IJMRSETM

e-ISSN: 2395 - 7639

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING, TECHNOLOGY AND MANAGEMENT

Volume 10, Issue 6, June 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.580

+91 99405 72462 +9163819 07438 ijmrsetm@gmail.com www.ijmrsetm.com

International Journal of Multidisciplinary Research in Science, Engineering, Technology & Management (IJMRSETM)

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.580| A Monthly Double-Blind Peer Reviewed Journal |

| Volume 10, Issue 6, June 2023 |

Text Mining in Healthcare: Electronic Health Records

Miss. Priyanka Tadke, Mr Shripad S Bhide

Student, Department of M.C.A, P.E.S Modern College of Engineering, Pune, India

Assistant Professor, Department of M.C.A, P.E.S Modern College of Engineering, Pune, India

ABSTRACT: Electronic medical records (EMR) are frequently used by medical facilities to record several elements of a patient's condition, such as diagnostic data, carried out procedures, and treatment outcomes. For extensive examination, EMR data is regarded as important. However, EMR has some features that make it tough to directly undertake data mining and analysis, including diversity, incompleteness, redundancy, and privacy problems. To increase data quality and the outcomes of text mining, it becomes necessary to preprocess the raw EMR data.

Different preparation methods are required for various types of data. Traditional preprocessing techniques like data purification, integration, transformation, and reduction are frequently used with structured data. However, more intricate and difficult process techniques are required when working with semi-structured or unstructured data, like medical language that provides a wealth of health-related information.

Named-Entity Recognition and Relation Extraction are the two main tasks involved in the extraction of pertinent information from medical texts. NER focuses on locating and categorizing particular named entities—like illnesses, therapies, or drugs—within the text. The goal of RE, on the other hand, is to pinpoint the connections among the various concepts discussed in the text, offering insightful information about how diagnoses, treatments, and patient outcomes are related.

The paper focuses on the preparation of EMR data and gives a thorough overview of the key methods used. It also explores the medical applications that have been created by text mining, as well as the problems that still need to be solved and the areas for further study.

The preparation of EMR data is examined in the paper, emphasizing the importance of resolving its variety, incompleteness, redundancy, and privacy problems. It emphasizes the necessity of using various processing technologies depending on the kind of data, with structured data typically utilizing traditional preprocessing techniques and semi-structured or unstructured data necessitating more complex approaches like NER and RE. Additionally, the study covers text mining's applications in the field of healthcare and points out unresolved problems and areas for future research.

I. INTRODUCTION

Describe the typical electronic medical report data preprocessing technique and the well-known Chinese word segmented systems. We go over a few traditional methods for preparing EMR data. We discuss the current state of research on named entity recognition, relation extraction, and information extraction for electronic medical records (EMR) based on text mining.. We present three key areas of text mining-based application development.

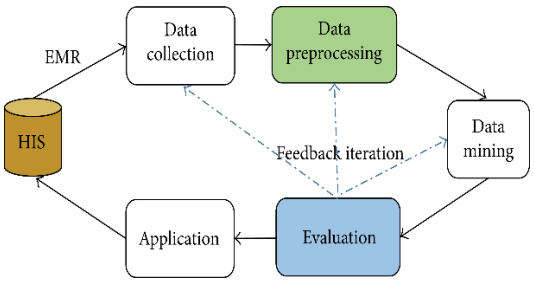

Fig: EMR Data Processing Flow

Government agencies and academic medical centers are the key organizations responsible for data gathering. Medical management and treatment program termination are more closely related to knowledge application, which is both aims and driving force behind text processing. Technologies used in data mining include clustering, association rules, regression, classification, and many more. We can only make a decision and create a predictive model after carefully

analysing the dataset. To evaluate a model, we must set up some tests for it in order to assess how well it performs. In addition, analysis and optimization are required for the discovered patterns and information. Data processing is therefore an involved iterative process that necessitates ongoing corrective feedback. We can only obtain a substantially superior knowledge model in this manner.It must be noted, nonetheless, that the EMR's data complexity makes it difficult to analyze data without first preprocessing it in a useful way. High-quality data increases the likelihood that the results will also be high-quality. According to statistics, preparation accounts for more than 60% of the entire data processing process' burden.

Data Processing on EMR

The data received from the EMR database is typically diverse, incomplete, and redundant, which has a significant impact on the final mining outcome. The EMR db is typically made up of a number of different data sources. It is compulsory to preprocess the EMR data in order to guarantee its accuracy, consistency, completeness, and privacy protection. Figure illustrates the data preprocessing process, which comprises, data transformation, privacy protection, data integration and data minimization. It should be noted that each preprocessing stage's tactics are interconnected. Therefore, it is important to choose the preprocessing techniques carefully, especially when working with medical data.

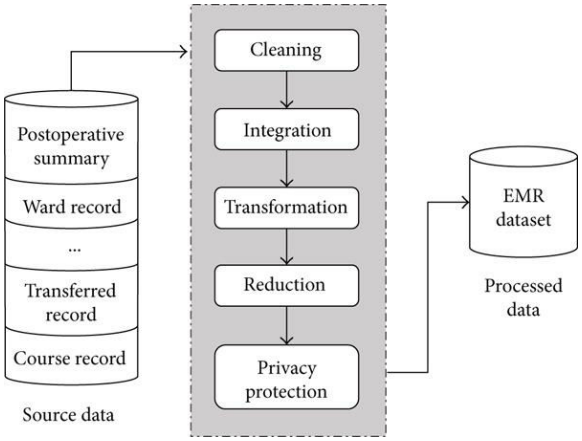

Fig: Data processing on EMR

II. DATA CLEANING

By quilting in failure, reducing dirty, and fixing facts inconsist, the incomplete, noisy, and inconsistent EMR data can be enhanced.

Data Cleansing: Some data properties may be lost during the EMR data collection process as a result of human error and system failure. There are numerous workarounds for default data. Missing data may be ignored, default values may be manually filled, attribute averages may be used to fill defaults with the most likely values, or we may obtain data from other sources.

The missing data is often disregarded unless it significantly affects how something is processed. For instance, while removing patient data, if procedure name absent, the data should neglected; but, if bed number details is absent, the fact can’t neglected. The defaults might need to be manually filled in if the dataset isn't very large.

When working with large sets that include more errors, it nonetheless fails. Due to its time and money requirements, this method is frequently not used. In the situation where the cost budget is minimal and the data distribution is uniform, the attribute averages could be utilised to fill the defaults. Regression, formal Bayesian methods, and decision tree induction are some machine learning techniques that can be utilised to determine the optimal value while working with default data. These techniques are nevertheless able to cope with data defaults better even though the prediction might in extreme circumstances indicate a pretty large variance. Furthermore, the data source should be obtained if another data source contains blank data.

Noise Processing

Noise in a data set refers to a value for an aberrant attribute, also called as an unlawful use occasionally. The patient's heat, for instance, is 27.8 degrees Celsius; her urine's pH is 3.26 (normal range: 5.00-9.00); and her specific gravity is 1.96 (normal range: 1.01-1.03). The various steps involved in processing noisy data include outlier analysis, regression, binning, and locating additional data sets.

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.580| A Monthly Double-Blind Peer Reviewed Journal |

| Volume 10, Issue 6, June 2023 |

Binning techniques smooth values of the ordinal data by looking at values surrounding data. The regression approach alters the dirty value by building a functional model reflects the value of the data attribute. Utilising clusters, outlier analysis can be produced using the clustering approach.
Even though all of the data points in a cluster have the same features, there is a significant difference in attribute values, the data point in various cluster.
Unreliable Data Process
There may variations between several sets of information or similar data, such as the reported values and measuring units. It is possible to correct the inconsistent data by looking at correlation between data and retrieve the information from other source.

III. DATA INTEGRATION

data unification phase, where data contained numerous data origin need to integrated, challenge is dealing with different data and repetitions. Data unification can boost data mining's precision and speed.

Heterogeneous Data Process:

Multiple Electronic Medical Records systems used to collect EMR data, and the diverse data sources will inevitably result in heterogeneous issue. Inconsistency data property, such attribute name and measurement units, serve as a primary representation of heterogeneous issues. For instance, the quantity of triglycerides can in mole/L, but occasionally mg/dl, and expression specific gravity of urine can specific gravity.

Redundant Data Processing
In other words, if attribute can take from another attribute, it should be cleaned up because it is redundant. Redundancy is primarily manifested in inconsistent attribute expression or repeated records of data attributes. For instance, when patient must transferred to another hospital for treatments, inspections would need to redone with new facility, leading to redundant and repeated medical records
Correlation analysis can find the majority of redundant data. When given two attributes, we can use the available data to determine how much one attribute is relevant to the other. The chi square test popular analysis technique for normal data.

IV. DATA REDUCTION

Data remove can shrink dataset size while still upholding data integrity, which helps facilitate and accelerate data mining. Every day, a sizable amount of EMR would be produced. Data reduction must be done in the current conditions. Data compression, quantity reduction, and dimension reduction are all examples of data reduction techniques. Dimension reduction is one of them and is a fairly common technique because it is simpler to implement with better results.In general, the dimension reduction strategy reduces the number of random variables to regulate the dataset quantity. Principal component analysis and wavelet transform are two techniques for reducing the size of the dataset by projecting the source data into it. A strategy of reducing the size of a dataset by identifying and eliminating unnecessary, weakly linked, or redundant attributes or dimensions is called attribute subset selection.

V. DATA TRANSFORMATION

A dataset is transformed into a standardized format through the process of data transformation so that it can be used for data mining. Techniques for transforming data include noise smoothing, data aggregation, and data normalization. Filtering and summarizing EMR data in accordance with the aim and goal of data mining is done using the data transformation approach. Data analysis can be more successful when there is directional, planned data aggregate.
In sequence to prevent the dependent of attributes of data on the measurement unit, data should normalize to make data fall in smaller common spaces, such [0,10], which are lot of intelligible. Normalization can be divided into three different categories: fractional scale normalization, min-max normalization, and zero-mean normalization. For classification algorithms using neural networks methods or distance measures, the normalization method works better.

Information Extracting of EMR Based Text mining:
The goal of text mining, also called as text data mining, o find implicit knowledge that hide in unstructure language. A wealth of helpful information can found in bio medical text that can be used to identify negative drug reactions or make early assessments of a patient's symptoms.
Figure shows the information extraction, information retrieval, , knowledge application and knowledge discovery phases of the text mining processes. Similar to conventional data processing, text mining is a method of data

processing. Data collection is similar to the process of locating required texts via information retrieval. To extract predetermined information, the data must first be prepared through information extraction. With the use of knowledge discovery, we can glean additional information from text.

Knowledge application is ultimate aim of putting unknowable fact infe from texts practice. Because medicine text mining predominantly uses semi-structured and unstructure texts the profession medical industry, traditionally preprocessing tools can't applied directly.

The fundamental strategies is converted semi-structured and unstructure text into structure data computers can understand using information extract and NLP technologies. For this processes to work, the named entity recognition and relation extract technologies are essential.

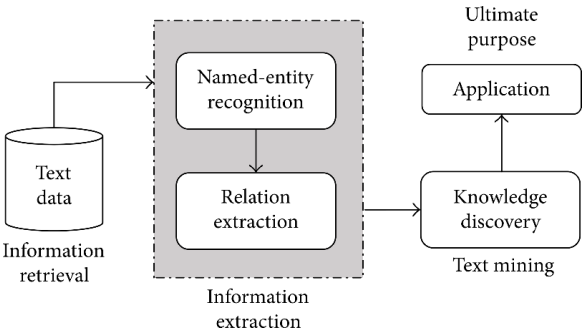

Fig: Progress of Text Mining

Applications:

Medication Decision Support and Diseases Risk Prediction:

Maintaining doctors' complete knowledge of the patients' treatments requires a lot of work. This is where medication decision support and diseases risk prediction come in. Medical decision support systems give medical professionals guidance on the best course of action for treating symptoms based on logical information. If the mechanism can be improved upon and put to use, it will be crucial for diagnosing illnesses by medical professionals, especially those with less clinical experience.

Drug Reaction Detection: Medical data mining technology plays a supporting role in disease diagnosis and therapy by immediately determining the disease's medical trajectory through time and studying its natural history. Medical data mining tools, for instance, can precisely identify the risk variables in specific regions with a high prevalence of epidemic diseases. Additionally, after the creation of new pharmaceuticals, significant resources and time would expended research their effect, but medication data mining technologies identify adverse drug occurrences an efficient manner.

Integration with Electronic Health Records and Clinical Decision Support Systems:

Strengthening the integration of text mining capabilities with EHRs and clinical decision support systems will provide seamless access to textual data and support real-time clinical decision-making. Text mining algorithms can assist in extracting relevant information from EHRs and provide decision support to healthcare professionals at the point of care.

VI. FUTURE ENHANCEMENT

The use of text-mining in EHRs for patient follow-up and outcome assessment in clinical trials. They also suggest that the accuracy of text-mining could be improved by using natural language processing and machine learning techniques.

Integration with Electronic Health Records and Clinical Decision Support Systems:

Strengthening the integration of text mining capabilities with EHRs and clinical decision support systems will provide seamless access to textual data and support real-time clinical decision-making. Text mining algorithms can assist in extracting relevant information from EHRs and provide decision support to

Knowledge Graphs and Semantic Linking: Leveraging knowledge graphs and semantic linking techniques can enhance the representation and understanding of healthcare concepts and relationships within text mining applications. Building comprehensive healthcare knowledge graphs and utilizing ontologies can improve information retrieval, concept extraction, and knowledge discovery from textual data.

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.580| A Monthly Double-Blind Peer Reviewed Journal |

| Volume 10, Issue 6, June 2023 |

VII. CONCLUSION

Clinical data from electronic medical records must be extracted before being used for clinical research or quality assurance. The data that is extracted from EMRs may be unstructured, lacking, and rife with inaccuracies. The data was cleaned up and made available for statistical analysis using conventional database administration methods.
Laboratory, medication, and diagnostic data could be used immediately, but vital sign data required extensive searching before it could be used. As long as they are screened and controlled in models, it is possible to use current patient data in the various electronic medical records systems for quality control and clinical research. This kind of cleaned data can show us where adjustments are needed and can teach us a lot about the standard of care of healthcare systems.

REFERENCES

1. Data processing and text mining technologies on electronic medical reports(EMRs).
2. https://typeset.io/papers/data-extraction-from-a-semi-structured-electronic-medical-3kjx95ssbs
3. https://ieeexplore.ieee.org/document/9593998/authors#authors
4. https://innovate.ieee.org/ieee-journals-conference-proceedings/
5. text-mining-in-electronic-healthcare-records-can-be-used-as-3gxwonz7td
6. (pdf).researchgate.net/publication/366921083_Design_and_Implementation_
7. google Wikipedia images(diagram)
8. typeset.io/papers/privacy-preserving-process-mining-in-healthcare-4g62tfx1vl
9. https://typeset.io/explore/papers/text-mining-in-healthcare-applications-and-opportunities-3d65pskbd5
10. https://dl.acm.org/doi/abs/10.1145/1089815.1089824
11. Text-mining in electronic healthcare records can be used as efficient tool for screening and data collection in cardiovascular trials: a multicenter validation study

Impact Factor: 7.580

INTERNATIONAL JOURNAL
OF MULTIDISCIPLINARY RESEARCH
IN SCIENCE, ENGINEERING, TECHNOLOGY AND MANAGEMENT

+91 99405 72462 +91 63819 07438 ijmrsetm@gmail.com

www.ijmrsetm.com