Text Mining on Electronic Medical Record

Dhanalakshmi T S
Dept. of Information Science & Engineering
RV College of Engineering®
Bengaluru, India

Merin Meleet
Dept. of Information Science & Engineering
RV College of Engineering®
Bengaluru, India

Abstract—Fast advancement in computerized information obtaining procedures have prompted immense volume of information extraction of text. Most of the data is composed of either unstructured or semi-structured form of text. To make this unstructured form of data into structured form using text mining, natural language process (NLP) techniques and machine learning algorithms are used. Cancer based text are in the form of Electronic Health Record (EHR/EMR) and there are tools to extract the text. Health care and clinical practice create a lot of content manifestations, test results, analyse, medicines, also, results for patients. This clinical content, reported in wellbeing records, is a potential wellspring of information and an underused asset for improved social insurance. To improve understanding consideration, information on demonstrative, prognostic, inclining, and medication reaction markers are fundamental. In this paper explored different text mining approaches using machine learning, natural language processing and data mining techniques.

Keywords—Clinical text, Machine learning, Data mining, EMR/HER, NLP

I. INTRODUCTION

EMR is been popularised with the development of Hospital Information System (HIS) and information technology. Electronic health record used charts, symbols, text, data and other digital information which can be stored, managed, reproduced and transmitted efficiently.

Malignancy is a dangerous sickness that has caused a huge number of human passing’s. Its examination has a long history of well more than 100 years [1]. There have been a colossal number of distributions on malignancy explore. This coordinated yet unstructured biomedical content is of incredible incentive for malignant growth diagnostics, treatment, and avoidance. Biomedical content mining on malignancy explore is computationally programmed and high throughput in nature [1].

Three kinds of data are divided in EMR data such as: structured, semi-structured and unstructured form [1]. Generally structured data is in fixed mode contains information such as (eg. in binary form for prediction). Usually flow chart forms are the type of semi-structure data. Unstructured data which contains a clinical note which will be in text form such as discharge patient records [2].
II. LITERATURE REVIEW

Literature review briefs about the techniques and methods used for text mining using EMR.

1 Text Mining Process

In this paper [3] authors give a brief introduction to the concepts of text mining in EMR. Extracting the useful data, which is in unstructured form and uses classification techniques. Its applications vastly used in the many application especially most in the medical field which uses EMR data.

Models of text mining is composed into three categories: text pre-processing, text mining operations, post-processing. Pre-processing technique which converts text into intermediate form. Text mining system includes rule-based association, analysis and NLP techniques.

Knowledge discovery in text (KDT) proposed by Feldman et al. KDT general architecture which takes two input knowledge-labelled collection of documents and directed acyclic graph keywords shown in figure 3.

![Figure 3: System architecture of KDA](image)

In this paper [3] authors give a brief introduction to the concepts of text mining in EMR. Extracting the useful data, which is in unstructured form and uses classification techniques. Its applications vastly used in the many application especially most in the medical field which uses EMR data.

Text clustering [15] gives a topic analysis method. Initially extracting the name entities from the EMR Hypergraph based method [16] is based on named entity.

![Figure 5: Model mining system](image)

1.3 Text Association is rule based extraction proposed by Agarwal [17] finds a relation between features of words from the different text collection. Two basic measures for association rule [18].

\[
\text{Support}(W,W_i) = \frac{\text{Support count of } W,W_i}{\text{Total number of documents } D}
\]

\[
\text{Confidence}(W_i \backslash W_j) = \frac{\text{Support}(W,W_i)}{\text{Support}(W_i)}
\]

1.4 Text Analysis is when-evolutionary theme graph is generated from the word clustering. Themes form the text are summarized through probabilistic approach and patterns from the text [19]. This allows to compare the overtime themes in different relative strength. The changing trend uses for calculation of keywords distance.

2 Classification Data pre-processing

EMR consists of different types of data sources and retrieval which may be incomplete redundant. So, required to do pre-processing of the data in order to get its accuracy. Pre-processing steps includes data cleaning, data integration, data transformation, data reduction and privacy protection [3].

2.1.1 Data Cleansing.

Due to manual errors, system failure data may lose its attributes. Missing data can be ignored by filling manually the values to retrieve data sources. Ignore missing data for processing when the missing value has great influence. The data can be ignored when operation name is lost for patient information extraction, in case bed number is lost data cannot be ignored. Default value can be filled during such cases for small data sets but not the same case can be applied for large data set which is time consuming and costly. Machine learning methods can be used for getting optimum value. Decision tree induction, Bayesian
method. The data source should be retrieved for other data sources if it is missed.

2.1.2 Noise Processing
Illegal value in a data source. The regression method can be used to classify function model by modifying the noise value. A large deviation with the attributes between clusters in a data point of attribute values.

2.1.3 Inconsistent Data Processing
Some inconsistencies would be present in recorded values these data can be corrected by analysing data correlation and retrieving data sources.

2.2 Data Integration
Consolidation of the different data should be taken care for dealing with heterogenous data and its redundancy. Improves the speed of data mining, integration and its accuracy through data integration.

2.2.1 Data Processing Heterogeneously
Electronic health records are collected from many EMR system combining which leads to semi-structured form of data. These further may create inconsistency in the data attributes. So, need to process this kind of data.

2.2.2 Data Processing Redundantly
When extracting data from another attribute is redundant should be clean up to maintain the consistency expression of attribute. For example, when patient shifting the hospital that patient’s record is redundant because takes the data of the available record of the patient for further treatment.

2.3 Data Reduction
Can reduce the dataset size which supports data mining in order to get efficiency. Reduction of dimension by reducing the random variables controls the dataset size.

2.4 Data Transformation
Converting dataset into unified structured for data mining is the data transformation. Includes data normalization and aggregation. Data transformation summarises the EMR data.

2.5 Protection through Privacy
EMR has a privacy issues of its vital data if misused. Can be protected through data protection protocols, access control methods and SDN technology is used [20], [21], [22], [23].

3. Text mining based EMR Information Extraction
Four stages of text mining are composed: Information retrieval, Information Extraction, Knowledge discovery and knowledge application as shown in the below figure 6.

Information retrieval is like classical data processing as data collection process. Information extraction pre-processes the collected data using predefined information. Knowledge discovery helps to discover new set of data from the EMR collected data. Knowledge application helps to pre-process unstructured form of data into structured form uses NLP techniques to extract the required data.

3.1 NER (Named Entity Recognition)
Different writings of medical terms are used in the medical field in order to encounter such words NER are used. The F-score average precision rate can be calculated by,

$$P = \frac{\text{the number of entities identified correctly}}{\text{the number of entities identified}}$$

$$R = \frac{\text{the number of entities identified correctly}}{\text{the number of entities present in the test set}}$$

$$F_{\text{score}} = \frac{P \times R}{P + R}$$

3.1.1 Rule Based NER Approach
Identified rules are valid in specific datasets from medical text [24]. Information extraction from EMR through Open source natural language processing system [25]. [26] Combines rule-based approach and machine learning approach to extract relevant data from clinical text.

3.1.2 Relation Extraction
According to evolution conference of I2B2 [27] EMR entities are divides into 3 categories, disease relation, disease relation and medical recommend and treatment of disease. EMR data of medical filed includes pattern-based medical learning and co-occurrence based [28]. [29] author presents convolution neural network (CNN) for synonyms and hyponyms extraction. Hybrid temporal extraction approach [30] proposed for combing the patient records ad random fields.

III. CONCLUSION
Briefs overview and detailed explanation about the EMR and its techniques. Text mining classifications and its types, Pre-processing techniques for EMR data in medical field

REFERENCES


