Automatic Generation of Repeated Patient Information for Clinical Notes

Prathiksha P Pai[1], Dr. Sarika Hegde [2]

[1]*MTech Scholar, Department of Computer Science and Engineering, N.M.A.M Institute of Technology, NITTE, Karnataka, India*

[2] *Associate Professor, N.M.A.M Institute of Technology, NITTE, Karnataka, India*

Abstract: *Clinical notes consist of set of patient's information like surgical information, lab test reports, medical reports, demographics and reason for current visit etc. Generating clear and accurate clinical notes manually will be a time-consuming task for physician. To this end, the system can automatically generate a new clinical note, when the sentences from the past clinical notes are stored in the excel file are given as the input. The newly generated clinical notes will be timely, accurate, readable and clear for the future references. We process past clinical notes of the patient using the concept of Natural Language Processing and Naïve Bayes Algorithm to generate the new clinical note. At the end, system will show the accuracy of the algorithm based on the diseases score count present in the past clinical notes of the patient. This process will help physician to retrieve any patient's information easily and give the proper treatment to the patient based on the past clinical data.*

Keywords: Natural language processing, Clinical report Processing, Clinical notes generation, Naïve Bayes Classifier.

1. Introduction

Clinical notes are the document which will provide all the important medical information of the patient and these records can be maintained for future references. Clinical notes will include patient's information like treatment details, lab test reports, medical reports, surgical history, demographics, reason for current visit and recommendation of future treatments. Clinical notes will provide the complete, accurate information about the patient's health record. This will help the patient without repeating the test again and again. Physician can give prescribed treatment to patient by referring to patient's past clinical notes and also it will facilitate a good decision making for a single patient by physician. To generate patient's

new clinical notes, sentences from the past clinical notes are given as input. The information present in past clinical notes are processed to automatically generate new clinical notes. Newly generated clinical notes will be timely, accurate and also it should attain continuity to give proper treatment to the patient.

Generating clear and readable new clinical notes manually from patient's past health record is time consuming task for the physicians, so this system is used to automatically generate the patient's new clinical notes from set of past clinical notes, which will also save the physician's time. The generated clinical notes will give the accurate results.The accuracy of the system depends on the past clinical data. This system will provide the security to the patient's new generated clinical notes. In the proposed research work Natural Language Processing and Naïve Bayes classifier is used to extract the important features from the sentences of patient's past clinical notes and to generate a new clinical note.

2. Literature Review

A literature review is the process of searching or evaluation of the present or available literature in the chosen topic area. Literature review gives the details of the work conducted on the chosen topic, so readers can start the new work rather than repeating the same work.
In the paper [1], the method of summarizing the clinical notes of the psychiatric patient will help to serve the patient care. Summary report includes patient's demographics, time line for clinical visits, treatment history and hospital stays. The author's aim is to reshape psychiatric records by encouraging the mental health professionals.
In the paper [2], the system will generate text document of patient's medical progress notes. After examining the patient, the handwritten alphanumeric code is generated. This alphanumeric code is entered into computer and converted to text. The output text is used to form patient's medical progress note. In this method doctor will generate the supplementary notations and set of code subsequent to patient for medical notes. The generated medical notes by doctor is fed as input to programmable computer. This process is useful for doctors, as it consumes less time to convert code into text for clinical notes.
In paper [3] new subsequent patient's clinical notes are generated from past clinical notes. The system will generate clear, accurate and readable new clinical notes for physicians. The clinical notes will include patient's information like medical history, name, age, location, surgical history, treatment details and this clinical notes details will vary from patient to patient. This approach will generate repeated patient information by using semantic patterns and approximate sequence matching algorithm, which will capture the discourse role of sentences. This feature is useful to determine whether the sentences are repeated or not. In paper [4] textual summaries will help to make a better decision for doctors to give treatment

to the patient. The Baby Talk project (BT-45) generated textual summaries of sick babies in neonatel Intensive Care unit for about 45 minutes of discrete values and continuous physical signals. The data generated by BT-45 is similar to quality decision making as visualization. The computer system can generate the textual summaries of complex continuous clinical notes.

In paper [5] author presented the results from automatic and manual evaluation of seven different methods for automatically generating clinical text summaries. The handwritten clinical notes were used as summaries for doing automatic evaluation. Among these summarization methods were the control methods. Random and oracle and four rouge metrics were used for automatic evaluation. The datasets used here contains electronic health records of approximate 26,000 patients admitted to hospital between the year 2005 to 2009. Majority of the data was stored in Electronic Health record system.

In paper [6] large amount of biometric data is generated and is available for healthcare experts. The aim is to achieve GUDM (Global Unified Data Model), this is achieved by making use of "Data Modeler" tool. This framework has easy way to use GUI and provide easily access to the knowledge engineers to obtain unified dataset. The tool (Data Modeler) is illustrated using sample diabetes mellitus data. This tool was tested and showed that it reduced the time effort need to locally create diverse dataset by experts and knowledge engineer to 94.1%.

In paper [7] the patient's medical documents are automatically summarized using extractive text summarization method. Text summarization deals with study of summarizing the text document. Medical documents consist of all major source of patient's medical information, this is the time-consuming task for physicians. Hence the medical documents are summarized, which will reduce the task of the physicians. It deals with the study of summarizing heart patient's medical documents. The method used in this paper depends on word space models that are constructed using distributional semantic models. Heart related patient's document is summarized based on manual evaluation and automatic evaluation of summarizing process and result is compared. The automated evaluation is more accurate and consumes less time.

In paper [8] medical images are used to diagnose and treat a patient; the best example is x-ray report of fracture. Doctor will analyse the medical image and write textual reports for findings. But manually writing to prone to error. So, this paper has provided a methodto automatically generate textual reports for the medical images. It makes use of hierarchical LSTM network that can capture semantics and produce long text. The dataset used contain radiology and pathology images. In paper [9] the diabetes register of Bulgaria region patient is generated automatically using the outpatient records. Outpatient records contain details

of lab data and values of clinical tests. The four free text fields are anamnesis, status, clinical tests and prescribed treatment.

This paper [10] describes about the value of clinical notes in the field of healthcare. The important factor in medical domain is clinical data and clinical notes. One of the major drawbacks in research on healthcare is lack of datasets. The reason for this is patient's privacy. Deidentification method can be addressed to solve privacy problem but it leads to adversarial attacks. This paper makes use of NLP to generate synthetic clinical notes from the vast amount of unstructured data stored in original medical report and make sure patient's information and privacy is not compromised. Hence by making use of synthetic generated report the models can be trained, this will help in medical NLP research. The example of experiments done using natural language model is shown in this paper which gives almost similar notes that of the real ones in some of the clinical NLP tasks.

3. Objectives

i) To generate accurate, clear and readable clinical notes from patient's past clinical notes.

ii)To easily access or retrieve the required information of the patient for clinical notes.

4. Materials and Methods

Datasets

Example 1

Figure 1 is the clinical note of one patient having PID as 1, who visited clinic three times. This particular clinical note consists of the patient details like name, Patient ID (PID), age, first visit, second visit and third visit details. Natural language and Naïve Bayes classifier are used to generate the new clinical notes of the patient.

Name: Suma
PID: 1
Age: 42
First entry: Patient is suffering from viral fever and migraine frompast 4 days. Prescribed medicinesgiven are paracetamol and rizatriptan.
Second entry: Patient is suffering from cough and diarrhea, admitted to hospital for about 3 days. Prescribed medicines given are benzonatate and lomotil.
Third entry: Patient is suffering from heartdisease and have undergone bypass surgery one time. Patient is having high cholesterol, admitted to hospital for about 7 days. Prescribed medicines given are aspir, atorvastatin and crestor.

Figure 1. Past Clinical note of Patient having PID as 1

Example 2

Figure 2 is the clinical note of one patient having PID as 19, who visited clinic two times. This particular clinical note consists of the patient details like name, Patient ID (PID), age, first visit and second visit details.

Name: Karan Patel
PID: 19
Age: 20
First entry: Patient is suffering from leukemia and prescribed medicine given are sprycel.
Second entry: Patient is suffering from malaria from past 15 days and admitted to hospital for about 11 days. Prescribed medicine given are coartem.

Figure 2. Past Clinical note of Patient having PID as 19

The details from the past clinical notes is stored in excel file as row and column format under its attribute like Name, PID, Age, First entry, Second entry, Third entry etc. Similarly collected 101 patients details with unique PID and stored it in the excel file for easy processing of the data.

Methods

In figure 3, the sentences that are stored in excel sheet from the past clinical notes is given as the input and new clinical notes is generated as the output. In the first step the sentences of the past clinical notes stored in excel file are pre-processed using the concepts of Natural language processing.

Spelling Checker: It will check the spelling of the technical words that are present in the sentences of input clinical notes. It will compare with the spellings present in the dictionary.

Tokenizer: Tokenizer is the process of breaking the paragraphs into sentences and sentences into words. So, in the given datasets i.e. Paragraphs from the past clinical notes are broken down into sentences and stored in the form of array, this process is known as sentence tokenizer. Then the sentences are broken down into words and stored in the form of array, this process is known as word tokenizer.

Stop word remover: NLTK English dictionaries consists of many stop words like from, and, this, my, myself, very, they, them, is, are, into, who, what, which etc. Stop word remover will remove all stop words present in the input datasets.

Punctuation mark remover: In this step, all the punctuation marks like .,?;!:’”()[]”’-_/@{}*… are removed from the input datasets i.e. from the sentences stored in excel sheet.

Create Dictionaries: Created two dictionaries for medicine and diseases, included almost all the words in the dictionaries. These two dictionaries consist of case sensitive words and the words should match the input datasets.

Category tagging: Category tagging is important step for tagging the technical words that are present in sentences of the clinical notes. Words present in clinical notes should match the words present in dictionary.

Naïve Bayes classifier: Naïve Bayes classifier is used for score calculation. It will calculate the score of sentences based on the diseases present in the sentences of past clinical notes. It will also calculate the model's performance and give the models accuracy based on the diseases count. Training datasets given to the classifier are words from the created dictionaries.

It will extract some of the important features from the sentences, so sentences of past clinical notes are given as testing datasets. Naïve Bayes classifier will also help the physicians in decision making about the treatment given to the patient based on their past clinical notes. This can help the patient also without undergoing lab tests again and again. Classifier will calculate the overall score based on the diseases count. It will count the number of diseases in the sentence and estimate the accuracy of the algorithm.

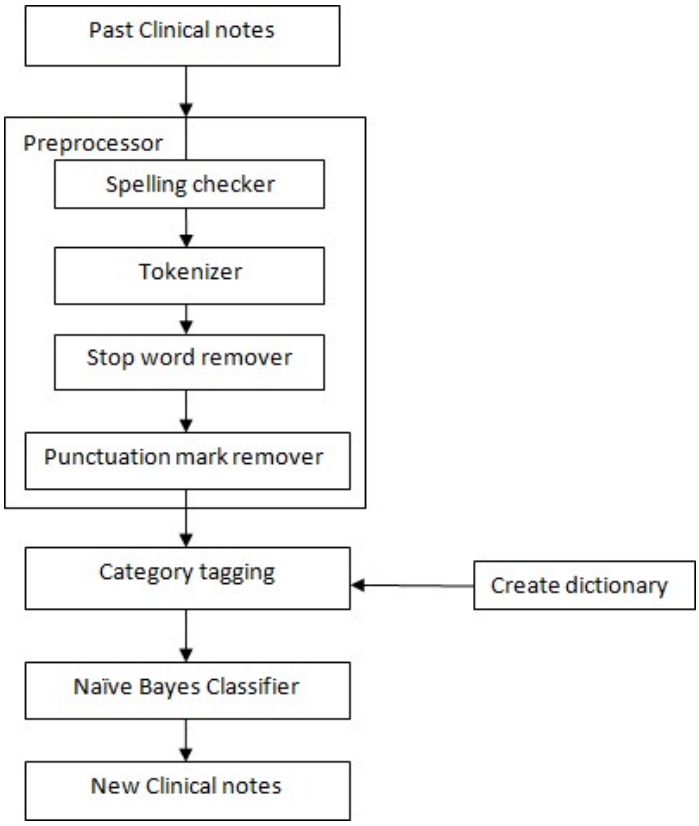

Figure 3. Architectural diagram of system

In final stage the output summary of past clinical notes is generated. The new clinical notes generated from past clinical notes will be accurate, timely and clearly readable document.

5. Results

The final result obtained in this work is repeated patient information can be easily accessible and automatically generated, when the valid patient ID (PID) is entered. The newly generated report is accurate, timely and clearly readable document which can be used for future references. The model's accuracy is shown in the form of graph. Naïve Bayes Classifier model performance is calculated based on the number of the diseases present in the given input dataset. The doctor can enter the patient id, if the patient record is present it will show the message, patient record exists and generates the patient's new clinical notes. If patient record is not found then it will not generate any patient's clinical record, it will show the message, no record found.

The output of new clinical notes is generated automatically from the past clinical notes of patient ID 1 (i.e. Example 1) is shown in figure 4. Automatically generated new clinical note is clear and accurate, this record is maintained by physicians for future reference.

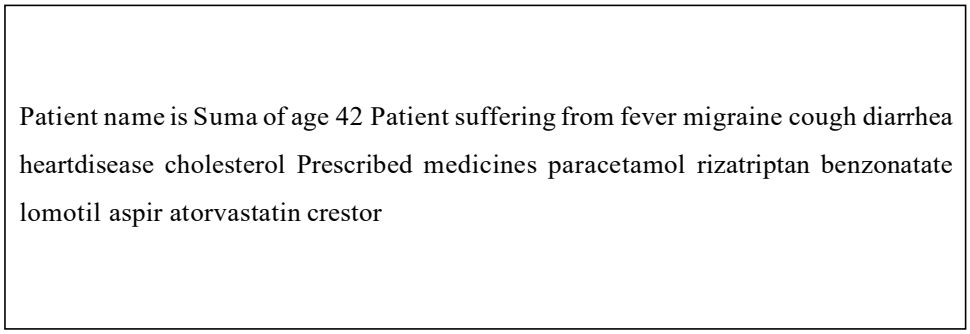

Figure 4. New clinical notes generated automatically from Past clinical notes having PID as 1

A graph is plotted, which gives the models performance. In the figure 5 it shows that the model is 97% accurate for the given dataset. The accuracy of the model depends on the disease count of the past clinical notes in figure 1 is 6 i.e. fever, migraine, cough, diarrhea, heartdisease, cholesterol.

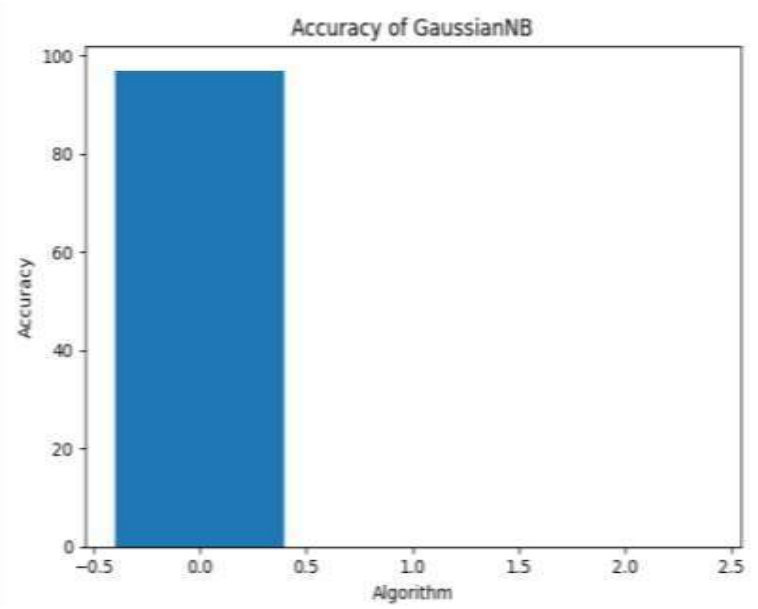

Figure 5. Prediction performance results obtained from Naïve Bayes Classifier for example 1

Similarly, the output of new clinical notes is generated automatically from the past clinical notes of patient ID 19 (i.e. Example 2) is shown in figure 6.

Patient name is Karan Patel of age 20 Patient suffering from leukemia malaria Prescribed medicines sprycel coartem

Figure 6. New clinical notes generated automatically from Past clinical notes having PID as 19

A graph is plotted, which gives the models performance. In the figure 7 it shows that the model is 96% accurate for the given dataset. The accuracy of the model depends on the disease count of the past clinical notes in figure 2 is 2 i.e. leukemia and malaria.

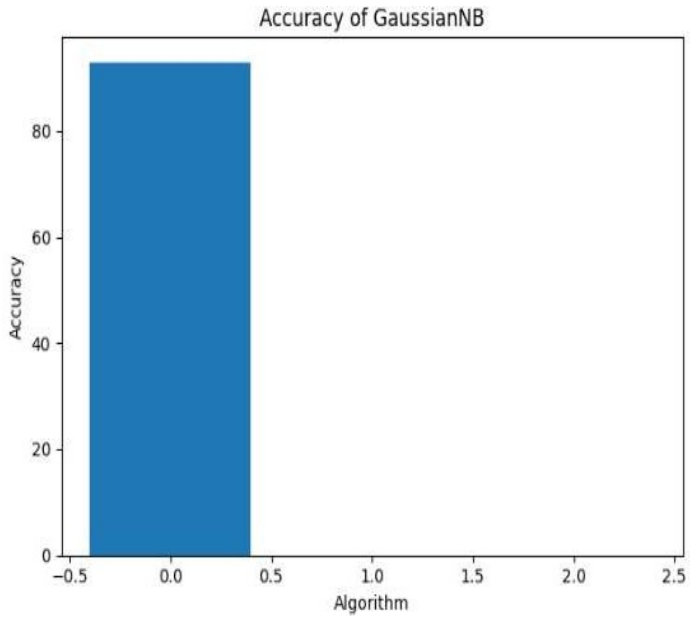

Figure 7. Prediction performance results obtained from Naïve Bayes Classifier for example 2

Manually generated clinical notes by doctor may contain error and not clearly readable, but the automatically generated patient's clinical notes by the system will be more accurate when compared to manual method. Automatically generated clinical notes will be easy to read, understand and error free.

Manually generating clinical notes will be time-consuming task for physicians and also there will be high risk of errors in the result. Automatically generated results will be quickly generated by the system based on its past clinical notes.

6. Conclusion and Future work

In the given study, it has been demonstrated that how the repeated patient information is automatically generated and also the required patient information can be easily accessible using the concept of Natural Language Processing and Naïve Bayes Classifier. The model's performance is shown in the form of graph. Automatically generated new clinical note will be more accurate, complete, timely and readable. This will reduce the burden of physicians without generating the report manually and the results obtained are error free.

For the future work, the proposed work can be extended using LSTM Neural network model, it will use bidirectional wrapper with LSTM layer this can propagate the input forward and backward and also it can concatenate the output. These types of models can give more accurate results. Therefore, possible solutions for considering such problems are great area for research domain.

REFERENCES

[1] Powsner and Tufte, "Summarizing Clinical Psychiatric Data", November 1997

[2] Marc Edward Chi corel, "Computer keyboard generated medical progress notes via a coded diagnosis-based language", Feb 2001

[3] Frank Meng, Ricky K. Tiara, Alex A.T. Bui, Hooshang Kangarloo, Bernard M. Churchill, " Automatic generation of repeated patient information for tailoring clinical notes", March 2005

[4] François Portet, Ehud Reiter, Albert Gatt, Jim Hunter, Somayajulu Sripad, Yvonne Freer, Cindy Sykes, "Automatic generation of textual summaries from neonatal intensive care data", December 2008

[5] Hans Moen, Juho Heimonen, Laura-Maria Murtola, Antti Airola, TapioPahikkala, Virpi Terava, Riitta Danielsson-Ojala, TapioSalakoski, and Sanna Salanter, "On Evaluation of Automatically Generated Clinical Discharge Summaries", 2014.

[6] Rahman Ali, Muhammad Hameed Siddiqi, Muhammad Idris, Taqdir Ali, Shujaat Hussain, Eui-Nam Huh, Byeong Ho Kang, Sung young Lee "Automatic Generation of Unified Datasets for Learning and Reasoning in Healthcare", July 2015.

[7] Hans Moen, Laura-Maria Peltonen, Juho Heimonen, Antti Airolab, Tapio Pahikkala, Tapio Salakoski, Sanna Salanterä, "Comparisonof automatic summarisation methods for clinical free text notes", January 2016.

[8] Baoyu Jing, PengtaoXie, Eric Xing, "On the Automatic Generation of Medical Imaging Reports", Nov 2017.

[9] Dimitar Tcharaktchiev, Zhivko Angelov, Svetla Boytcheva, Galia Angelova "Automatic generation of a national diabetes register from outpatient records", 2018, pp. 163-166

[10] Melamud and Shivade, "Towards Automatic Generation of Shareable Synthetic Clinical Notes Using Neural Language Models", June 2019.