**Group Member Details (Name – ID):**

* Mitanshi Patel- 0775260

**A Systematic Approach for Automated Extraction of Data from Clinical Reports using NLP and Artificial Intelligence Predictive Analytics Tools**

**Healthcare Industry Needs**

Healthcare industry accumulates enormous data, which is continuously growing, and is largely semi-structured and/or unstructured in nature. This leads to certain potential challenges for the healthcare stakeholders in utilising it efficiently. Medical data that is collected is in various forms including but not limited to text reports, images, prescriptions, signals, discharge summaries, etc. and this huge unstructured data as per the National Center for Biotechnology Information (NCBI) comprises of about 80% of the entire medical records, which in its raw form cannot be efficiently utilised to extract information.

The medical knowledge has been drastically increased over the recent years and this has resulted in additional diagnostic and treatment possibilities. But at the same time, it cannot be denied that every patient has their own complex medical needs and for this it is essential for healthcare providers to obtain the entire medical history of the patient in a format which could help in gaining valuable health insights. It is absolutely necessary at the point of care (POC) to gather accurate, complete, and most up-to-date information about patient’s health so that finest quality and safe care could be delivered.

The possible solution to maintain the patients’ data over time is through creating Electronic Health Records (EHR), which has the potential to capture the entire history of the patient’s health. The electronic version works by digitising the patients’ medical history and further making it consistent and easily accessible to the healthcare providers. Application of machine learning and artificial intelligence (AI) can be served as an important tool in leveraging text data in the form of clinical reports and can further help in better clinical decision making. These techniques can help healthcare provider to better understand the patents’ disease, and thereby effective care could be extended which can improve the outcome.

**NLP as a Solution**

In healthcare industry, deep leaning is being widely deployed to assist clinicians and thereby, patients through multiple ways. In the past, human involvement was essential to utilise the unstructured data, where clinical narratives and reports were considered as the major form of communication. Recently, artificial intelligence though its powerful tool – natural language processing has been helping the healthcare industry to convert those medical text records to a format which is both useful and easily analysable.

As a solution, two important techniques of natural language processing have been used to analyze the clinical texts - extraction and summarization. This will open up the pathways for healthcare industry to make use of analytics driven opportunities. The clinical reports are usually available in the free text form, clearly this lacks uniformity and consistency. The medical records include but not limited to doctor’s prescription, discharge summaries, diagnosis reports, treatment recommendations, etc. The key information from those text can be potentially taken out using NLP extraction. The optical character recognition (OCR) tool has been deployed to recognize the text from the images which are there in the documents contained in the EHR. This tool is efficient enough to support various file types and hence, will be used in the analysis. Once the key information has been extracted from the large textual data, summarization technique is applied to obtain concise, accurate and meaningful synopsis. In the analysis, the focus was to apply these machine learning techniques efficiently so that better technical solutions can be extended to healthcare industry.

**Benchmark**

In a study “Automated Methods for the Summarization of Electronic Health Records” which was published in National Center for Biotechnology Information (NCBI), it was emphasised that the revolution of electronic health records has led to an unprecedented increase in the patient’s health information that is stored in an electronic format. As a solution, clinical summarization is adopted for chronically ill patients as they usually have large datasets. They have defined clinical summarization as the act of collecting, describing, and shortening the patient’s information by taking the required points into consideration. This study showcases that for these patients, medical information is difficult to present in a consistent way, which highlights the need for comprehensive EHR summarization. This paper also outlines the techniques and important methodological challenges in the area.

The challenges that were considered in this study are:

• Information redundancy

• Temporality

• Missing data

• Salience detection

• Rules and heuristics

• Deployment of summarization tools

The article defined the summary in four different categories., which are extractive, abstractive, indicative, and informative. The study founds out that, so far, all the summarization models fall under informative and indicated category. But unfortunately, all the summarization models have lacked in evaluation.

**Work-Flow Chart**

Diagram

Description automatically generated

**Data Collection and Preprocessing**

Data in healthcare industry increases with high velocity and in great volume which makes it big data, which is in various formats such as reports, text records, images, radiation frequencies, audio etc. and is generally unstructured in nature which has to be organized in order to draw information from it. This also creates a challenge for doctors to go through multiple pages for every patient to get to a conclusion.

The data was obtained from the medical professionals, which are in the form of actual patients’ reports from Windsor Regional Hospital. It was in the form of PDF that contained scanned images of the patient’s progress notes, diagnosis, prior and current therapy, imaging investigation results, laboratory reports, so on and so forth.

These images were unclear and contained different font types that made it difficult for the tool to extract the text. So, preprocessing of these images was required so that Optical Character Recognition (OCR) technique could be applied to extract the text.

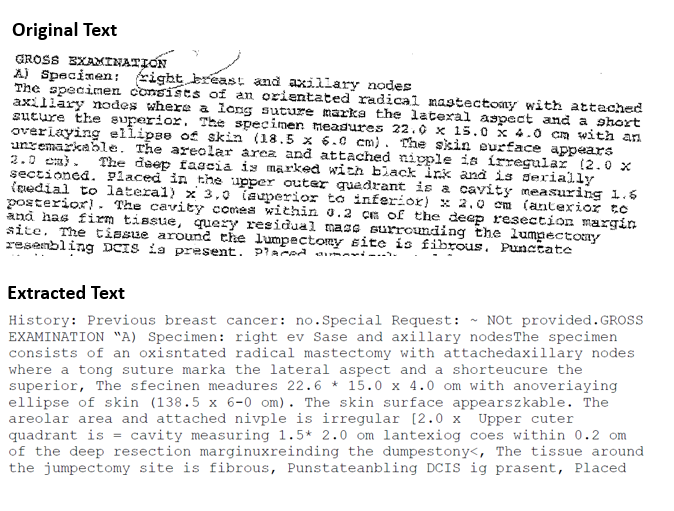
OCR is a technique used to recognize the text which is converted into image. This technology can transform any type of image with text in it into a machine readable format. This technology has been implemented with the help of an OCR engine called Tesseract. This engine looks for models in pixels, letters, words, and sentences. It deploys a two-step method known as adaptive recognition. It needs one data stage for character recognition, followed by a second stage to fill in any missing letters with letters that fit the word or sentence meaning.

To achieve this OCR technique in python, an AWS associated library TEXTRACT has been used. This also works using tesseract OCR. The code used will check if the images are present in the document and if yes, it will use tesseract OCR to convert that images into text. Using encoding the text is converted back into string so that it makes the pre-processing easy. This text extraction process is kept in a python function so that it will be easy to just call and reuse the method.

Pre-processing of the text in an NLP task is very important, because it removes any noise in the data. A library called regex is used to apply the different cleaning techniques. Regex is short form for regular expressions and helps in finding specify patterns and replace them with user specified string. Generally, in text pre-processing, filtering numeric values out is must. But, since our task is summarising the health reports numeric values (Blood Pressure, Heart Rate etc.) are important. The escape sequences were also removed from the extracted text using inbuilt python functions. Further, the special characters and punctuations were also eliminated without affecting the medical terminologies and the sentence formatting.

Next thing in pre-processing is checking the spelling of the words extracted. For this a python library spellchecker is used. This library gives user to use their own word library to verify our text against with. We have found a dictionary with most used medical words since our project is related to medicine. Using this word dictionary in the spellchecker library, we have checked the words for misspelled. Now, we found the misspelled word and its corrected word. To replace the misspelled word with its associated corrected word, we created a dictionary in python which can store misspelled word as key and corrected word as its value. Now a simple for loop will replace the words. This text pre-processing process is kept in a python function so that it will be easy to just call and reuse the method.

Here is an example to illustrate the original image and the text extracted from it:



**Methodologies**

NPL summarization which is a technique to prepare a synopsis from huge textual records has been implemented in the analysis. To execute this, two language processing python libraries were used Gensim and spaCy.

Gensim is an open source library in python tool used mainly for topic modelling. In the analysis, genism implementation is done based on TextRank algorithm, which extracts semantic text from documents automatically. This model has the potential to highlight critical information in a large corpus. The summarizer was adjusted to provide apt proportion of the output and it was analysed that 30 percent of the entire medical text record can summarize the medical document most efficiently.

Another model has been built using spaCy package, which is an inbuilt python application that can understand and process large amount of text data. The medical text records were preprocessed to apply summarizer effectively. For this, the input text was first converted to lower case and tokenized using spaCy’s language model. The loop was created to process each of the tokens and the model was prepared in a way which could potentially ignore the tokens if those were stop-words or punctuations. The keywords were normalised on the basis of their individual weightage. The keywords which were converted to lists were stored in a dictionary using a loop function along with their frequencies.

The results returned the frequency value of 1 for the keywords with highest weightage. Further, the spaCy model has divided the sentences based on full-stop punctuation. The normalized keyword values were added to the key-value pair of the sentence, which denotes the sentence strength and the sentences with more strength were considered in the summary and less important ones were eliminated on the basis of the limits assigned. This limit is basically the number of words to be presented as a summary.

The summaries from both the models, Gensim and spaCy, were compared and because of spaCy’s reliability to obtain better output in the form of summary, it was taken as a final model to be used to prepare Graphical User Interface.

**Graphical User Interface**

**Results**

**Challenges**

**Future Scope**

**References**

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4607324/>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6372467/>

<https://info.nhanow.com/blog/whats-ehr-and-why-is-it-important>

<https://www.healthit.gov/faq/what-are-advantages-electronic-health-records#:~:text=The%20benefits%20of%20electronic%20health%20records%20include%3A,timeliness%2C%20efficiency%2C%20and%20equity>.

<https://hitinfrastructure.com/news/unstructured-healthcare-data-needs-advanced-machine-learning-tools#:~:text=About%2080%20percent%20of%20healthcare,of%20Things%20(IoT)%20devices>.

<https://towardsdatascience.com/natural-language-processing-nlp-for-electronic-health-record-ehr-part-ii-1b2a024ea03a>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7157505/>

<https://towardsdatascience.com/how-to-extract-text-from-images-with-python-db9b87fe432b#:~:text=The%20Python%20Library&text=Python%2Dtesseract%20is%20an%20optical,for%20Google's%20Tesseract%2DOCR%20Engine>.

<https://www.analyticsvidhya.com/blog/2018/11/introduction-text-summarization-textrank-python/>

<https://medium.com/@oyewusiwuraola/simple-natural-language-processing-projects-for-health-sciences-how-to-extract-and-summarize-53ea2e7b3f1d>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4986665/>