iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VII **Month of publication:** July 2021

DOI: https://doi.org/10.22214/ijraset.2021.36972

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VII July 2021- Available at www.ijraset.com

Novel Approach to Transform Unstructured Healthcare Data to Structured Data

Anusha A R[1], Meghana H D[2], Namitha G N[3], Spoorthi R S[4], Naresh Patel K M[5]
[1, 2, 3, 4, 5]*Department Computer Science and Engineering, BIET Davanagere*

Abstract: ***With the rapid growth in number and dimension of databases and database applications in Healthcare records, it is necessary to design a system to achieve automatic extraction of facts from huge table. At the same point, there is a provocation in controlling unstructured data as it highly difficult to analyze and extract actionable intelligence. Preprocessing is an important task and critical step in Text Mining, Regular Expression and Information retrieval. The accession of key data from unstructured data is often difficult. The objective of this project is to transform the unstructured healthcare data to structured data particularly to gain perception and to generate appropriate structured data.***
Keywords: Healthcare, Unstructured data, Structured data, Text mining.

I. INTRODUCTION

Unstructured data is data that both does now no longer have pre-described data version or it isn't prepared in a pre-described manner. It is commonly textual content heavy, however might also additionally incorporate data which include images, audios, videos, dates, numbers, PDFs, etc. This consequences in irregularities and ambiguities that makes it hard to apprehend the usage of conventional packages in comparison to data saved in fielded shape in databases or annotated in documents.

Structured data is the data which follows information version, has properly described shape, follows a constant order and may be effortlessly accessed and utilized by character or pc program. It is commonly saved in properly-described schemes which include databases. So, that its factors may be made addressable for extra powerful processing and evaluation.

Data mining is a policy of withdraw and coming across styles in huge data units regarding strategies on the intersection of device learning, records and database systems. Data mining is an interdisciplinary subfield of pc technological know-how and records with an average aim to extract data (with clever strategies) from a data set and remodel the data right into a understandable shape for similarly work.

Data mining is the evaluation step of the "understanding discovery in databases" procedure or KDD. Aside from the uncooked evaluation step, it additionally entails database and data control aspects, data pre-processing, version and inference considerations, interestingness metrics, complexity considerations, post-processing of found structures, visualization and updating.

Text mining (additionally referred to as textual content data mining or textual content analytics) is a way for extracting beneficial data from unstructured data thru the identity and exploration of big quantities of textual content in unstructured changing into layout established textual content to discover significant styles and new perception. It allows companies to locate probably precious perception in company documents, consumer emails, name middle logs, verbatim survey comments, social community posts, healthcare data and different reassets of textual content-primarily based totally data.

Text mining applies strategies which include categorization, entity extraction, sentiment evaluation and herbal language processing to convert textual content into data that may be used for similarly evaluation. However, has proved to be a dependable and cost-powerful manner to obtain accuracy, scalability and to spend brief reaction instances for the documents.

II. LITERARTURE SURVEY

Agostino Forestiero, et. al. “Natural language processing approach for distributed health data management”. In: Euromicro International Conference on Parallel, Distributed and Network- Based Processing(2020). In this paper, they proposed Natural language processing approach for distributed health data management. The increasing use of digital health data, like electronic health records (EHRs), has led to store an unprecedented amount of information. Managing this large amount of data can often introduce issues of information overload, with potential negative consequences on clinical work, such as errors of omission, delays, and overall patient safety. Health data are represented with vectors obtained through the Doc2Vec model. The effectiveness of the approach was proved performing a set of preliminary experiments exploiting a tailored implemented simulator.

Veena Bansal, Abhishek Poddar, et. al. "Identifying a Medical Department Based on Unstructured Data: A Big Data Application in Healthcare". In: Information (Switzerland) (2019). In this paper, they proposed a system that will scan prescriptions, referral letters and medical diagnostic reports of a patient, process the input using OCR (Optical Character Recognition) engines, coupled with image processing tools, to direct the patient to the most relevant department. We have implemented and tested parts of this system wherein a patient enters his symptoms and/or provisional diagnosis; the system suggests a department based on this user input.

Jagruti Jangal Wagh et.al. "Unstructured Data Mining and Its Application"(2016). In this paper, they presented that about 90% of the data in the world has been generated over the last two years. If this data is left unmanaged, then it becomes overwhelming, making it difficult to get information from it whenever it is needed. Unstructured data is that which has no identifiable internal structure. The basic goal of a data mining process is the extraction of information from a large dataset and convert or transform it into understandable form for future use.

III. EXISTING SYSTEM

The present industry is focused on structured data, information retrieval from the structured data is easier as they will reside in a predefined model but unstructured healthcare data doesn't have any predefined model where a lot of preprocessing is required. So, there is required to transform the unstructured healthcare data to structured data.

IV. PROBLEM STATEMENT

Preprocessing of unstructured healthcare data is a challenging task as it doesn't reside in a predefined model, as compared with the structured data. So there is a required to transform unstructured healthcare data to structured data. In order to transform we need to take consideration of quasi-identifier, sensitive attributes and names of the patient, where names will reside along the row side and quasi identifiers and sensitive attribute will be an column side.

V. PROPOSED SOLUTION

Analyzing the problem of extracting facts from unstructured healthcare data. Decide data pre – processing techniques that can be applied on the target unstructured healthcare data to reduce the size of our data which will increase the effectiveness of information retrieval system and choose the appropriate approach to analyze our unstructured healthcare data and then converting it into structured data.

VI. OBJECTIVES

A. To collect the unstructured healthcare data.
B. To apply data pre-processing techniques on the target dataset which will increase the effectiveness of information retrieval system.
C. To apply suitable algorithm to transform unstructured healthcare data to structured data.

VII. METHODOLOGY

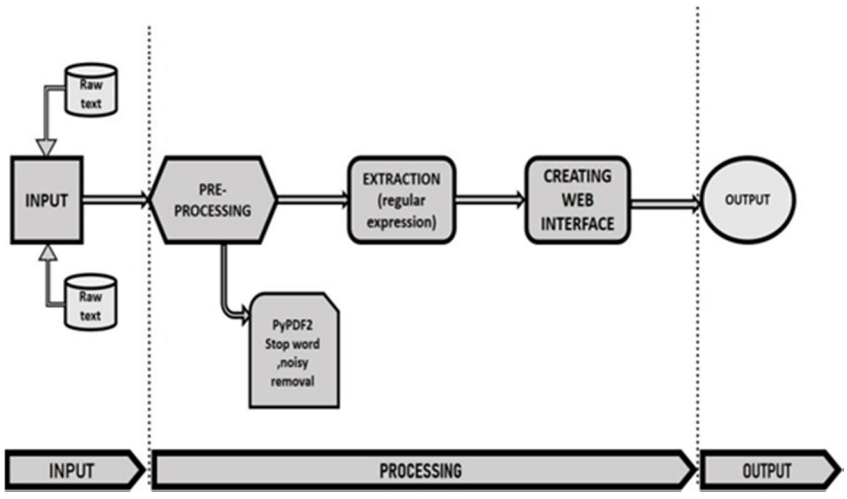


Fig 1: Methodology Diagram

A. Description Of Each Phase

1) *Raw Data:* Dataset incorporates a discharge precis of affected person that's shape of unstructured way that's attain via way of means of the healthcare center.
2) *Preprocessing:* Pre-processing approach performs a completely critical position in textual content mining strategies and applications. A real-world facts normally incorporates noises, lacking values, and perhaps in an unusable layout which can not be at once used for device getting to know models. Data pre- processing is needed obligations for cleansing the facts and making it appropriate for a device getting to know version which additionally will increase the accuracy and performance of a device getting to know version. We will convert PDF to textual content record the usage of PyPDF2 library. The PYPDF2 package deal is a pure-python pdf library that you could use for splitting, merging, cropping, and remodeling pages in pdfs.

PyPDF2 Library is a pure-python library constructed as a PDF toolkit. It is succesful of:

a) Extracting report data
b) Splitting files web page via way of means of web page
c) Merging files web page via way of means of web page
d) Cropping pages
e) Merging more than one pages right into a unmarried web page
f) Encrypting and decrypting PDF files.

3) *Extraction:* Information is extracted from discharge precis of affected person the usage of normal expressions are used to extract a required a part of the textual content via way of means of the usage of superior manipulations. We will extract the elements of the string that in shape a normal expression and save it in python dictionary. Regular Expression is a series of person that specifies a seek sample. In normal expression there's a specialised syntax that's enclosed in a sample. In this syntax we will alternate the styles in it which ends up in one-of-a-kind varieties of facts extraction in line with the sample we specify. Regular expressions can incorporate each unique and everyday characters. Most everyday characters, like 'A', 'a', or '0', are the best normal expressions they genuinely in shape themselves. By matching of strings we attain required data to be extracted.
4) *Create Data Frame:* Data Frame might be created via way of means of loading the datasets from present storage, (SQLite Database). Pandas Data Frame is two-dimensional size-mutable. i.e., facts is aligned in a tabular style in rows and columns. Where we attain primary required data of affected person in dependent facts layout.

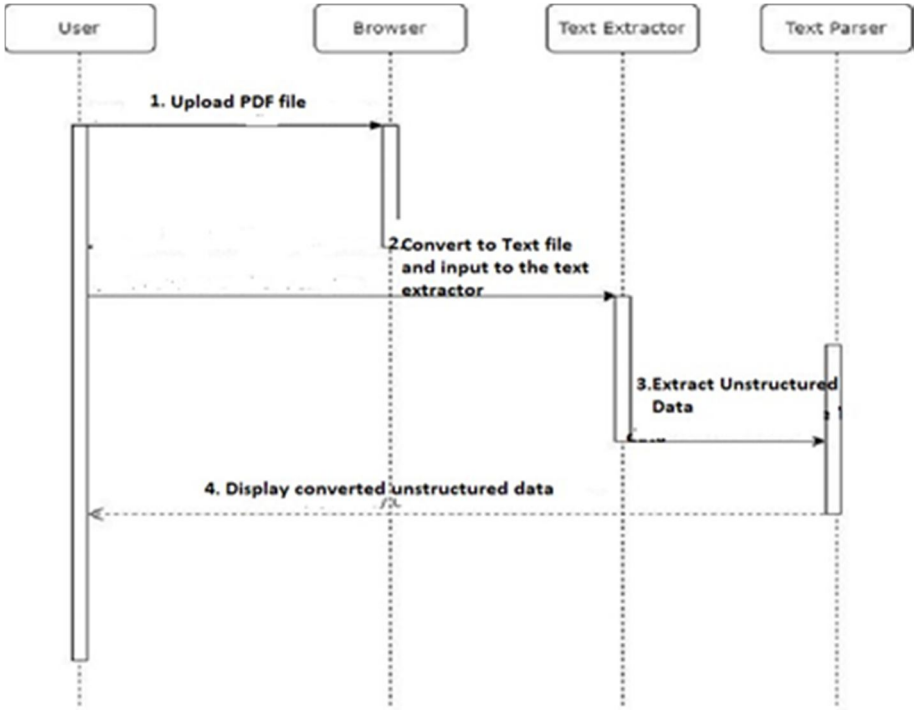


Fig 2. Sequence Diagram

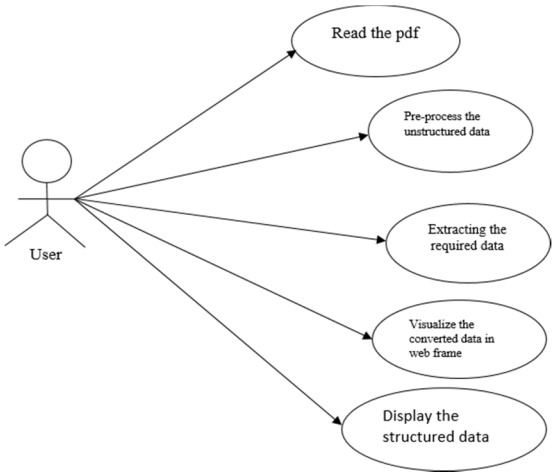


Fig 3: Use Case Diagram

VIII. ALGORITHM

1) *Input:* Unstructured Discharged Summary
2) *Output:* Structured Data
3) *Begin*
a) *Step1:* Read the Input
b) *Step2:* Performs the Preprocessing: Import PyPDF2 python library
c) *Step3:* Feature Extraction:

Import re library

- re.compile(r'(.*)attribute name:(?P<attribute name*[]+\+'))
- re.search (rx file)

4) *Step4:* if key=match

We obtain the data

else

We obtain unbound local error

5) *Step5:* Otherwise repeat step2 until the end of the file

End

IX. RESULT

CONSULTANTS

Dr. SENDIL .G - Cardiologist
Dr. Sandeep Shankar - Cardiologist
Dr. Shashikumar – Physician

DIAGNOSIS

ACS - UNSTABLE ANGINA
CAD - LMCA WITH TVD WITH CALCIFIED VESSELS
LRTI WITH TYPE I RESPIRATORY FAILURE
? H1N1 INFLUENZA
SEPTIC / CARDIOGENIC SHOCK

PROCEDURE
CAG DONE THROUGH RIGHT FEMORAL ARTERY APPROACH ON 5-10-2017

HISTORY OF PRESENT ILLNESS

Mrs. Gangamma aged 70 years presented with the complaints of chest pain, fever, cough since 3 days associated with breathing difficulty since 1 day. Initially treated at Adarsha Speciality Hospital in Tumkur and referred here for further treatment..

PAST HISTORY

A known case of Hypertension on treatment.

GENERAL EXAMINATION

Conscious, Oriented
Temperature : 102
degree F BP : 100/60
mm Hg
PR : 75/min RR : 26/min
SPO2 : 65% on RA on O2
90%

SYSTEMIC EXAMINATION

CVS : S1 S2 +
RS : Bilateral
crepts PA : Soft
CNS : No FND

Fig 4: Input

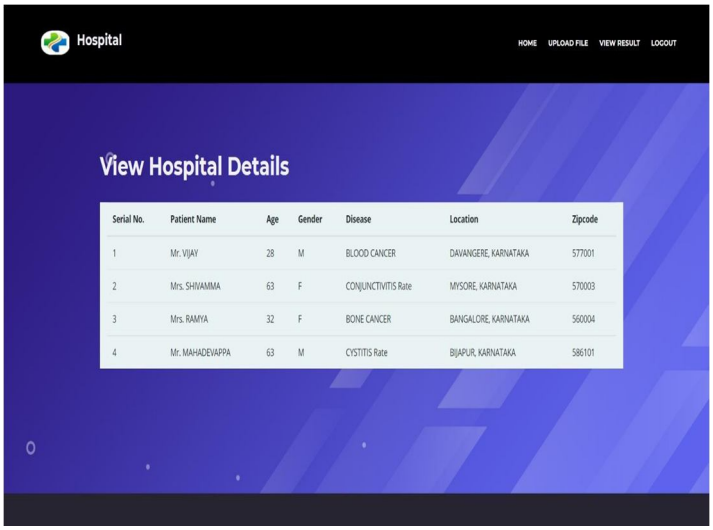


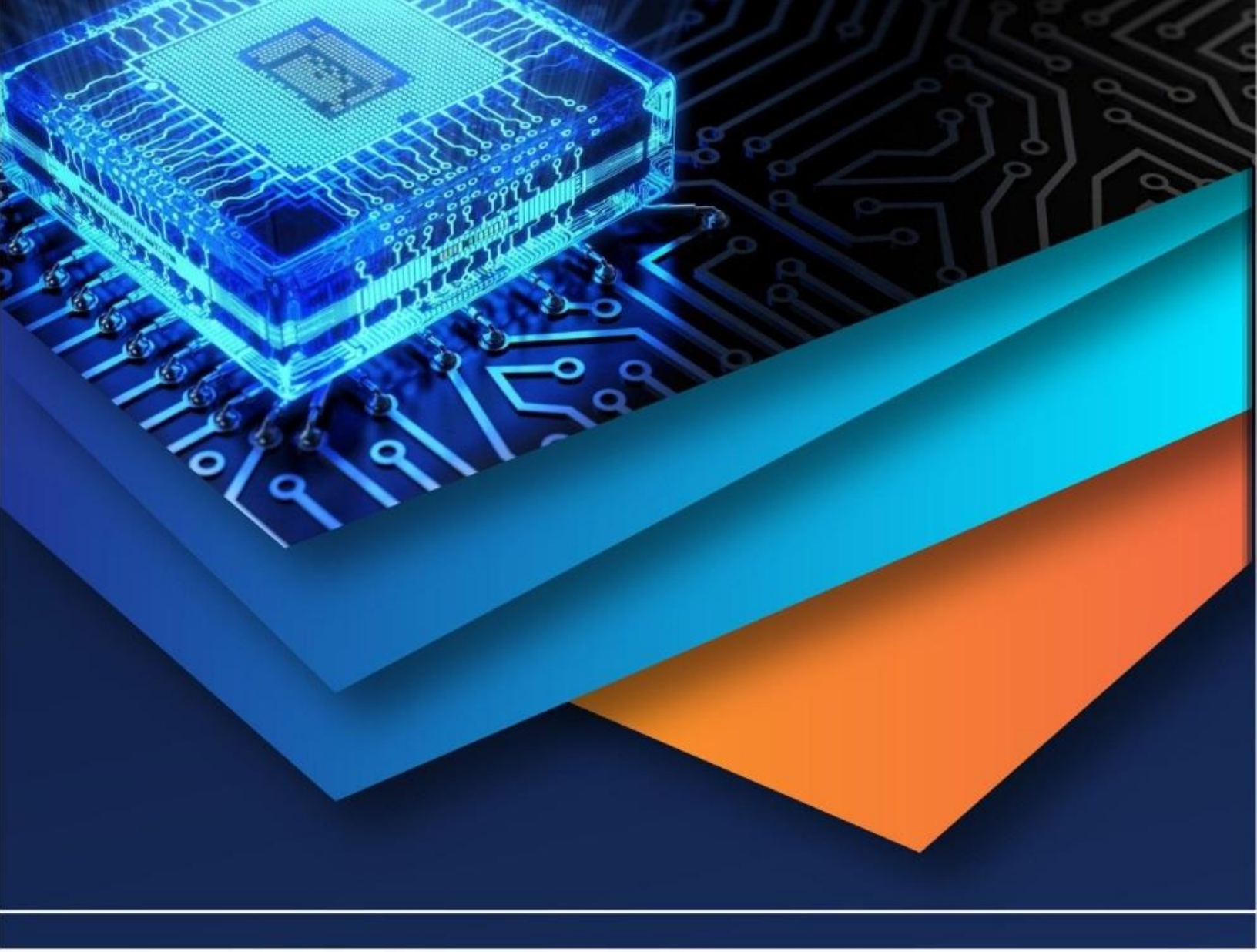
Fig5: Output

X. CONCLUSION

A literature review was carried out to understand the drawbacks and shortcoming of the existing system for the transform of unstructured data to structured data. Our project goal is to transform the unstructured data to the structured data by extracting the most prominent features which providing an efficient storage solution. A systematic procedure along with the methodology has been designed to achieve the project goal.

BIBLIOGRAPHY

[1] Agostino Forestiero, Giuseppe Papuzzo, “Natural language processing approach for distributed health data management”, 2020.
[2] Veena Bansal, Abhishek Poddar, R. Ghosh_Roy, “Identifying a Medical Department Based on Unstructured Data: A Big Data Application in Healthcare”, 2019.
[3] Jagruti Jangal Wagh, Jidnyasa Dharmik Gongane, Ashvini Tulshiram Dukare, “Unstructured Data Mining and Its Application”,2016



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY