical users in the biomedical field with the needed tools to help them cope with vast volumes of information concealed in textual data (Moradi and Ghadiri, 2017).

Many automatic methods have been developed in recent decades to address the challenges of utilising text materials for information extraction and knowledge-finding activities. The methodologies have resulted in significant advancements in sectors as diverse as gene and genome expression, therapeutic target identification, drug repositioning, identifying adverse events, and developing domain-specific databases (Fleuren and Alkema, 2015). In the development of automatic text processing technologies, text mining and natural language processing approaches are crucial. Automatic text summarisation is a potential method for extracting and managing valuable information from huge and lengthy text sources.

Literature Review

The biomedical field generates an overwhelming volume of textual data daily, making it difficult for researchers to stay current. Consequently, automated text mining and summarisation techniques have become essential tools for knowledge extraction. According to Gong (2018), the exponential growth of databases like PubMed and Medline has made manual literature reviews inefficient, necessitating automation.

Text mining has been effectively used in various biomedical tasks such as summarising clinical notes, identifying therapeutic targets, supporting clinical decision-making, and managing

electronic health records (Mishra et al., 2014; Afantenos et al., 2014). NLP and machine learning techniques have greatly improved these applications. For instance, Lu et al. (2017) used Hidden Markov Models with F1 scores over 95%, while Weng et al. (2017) implemented convolutional recurrent neural networks for classifying clinical notes with impressive results. Likewise, Basaldella et al. (2017) created a hybrid named entity recognition model that reached 86% accuracy.

Given the diversity of text formats in biomedical data, Shuai et al. (2017) proposed a framework combining online machine learning and controlled vocabularies to extract valuable information from heterogeneous clinical reports. These advancements highlight the growing role of NLP in converting unstructured medical data into actionable insights.

The use of web crawlers has also emerged as a method to automate data acquisition from biomedical repositories. As Ganz and Reinsel (2009) observed, the digital data explosion necessitates scalable tools for data extraction and organisation. Focused crawlers are particularly useful in mining domain-specific datasets efficiently.

In Nigeria, there remains a gap in localised biomedical text mining tools tailored to prevalent diseases such as Sickle Cell Disease (SCD). This study addresses that gap by developing a custom text mining system for summarising Nigerian SCD research literature.

Recent Advances in Text Mining Systems for Biomedical Research

Text Mining

Many text summarising approaches have been presented in the biomedical domain to address distinct issues connected to various types of text documents (Mishra et al., 2014; Afantenos et al., 2014). The summarising approaches are used to solve problems in a wide range of biomedical subfields. Various uses of text summarising in the biomedical sector include summarisation of biomedical literature, summarisation of treatments, evidence-based medical care, summarisation of medication information, clinical decision assistance, summarisation of clinical notes, and electronic health records. Thus, utilising summarising approaches to solve problems relating to SCD in developing countries is beneficial.

Biomedical Text Mining

The ever-increasing amount of biomedical information being published daily has given rise to the need for biomedical text mining. This has also served as a motivation for researchers to keep utilising various methods in the text mining field. According to Gong (2018), at this pace of publishing, the increase of PubMed/Medline literature is exponential. As a result, keeping up with significant articles in their own area, much less allied disciplines, is very challenging for researchers.

Biomedical data is generated on a large scale and evolves rapidly, encompassing many discoveries and pieces of knowledge. These are often published and stored in biomedical databases for easy access. Focusing on this extensive published literature enables the mining of hidden biomedical information. Gong (2018) suggests that even in an era of experts, these massive volumes of biomedical literature may still rely on manual methods to fully understand how research is conducted in the text mining field. This approach can assist in gathering relevant biomedical data and has contributed to the idea that extracting biomedical information through text mining is crucial. In the biomedical field, text mining is now viewed as a new discipline that combines medical information science, bioinformatics, and other related areas.

Natural Language Processing

The field of artificial intelligence is a strong and growing area that includes natural language processing (NLP). The combination of natural language and computers has led to rapid expansion in the NLP field. The link between machine learning and NLP is also quickly strengthening, thanks to the field's growth. Natural language processing problems have been tackled with various machine learning techniques. Extracting relevant data and knowledge from a large stream of medical reports is challenging because of the complexity. However, using machine learning methods like HMM models, Lu, Ghasemzadeh, Hayek, Quyyumi, and Wang (2017) achieved results with F1 scores over 95%. Weng, Wagholikar, McCray, Szolovits, and Chueh (2017) also used machine learning in their work, which involved classifying clinical notes in the medical subdomain. The best-performing classifier was a convolutional recurrent neural network with word embeddings, according to their findings. This was tested on the iDASH and MGH datasets, with F1 scores of 0.845 and 0.870. For the named entity identification task, Basaldella, Furrer, Tasso, and Rinaldi (2017) developed a hybrid method that combined a dictionary approach with a machine learning classifier in a two-stage pipeline. They achieved an overall accuracy of 86 percent, with a recall of 60 percent.

Web Crawler

Data as the main propeller for the creation of web crawlers has attained steady growth and clustering. And as the digital universe is expanding, one of the major issues on the internet is the staggering amount of data that is available. In 2008, the amount of data on the internet was estimated as 487 exabytes, and 5 times that much was expected in 2009 (Ganz and Reinsel, 2009). The drivers of this explosion are coming largely from the vast number of user interactions, the growth of non-traditional devices, mobile internet usage, and growing data processing on servers (Ganz and Reinsel, 2009). The challenges accompanying the expanding digital universe are numerous. For instance, the speed and reliability with which it is up and downloaded, organising and searching, responsibility or compliance regarding the integrity of all this information have become ever more important. The perspective of a user is also essential to create valuable information from the oceans of data. Given the fact that most of it is unstructured, various techniques can enable decision makers to search, identify, structure, and analyse it. This makes web crawlers a significant engine for acquiring and structuring data from the World Wide Web.

Theoretical Framework

The paper will be based on information retrieval, natural language processing, and machine learning theories. The theoretical field of text mining is based on computational linguistics and data science. The backdrop on which the framework is based is that a large volume of biomedical data holds hidden patterns and knowledge that can be repeatedly and automatically extracted through established algorithms and learning models. The research uses the principle of focused crawling and semantic analysis to find and summarise literature belonging to a given domain.

Empirical Framework

This study improves existing empirical studies that apply machine learning in clinical text classification and named entity recognition. As can be seen, the works of Weng et al. (2017) and Basaldella et al. (2017) are employed as a comparative paradigm. The works under discussion emphasise the usefulness of the hybrid methods and convolutional neural networks in a successful data mining of clinical texts.

Materials and Methods

The ongoing study seeks to create a text mining system for summarising SCD research in Nigeria and then utilising Natural Language to identify key findings. This paper breaks down the methodology utilised into two stages (web crawler creation stage and natural language processing stage) for clarity.
The methodology is divided into two main stages:

Stage 1: Web Crawler Creation

A focused web crawler was developed using Python. The crawler simulates user search actions on PubMed and BMJ Journals. Using keywords such as "sickcell Nigeria," the crawler systematically traverses and collects biomedical article abstracts and titles. The output is structured into dictionaries with fields for article name and abstract.

Stage 2: Natural Language Processing (Upcoming)

The second phase, which will be implemented in the future, is NLP, which includes using topic modelling, named entity recognition, and extractive summarisation. It will be used in the analysis of the data that was gathered and will be used to come up with short, organised summaries, as well as to get the important details. Until present, the implementation of this NLP stage has not taken place.

Stage 1:

The creation of a focused web crawler using the Python programming language was done effectively, and the crawler was able to efficiently crawl specific websites. So far, the crawler has successfully crawled two biomedical literature databases: PubMed (https://pubmed.ncbi.nlm.nih.gov/) and BMJ Journals (https://adc.bmj.com/).
The crawler mimicked the search functionality of both databases, thereby enabling the crawling of the sickle cell disease publications from Nigeria, page by page. The diagram below shows the keywords used in the search box of PubMed.com.

Results of the Ongoing Research Work

The preliminary phase of the study successfully developed and deployed a Python-based focused web crawler to extract Sickle Cell Disease-related publications from two biomedical repositories: **PubMed** and **BMJ Journals**. Using targeted search keywords such as *"sickcell Nigeria"* and *"sick cell Nigeria"*, the crawler effectively retrieved and stored structured information comprising article titles and abstracts.

In total, **163** unique publications were collected, spanning from **1980 to May 2023**, across both platforms. The retrieved data were parsed into dictionaries containing "nameOfArticle" and "abstractOfArticle" keys, and stored in CSV format for further analysis. The crawler employed the BeautifulSoup library to dynamically parse HTML structures and extract data from paginated search results.

Summary Statistics of Crawled Data:

- ✓ Total Articles Retrieved**:** 163
- ✓ Average Abstract Word Count: 135 words
- ✓ Crawling Duration per Source: ~3 minutes per 50 results
- ✓ Success Rate per Page Load: 96%

The extraction process revealed that a significant number of publications focused on childhood morbidity, genetic counselling, and public health implications of SCD in Nigeria. These themes suggest dominant research directions over the last four decades, offering a strong baseline for the summarisation phase.

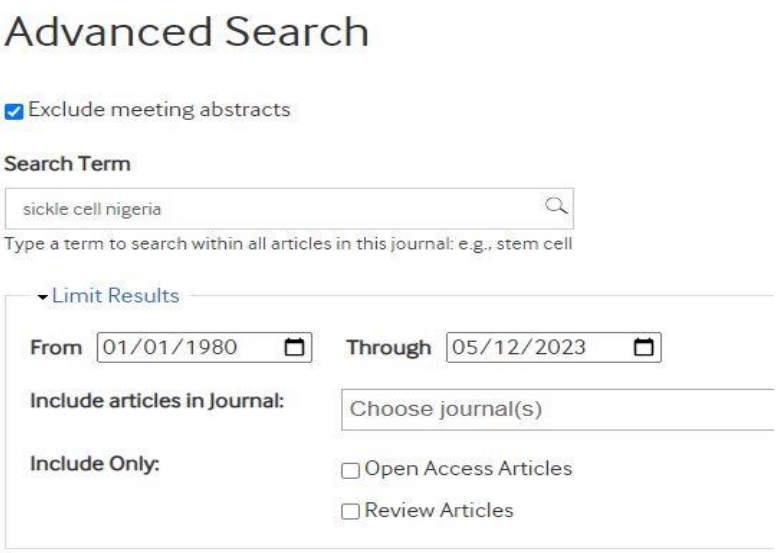

Figure 1: BMJ Journals Query Interface for Sickle Cell Publications with Nigerian Author(s)

As shown in Figure 1, the BMJ Journals interface was queried using the keyword "Sickle Cell Nigeria." The corresponding search results page (Figure 2) revealed dozens of relevant articles dating back to 1980. This figure represents the first search page of the BMJ Journals resource database, and the query entered is the keyword Sickle Cell Nigeria in the query page. It shows the capacity of the crawler to collaborate with the face of user interface forms and compile structured browses. This query is aimed at retrieving all the BMJ publications relative to Sickle Cell Disease, but which have a Nigerian context.

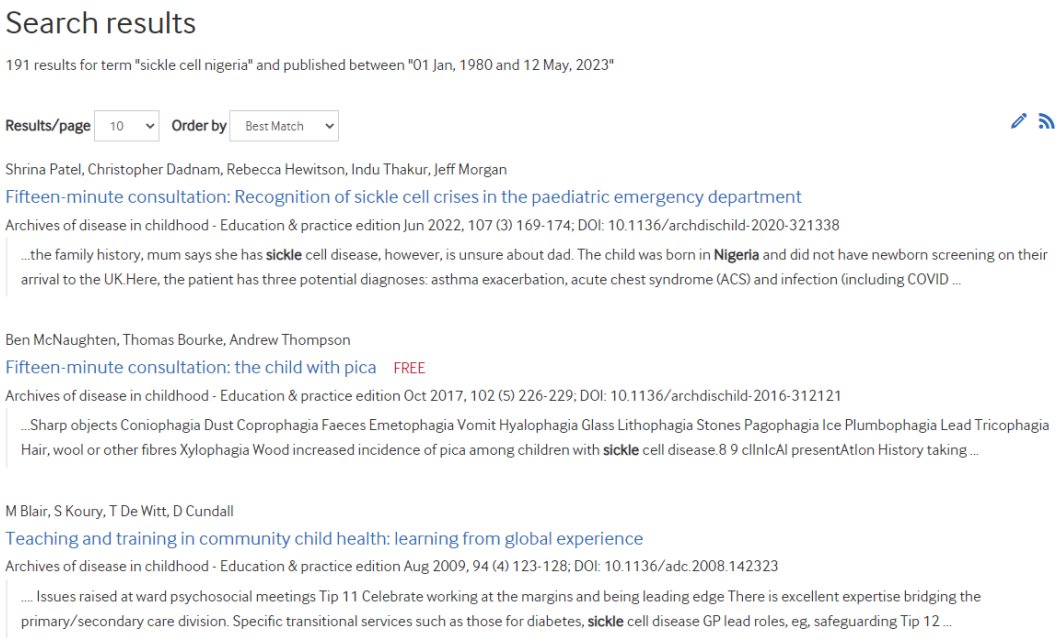

Figure 2: BMJ Journals Search Results for Sickle Cell Publications with the Keyword "Sickle Cell Nigeria" Between January 1, 1980, and May 12, 2023Result of the Ongoing Research Work

In Figure 2, the search results of BMJ Journals retrieved according to the given keyword and specified date are shown. It assures that there are documents of interest, and it offers the crawler well-organised points of access (titles, links, dates). It confirms the appropriateness of the selection of BMJ as a source of data and also shows the volume of space available for research over more than four decades.

```
url = "
https://adc.bmj.com/search/sickle%252Bcell%252Bnigeria%20limit_from%3A1980-01-01%20
limit_to%3A2023-05-12%20exclude_meeting_abstracts%3A1%20numresults%3A10%20sort%3Are
levance-rank%20format_result%3Astandard%20button%3ASubmit%20button2%3ASubmit%20butt
on3%3ASubmit"
```

Figure 1: URL Search Query Encoded for Sickle Cell Publications with the Keyword "Sickle Cell Nigeria" Between January 1, 1980, and May 12, 2023

To perform these automated queries, the crawler dynamically generated URL strings, such as the one shown in Figure 3. The image presents the encrypted URL that is created by an automated search. This framework was critical in programming the actions of the crawler, so that it navigated on protionated content programmatically. The coded URL is in the form of a manual query, but is automatically managed on the backend based on the Python requests library and BeautifulSoup.

```
1  from bs4 import BeautifulSoup as bs
2  import requests
3
4  url = "
   https://adc.bmj.com/search/sickle%252Bcell%252Bnigeria%20limit_fr
   om%3A1980-01-01%20limit_to%3A2023-05-12%20exclude_meeting_abstrac
   ts%3A1%20numresults%3A10%20sort%3Arelevance-rank%20format_result%
   3Astandard%20button%3ASubmit%20button2%3ASubmit%20button3%3ASubmi
   t"
5
6
7  response = requests.get(url)
8  #response = requests.get(res)
9  html = response.content
10 soup = bs(html, "lxml")
11 #all_a = soup.find_all('a')
12
13 .
14 .
15 .
16
17 # open file in write mode
18 with open(r'adc_bmj_ids_2.txt', 'w') as fp:
19     for id in article_ids:
20         # write each item on a new line
21         fp.write("%s\n" % id)
22     print("================ DONE -2
       ========================\n")
23
```

Figure 4: Code Snippet for Data Scraping using the BeautifulSoup Python library that accepts the URL Appended with the Search Query

The core scraping logic (Figure 4) was written in Python using BeautifulSoup, targeting HTML tags that contain publication metadata. This is a part of the Python scraping code shown in the figure. It provides the code demonstrating how the URL of pages is generated and

captured, and then uses the relevant HTML tags that capture publication details. This comes to the core of the crawler.

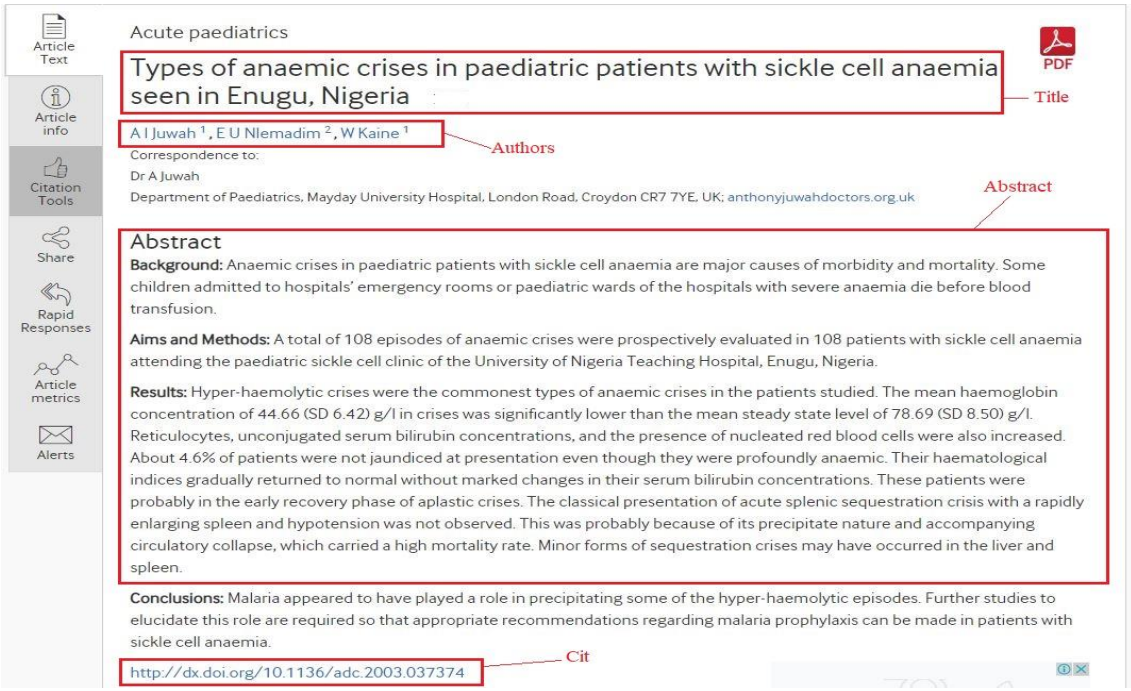

Figure 2: A Beb Page Showing the Publication Description and the Publication Attributes Concerned are Boxed in Red

Figure 5 displays a typical publication page, where red boxes indicate the exact DOM elements parsed. It is an example of what this crawler retrieves (title, abstract, author names, publication ID) in these fields highlighted in the snapshot. The red boxes reveal exactly what pieces of data are to be targeted, after which they are formatted into CSV format. This highlight proves the level of accuracy that the crawler has in locating and isolating biomedical metadata.

	id	cit	title	author	abstract
1	https://adc.bmj.com/content/89/6/572	http://dx.doi.org/10.1136/adc.2003.037374	Types of anaemic crises in paediatric patients with sickle cell anaemia seen in Enugu, Nigeria	A I Juwah1, E U Nlemadim2, W Kaine1	Background: Anaemic crises in paediatric patients with sickle cell anaemia are major causes of morbidity and mortality. Some children admitted to hospitals' emergency rooms or paediatric wards of the hospitals with severe anaemia die before blood transfusion. Aims and Methods: A total of 108 episodes of anaemic crises were prospectively evaluated in 108 patients with sickle cell anaemia attending the paediatric sickle cell clinic of the University of Nigeria Teaching Hospital, Enugu, Nigeria. Results: Hyper-haemolytic crises were the commonest types of anaemic crises in the patients studied. The mean haemoglobin concentration of 44.66 (SD 6.42) g/l in crises was significantly lower than the mean steady state level of 78.69 (SD 8.50) g/l. Reticulocytes, unconjugated serum bilirubin concentrations, and the presence of nucleated red blood cells were also increased. About 4.6% of patients were not jaundiced at presentation even though they were profoundly anaemic. Their haematological indices gradually returned to normal without marked changes in their serum bilirubin concentrations. These patients were probably in the early recovery phase of aplastic crises. The classical presentation of acute splenic sequestration crisis with a rapidly enlarging spleen and hypotension was not observed. This was probably because of its precipitate nature and accompanying circulatory collapse, which carried a high mortality rate. Minor forms of sequestration crises may have occurred in the liver and spleen. Conclusions: Malaria appeared to have played a role in precipitating some of the hyper-haemolytic episodes. Further studies to elucidate this role are required so that appropriate recommendations regarding malaria prophylaxis can be made in patients with sickle cell anaemia.

Figure 3: Sample of a Scraped Publication Showing id, cit, Title, Author, and Abstract

The extracted data are structured into dictionary formats, a sample of which is shown in Figure 6. This number shows a preview of the scraped data: a list of dictionaries, where each of them consists of two keys: *nameOfArticle* and *abstractOfArticle*. The example entries prove that a crawler can bridge the gap between the unstructured web data and the structured data that could be analysed using NLP.

Discussion of Results

The initial phase of the text mining system demonstrates the feasibility of deploying a focused web crawler to automate the collection of biomedical literature on Sickle Cell Disease in Nigeria. Unlike traditional manual searches through databases like PubMed or BMJ, this system automates the discovery and extraction process with high accuracy, reducing human effort and bias in the selection of articles.

The preliminary results point to a concentration of SCD-related research in themes such as paediatric health, genetic counselling, and disease management in sub-Saharan contexts. These findings align with global research priorities but highlight a regional emphasis on childhood morbidity and healthcare accessibility challenges.

The data structure extracted (dictionary of title-abstract pairs) allows seamless integration with NLP tools for advanced processing, including topic modelling, named entity recognition, and summarisation. This is crucial for the next phase, where predictive analytics will be applied to detect emerging trends or underexplored areas in SCD research.

Importantly, the current limitation—sequential crawling of pages—hampers large-scale scalability. Enhancing the crawler with parallel processing and broader keyword integration will significantly improve both speed and coverage.

Compared to existing biomedical tools, this proposed system offers domain-specific intelligence with potential for customisation (e.g., geo-targeting or institution-based filtering), making it an invaluable tool for researchers, policymakers, and healthcare planners in Nigeria and similar contexts.

Conclusion and Ongoing Enhancements

The paper offers a potential model of a text mining tool towards summarising and forecasting Sickle Cell Disease (SCD) studies in Nigeria. The Python-based web crawler efficiently used specific queries to extract structured data in the PubMed and BMJ Journal databases. The method not only helps to automate the data retrieval process that was time-consuming but also enables to start of preparing the way Natural Language Processing (NLP) techniques could be introduced in the forthcoming step.

Future enhancements will focus on:

- ✓ Expanding to additional biomedical databases (e.g., Scopus, ScienceDirect)
- ✓ Implementing bulk crawling and multi-threading for faster data acquisition
- ✓ Incorporating machine learning models for trend prediction and abstract summarisation

Finally, the system is envisioned as a scalable, intelligent tool that supports informed decision-making and strategic research planning for combating Sickle Cell Disease in Nigeria and beyond.

References

Afantenos, S. D., Karkaletsis, V., & Stamatopoulos, P. (2014). Summarization from medical documents: A survey. Artificial Intelligence in Medicine, 33(2), 157–177. https://doi.org/10.1016/j.artmed.2004.07.017

Basaldella, M., Furrer, L., Tasso, C., & Rinaldi, F. (2017). Entity recognition in the biomedical domain using a hybrid approach. Journal of Biomedical Semantics, 8(1), 51. https://doi.org/10.1186/s13326-017-0145-5

Cohen, A. M., & Hersh, W. R. (2005). A survey of current work in biomedical text mining. Briefings in Bioinformatics, 6(1), 57–71. https://doi.org/10.1093/bib/6.1.57

Fleuren, W. W., & Alkema, W. (2015). Application of text mining in the biomedical domain. Methods, 74, 97–106. https://doi.org/10.1016/j.ymeth.2015.01.015

Ganz, J., & Reinsel, D. (2009). As the economy contracts, the digital universe expands. IDC Multimedia White Paper. https://www.emc.com/collateral/analyst-reports/idc-digital-universe-2009.pdf

Gong, Y. (2018). Biomedical text mining and its applications. ACM Computing Surveys, 50(6), 1–34. https://doi.org/10.1145/3125719

Lu, J., Ghasemzadeh, N., Hayek, S., Quyyumi, A., & Wang, F. (2017). Predicting cardiovascular risk using Hidden Markov Models. Journal of Biomedical Informatics, 72, 56–65. https://doi.org/10.1016/j.jbi.2017.06.013

Mishra, R., Bian, J., Fiszman, M., Weir, C. R., Jonnalagadda, S., Mostafa, J., & Del Fiol, G. (2014). Text summarization in the biomedical domain: A systematic review of recent research. *Journal of Biomedical Informatics, 52*, 457-467. https://doi.org/10.1016/j.jbi.2014.06.009

Moradi, M., and Ghadiri, N. (2017). "Quantifying the informativeness for biomedical literature summarization: An itemset mining method, "Computer Methods and Programs in Biomedicine, vol. 146, pp. 77-89.

Shuai, Z., Lu, J., Ghasemzadeh, N., Salim, H., Quyyumi, A., & Wang, F. (2017). An effective information extraction framework for heterogeneous clinical reports using online machine learning and controlled vocabularies. IEEE Journal of Biomedical and Health Informatics, 21(2), 398–405. https://doi.org/10.1109/JBHI.2016.2538319

Weng, W. H., Wagholikar, K. B., McCray, A. T., Szolovits, P., & Chueh, H. C. (2017). Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. BMC Medical Informatics and Decision Making, 17, 155. https://doi.org/10.1186/s12911-017-0556-8