IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

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].

For partitioning the unstructured data uses word segmentation and storing the result in the database. There are many Segmentation tools such as SCWS, PhpanAlysis etc.

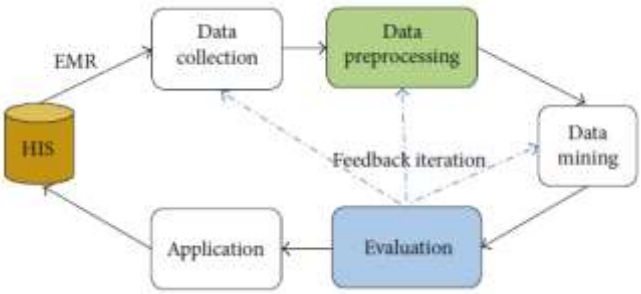

Figure 1: EMR data flow [2]

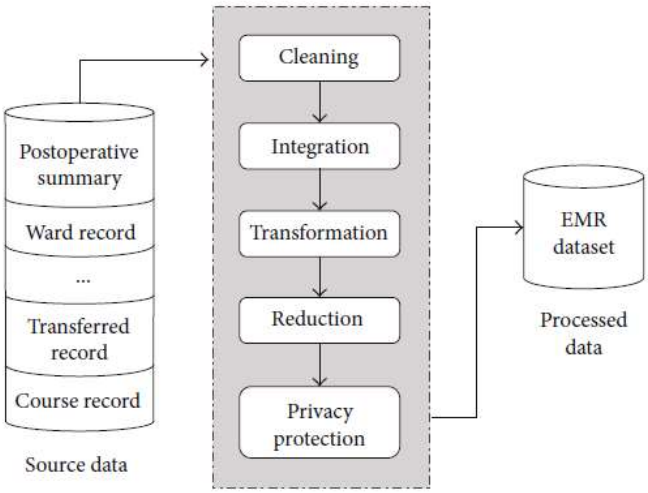

Figure 2: EMR Data pre-processing [2]

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, postprocessing. 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.

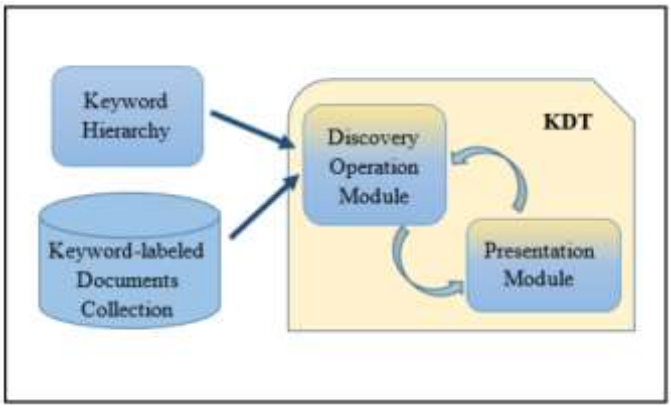

Figure 3: System architecture of KDA [10]

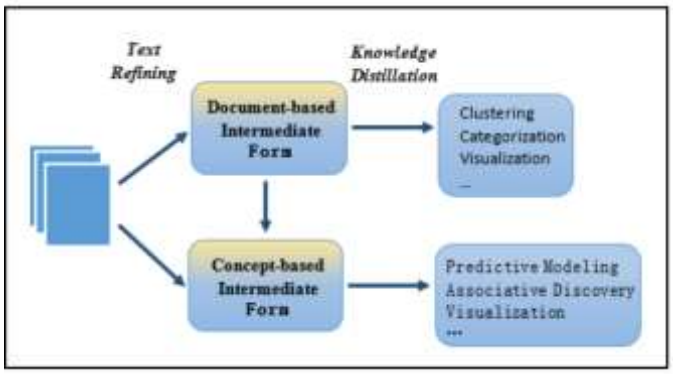

Figure 4: Framework of Text [11]

In figure 5 the model can find efficiently significant matching by providing appropriate semantics of sentences from the text.

In text mining classification which consists of Text categorization, Text clustering and text association and analysis.

1.1 Text categorization which determines text content of its classification. Due to increase in number of text data text categorization is developed. In Text disambiguation may occur can view this word occurrence and categorize its document. [4] , [5]. Uses Bayesian decision models as class for doing this task for information retrieval.

Sentiment analysis which calculates the similar words in the phrase and this technique is important in text categorization [6]. Bag of words approach to classify sentimental analysis.

1.2 Text Clustering is a process consisting of unsupervised form through which objects are classified into groups of predefined categories. It is based on cluster hypothesis used in data analysis, data mining, pattern classification and image segmentation [12], [13], [14]. 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.

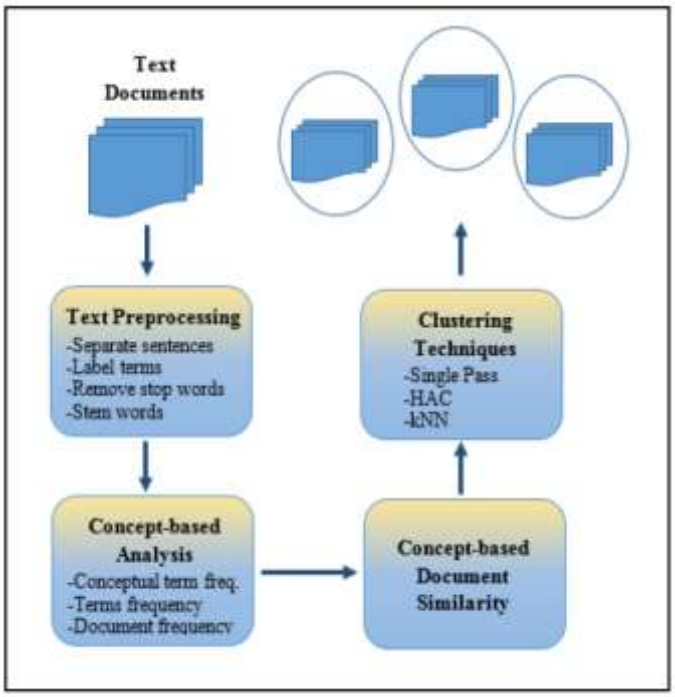

Figure 5: Model mining system [8], [9]

.

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],

$$Support\left(W_iW_j\right)=\frac{Support\ count\ of\ W_iW_j}{Total\ number\ of\ documents\ D} \quad (1)$$

$$Confidence\left(W_i \setminus W_j\right)=\frac{Support\left(W_iW_j\right)}{Support\left(W_i\right)} \quad (2)$$

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 process 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.

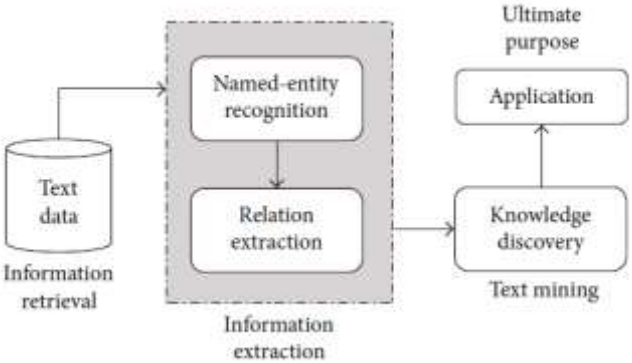

Figure 6: Text Mining Progress [3]

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,

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

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

$$F\text{-score} = \frac{P * R * 2}{P + R}. \quad (1)$$

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. In EMR data of medical filed includes pattern-based machine 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

[1] W. Sun, Z. Cai, F. Liu, S. Fang, and G. Wang, "A survey of data mining technology on electronic medical records," in 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom), pp. 1–6, Dalian, China, 2017.

[2] Wencheng Sun, Zhiping Cai ,Yangyang Li, Fang Liu, Shengqun Fang, and Guoyan Wang, "Data Processing and Text Mining Technologies on Electronic Medical Records: A Review", Volume 2018, Article ID 4302425, April 2018.

[3] Yu Zhang, Mengdong Chen and Lianzhong Liu, "A Review on Text mining" IEEE 2015.

[4] Gale, W. A., Church, K. W., and Yarowsky, D. (1993). "A Method for Disambiguating Word Senses in a Large Corpus." Computers and the Humanities 26(5): 415–439.

[5] Escudero, Gerard, L. Marquez, and G. Rigau. "Boosting Applied to Word Sense Disambiguation." IN PROCEEDINGS OF THE 12TH EUROPEAN CONFERENCE ON MACHINE LEARNING2000:129--141.

[6] Dun LI, Fu-Yuan CAO, Yuan-Da CAO, Yue-Liang WAN. "Text Sentiment Classification Based on Phrase Patterns." Computer Science. 35.4(2008):132-134. DOI:10.3969/j.issn.1002-137X.2008.04.037.

[7] Pang, Bo, L. Lee, and S. Vaithyanathan. "Thumbs up? Sentiment Classification using Machine Learning Techniques." Proceedings of Emnlp(2002):79--86.
[8] Shehata, S., F. Karray, and M. Kamel. "Enhancing Text Clustering Using Concept-based Mining Model." IEEE 13th International Conference on Data MiningIEEE Computer Society, 2006:1043-1048.
[9] Shehata, Shady, F. Karray, and M. S. Kamel. "An Efficient Concept-Based Mining Model for Enhancing Text Clustering." IEEE Transactions on Knowledge & Data Engineering22.10(2010):1360-1371.
[10] Feldman, Ronen, I. Dagan, and H. Hirsh. "Mining Text Using Keyword Distributions." Journal of Intelligent Information Systems10.3(1998):281-300.
[11] Tan, Ah Hwee, et al. "Text Mining: The state of the art and the challenges." Proceedings of the Pakdd Workshop on Knowledge Disocovery from Advanced Databases(2000):65--70.
[12] Rijsbergen, Van. C.J. "Information Retrieval." 14th International Symposium on Methodologies for Intelligent Systems. Volume 2871., Maebashi City, Japan, LNCS, Springer-Verlag12.2-3(1989):95.
[13] Douglass R. Cutting, David R. Karger, Jan O. Pedersen, John W. Tukey "Scatter/Gather: a cluster-based approach to browsing large document collections." Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval ACM, 1992:318--329.
[14] Oren Zamir, Oren Etzioni, Omid Madani, Richard M. Karp. "Fast and Intuitive clustering of Web documents." Proceedings of International Conference on Knowledge Discovery & Data Mining(1997):287--290.
[15] Clifton, Chris, and R. Cooley. TopCat: Data Mining for Topic Identification in a Text Corpus. Principles of Data Mining and Knowledge DiscoverySpringer Berlin Heidelberg, 1999:174-183.
[16] E.-H. S. Han, G. Karypis, and V. Kumar, "Clustering Based on Association Rule Hypergraphs," Proc. SIGMOD'97 Workshop Research Issues in Data Mining and Knowledge Discovery, 1997.
[17] Agrawal, Rakesh, T. Imieliński, and A. Swami. "Mining Association Rules Between Sets Of Items In Large Databases." SIGMOD '93 Proceedings of the 1993 ACM SIGMOD international conference on Management of data1993:207--216.
[18] Mahgoub, Hany, et al. "A Text Mining Technique Using Association Rules Extraction." International Journal of Computational Intelligence1(2008):21.
[19] Mei, Qiaozhu, and C. X. Zhai. "Discovering evolutionary theme patterns from text: an exploration of temporal text mining." Proceedings of Kdd '(2005):198-207.
[20] F. Liu and T. Li, "A clustering K-anonymity privacy-preserving method for wearable IoT devices," Security and Communication Networks, vol. 2018, Article ID 4945152, 8 pages, 2018.
[21] H. Zhang, Z. Cai, Q. Liu, Q. Xiao, Y. Li, and C. F. Cheang, "A survey on security-aware measurement in SDN," Security and Communication Networks, vol. 2018, Article ID 2459154, 23 pages, 2018.
[22] J. Xia, Z. Cai, and M. Xu, "An active defense solution for ARP spoofing in OpenFlow network," Chinese Journal of Electronics, vol. 3, 2018.
[23] Y. Li, Z. Cai, and H. Xu, "LLMP: exploiting LLDP for latency measurement in software-defined data center networks," Journal of Computer Science and Technology, vol. 33, no. 2, pp. 277–285, 2018.
[24] D. Rebholz-Schuhmann, A. Yepes, C. Li et al., "Assessment of NER solutions against the first and second CALBC silver standard corpus," Journal of Biomedical Semantics, vol. 2, article S11, Supplement 5, 2011.
[25] G. K. Savova, J. J. Masanz, P. V. Ogren et al., "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications," Journal of the American Medical Informatics Association, vol. 17, no. 5, pp. 507–513, 2010.
[26] A. Kovačević, A. Dehghan, M. Filannino, J. A. Keane, and G. Nenadic, "Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives," Journal of the American Medical Informatics Association, vol. 20, no. 5, pp. 859–866, 2013.
[27] M. Jiang, Y. Chen, M. Liu et al., "A study of machine-learningbased approaches to extract clinical entities and their assertions from discharge summaries," Journal of the American Medical Informatics Association, vol. 18, no. 5, pp. 601–606, 2011.
[28] Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall, "2010 i2b2/ VA challenge on concepts, assertions, and relations in clinical text," Journal of the American Medical Informatics Association, vol. 18, no. 5, pp. 552–556, 2011.
[29] R. Jelier, G. Jenster, L. C. J. Dorssers et al., "Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes," Bioinformatics, vol. 21, no. 9, pp. 2049–2058, 2005.
[30] J. Y. Lee, F. Dernoncourt, and P. Szolovits, "MIT at SemEval- 2017 task 10: relation extraction with convolutional neural networks," 2017, http://arxiv.org/abs/1704.01523.
[31] Y. L. Yang, P. T. Lai, and T. H. Tsai, "A hybrid system for temporal relation extraction from discharge summaries," in Technologies and Applications of Artificial Intelligence, pp. 379–386, Springer, Cham, 2014.
[32] A. Nikfarjam, E. Emadzadeh, and G. Gonzalez, "Towards generating a patient's timeline: extracting temporal relationships from clinical notes," Journal of Biomedical Informatics, vol. 46, pp. S40–S47, 2013.
[33] J.-W. Seol, W. Yi, J. Choi, and K. S. Lee, "Causality patterns and machine learning for the extraction of problem-action relations in discharge summaries," International Journal of Medical Informatics, vol. 98, pp. 1–12, 2017.
[34] A. Henriksson, M. Kvist, H. Dalianis, and M. Duneld, "Identifying adverse drug event information in clinical notes with distributional semantic representations of context," Journal of Biomedical Informatics, vol. 57, pp. 333–349, 2015.
[35] W. Sun, Z. Cai, F. Liu, S. Fang, G. Wang, and Y. Li, "Security and privacy in the medical internet of things," Security and Communication Networks, vol. 2018, Article ID 5978636, 9 pages, 2018.
[36] G. O. Barnett, J. J. Cimino, J. A. Hupp, and E. P. Hoffer, "DXplain. An evolving diagnostic decision-support system," JAMA, vol. 258, no. 1, pp. 67–74, 1987.
[37] M. J. Feldman and G. Octo Barnett, "An approach to evaluating the accuracy of DXplain," Computer Methods and Programs in Biomedicine, vol. 35, no. 4, pp. 261–266, 1991.
[38] Y. Liu, L. Wei, Z. Yao, and X. L. Fei, "The practice and experience of emergency information system construction," China Digital Medicine, vol. 11, no. 5, pp. 53–55, 2016.
[39] X. Gao, X. Yan, Y. Zhang, Q. Chen, and H. P. Hu, "Demand analysis of decision support system of grass-roots health," Chinese General Practice, vol. 19, no. 22, pp. 2636–2639, 2016.
[40] W. Shao, Y. Wang, G. T. Yan, and Y. Zhao, "Research on construction of a clinical decision making support system," China Medical Devices, vol. 31, no. 8, pp. 87-88, 2016.
[41] R. K. Lomotey and R. Deters, "Efficient mobile services consumption in mHealth," in Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM '13, pp. 982–989, Niagara, ON, Canada, 2013.
[42] K. I. Henrik Bostrom, "Predicting adverse drug events using heterogeneous event sequences," in 2016 IEEE International Conference on Healthcare Informatics (ICHI), pp. 356–362, Chicago, IL, USA, 2016.