Research article Available online www.ijsrr.org ISSN: 2279–0543

International Journal of Scientific Research and Reviews

Textual Analysis of Unstructured Medical Data

S Nayana [1*] and S Pushpalatha[2]

[1]*PG scholar, Department of ISE, Dr. AIT, Bangalore, Karnataka, India*
[2] *Assistant Professor, Department of ISE, Dr. AIT, Bangalore, Karnataka, India*

ABSTRACT

Textual analysis helps researchers to gather information which depends up on how humans feel the globe. One of the approach for information assembling process is textual analysis for those analyst who need to know the methods in which supporter of diverse cultures and sub cultures make meaning of who and how they will suitable to the globe. Textual analysis is one used by people performing in cultural studies, media studies, sociology and philosophy and mass communication. Interpreting texts (entertainment fields such as, clothes, television programmers, magazines) to attempt and gain ways in which, in specific cultures at certain times and during this time sensing of world is done by analyst or scientist. And notably, by watching the different methods attainable to interpret reality, also know the own cultures best to can start up to study feasible or restriction and supremacy of own sense conducting practices. The same concept used to read a lung cancer report. In which the report is an input for machine and that machine convert any file to image file ie; png, jpeg etc… the content of file is analyzed then the meaningful words are extracted from it. Based on the extracted data the next steps are analyzed. After analyzing medical terms in report the machine will suggest a next steps such as about doctors, hospitals, treatment, stages etc..

KEYWORDS: Textual Analysis, Assembling, Extracting Data, Analyzing

***Corresponding author**

S Nayana

Department of Computer Networks and Engineering,

Dr Ambedkar Institute of Technology,

Banglore-56, Karnataka, INDIA.

Email: snayana205@gmail.com

INTRODUCTION

Textual analysis[12] process is to narrate and also to break in features of a videotaped or perceptible message. The use and necessary is to narrate the information, formation and tasks of the messages within the texts. Textual analysis may involve consideration of audience, heed to the visual, written and verbal language, style and design elements, assessing the text for what it is trying to do, and response. Inspiration for textual analysis is narrating the text matter, form also tasks of text within content. Significant things that need to analyze in textual analysis is that selection of the ways of texts to analyzed, obtain acceptable messages and deciding particular fit method to employing survey them. Texts can be discriminated into two Reproduction of conversation and Results of conversation. Authentic conversations analyze contains explore conversation implant. Text accession is key which representative of the texts hand pick because sample text is typically used. Other case is predicting how absolute with detailing texts in sequence to perform a sound analysis. To report the structure, content and functions of the messages included within messages using the textual analysis is very much needed. Textual analysis may include, cogitation of audience, observation to the visual, written and verbal language, formation of elements and design elements, assessing the text for what it is venture to do and response. The same concept is used to read a unstructured medical report. In which the report is an input for a machine and that machine convert any file to image file ie; png, jpeg etc… the content of file is analyzed then the meaning full words are extracted from it. Based on the extracted data the next steps are analyzed. After analyzing medical terms in report the machine will suggest a next steps such as about hospitals, stages, anatomy, medications, medical condition, test treatment procedure, protected health information, etc.. Different textual analysis methods are rhetorical criticism, content analysis, interaction analysis, and performance studies. Rhetorical Criticism represents negative overtone timely used to grand, expressive, ranting or verbose discourse. Four steps for performing rhetorical criticism process are Selecting texts to study, Selecting a particular ways of rhetorical criticism, Examine text following with method chosen and Penning the difficult essay. Content Analysis recognizes, narrate and examine the occurrences of particular texts and text features implanted in messages content analysis is used. Qualitative content analysis is where analysts are more focused on attracting meanings connected with texts compared with count of times text variables occur. Quantitative content analysis is systematic in nature, procedure steps are used to reply analyst's queries and test hypothesis. Interaction Analysis is interplay same way the composite achievement which needs much more information depending on bit of discrete conveyers also capacity to harmonize reaction with others too. Performance study procedure

determines intermediate engaging to one's own and other's esthetic conveyance via means of performance. Analysts explicate messages as the way of query which sanctions and consultation about performances for expound stylish moneyed about messages. Natural Language Processing[11] examines semantic data, more regularly through the formation of textual data for ex, record using calibration ways. Natural language processing aiming to develop and representing of messages which puts form to the unformed natural language utilizing semantic insight. Natural Language Processing utilized in systems biology to build the procedure which combines the information collected from the documentation connecting to other origin of biological information.

RELATED WORK

In this, studies have been done on textual analysis applying on different purposes. Portable Document Format (PDF)[1,2] is the better way to publish information. PDF document is formless and given permission only to read not for modification purpose. Extraction task is much difficult in PDF document. Mechanically analyzing and recognition of PDF document shape specifically paragraph and tabular area is essential for extracting applicable information exactly for using in other domain applications. Plagiarism prevention and plagiarism detection is main consideration[3,4]. "Plagiarism as larceny of intellectual property" has been all around as long as human proposes work of art and research. Plagiarism can be elucidated as turning of someone's work as own without reference to actual source. It is very hard to oppose that problem of plagiarism moving more and more actual society's knowledge about plagiarism is constantly stepping up and plagiarism's problem actually starts to pull society's attention. Textual information gives chief semantic guide in video content analysis[7,8]. Procedure for detection and presenting text in video segments, video segmentation has seven steps: channel separation, image enhancement, edge detection, edge filtering, character detection, text box detection and text line detection. Results show that method can be applied to English as well as non English text with accuracy and recall of 85%. Plagiarism can have different natures, fluctuating from duplicating texts to designating ideas, without having credit to its originator[5,6]. Content presents taxonomy of plagiarism gives main differences between literal plagiarism and intelligent plagiarism, in the plagiarist's behavioral point of view. Algorithm is built for recognizing texts in images and video frames[9,10]. To recognize texts in images 3 steps are required: edge detection, text candidate detection and text refinement detection. Firstly, application of edge detection for getting four edge maps in horizontal, vertical, up right, and up left direction. Secondly, the features are extricated from four edge maps to give the texture property of text. Then k means algorithm is used to discern the beginning text candidates.

Endingly, the text areas are detected by the empirical instruction analysis and purified through profile analysis. Experimental effect reveals that built approach could efficiently use as a mechanized text discover system, which is robust for font size, font color, background complexity and language.

TEXTUAL ANALYSIS PROCESS

The major aim is to convert the unstructured medical into structured data using textual analysis. The process includes varieties of steps to convert unstructured to structured data. Figure 1 shows different process to convert the unstructured medical data into structured medical data which is every easy to analyse by doctor. Once the report is uploaded the report undergoes four processes ie; segmentation, tokenization, parts of speech and lemmatization. Sentence segmentation is where known as breaking of sentences also process of grouping hand booked text into meaningful units, such as words, topic or sentence. Natural language processing tools commonly need input to be split into sentences. Tokenization process contains of substituting sensitive data with distinctive identification symbols that keep all the required data about the data without compromising its security. The token is a referral that maps back to sensitive data through a tokenization system. Mapping from real data to a token system uses different methods which render tokens infeasible to bounce in the non appearance of tokenization system. Tokenization system permits data processing applications with the dominance and interfaces to request tokens or detokenize back to sensitive data.

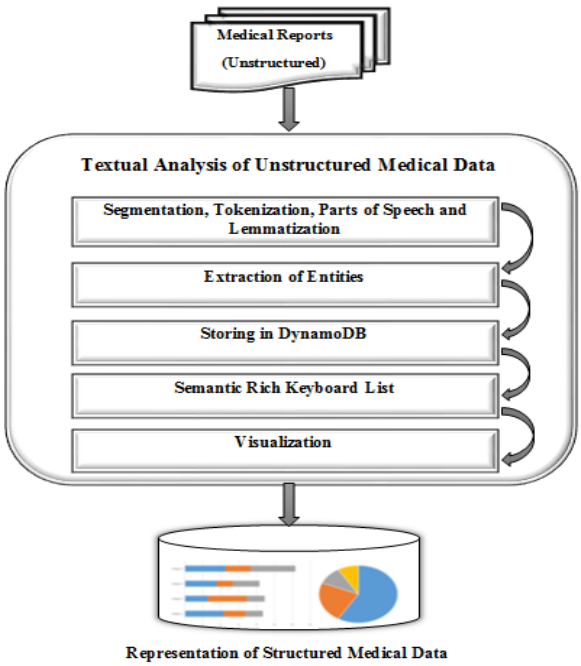

Figure1. Textual Analysis Process

Parts of speech tagging is process of making up a word in text as corresponding to particular part of speech depending on both its definition and its context. Parts of speech

tagging is now done in the context of computational linguistics, using of algorithms which associate discrete terms. Lemmatization in linguistics is the process of assemble together the curved forms of a word, so it can be examined as single item. In computational linguistics, lemmatization is algorithm process which determines the lemma of word depending on its intended meaning. It depends on properly recognizing the intended parts of speech and meaning of word in a sentence as well as within the larger context surrounding that sentence such as neighbouring sentences or even an entire document. The next process which the report undergoes is extraction of entities, relationships and entity trial automatically ie; medication, medical condition, anatomy, test treatment procedure and protected health information by amazon comprehend medical. The data which is extracted is called as raw data is stored in structured form ie; dynamo DB. Making it to analyse easily tag and identify commonalities with in clinical trials. Adding of semantic rich keywords list in to elastic search and indexing of medical records using of list. Visualization of structure data is in the form of pie chart and bar chart. There is a separate portal to know the information about the medication, medical condition, anatomy, test treatment procedure and protected health information. The major problem pops in a large font size.

IMPLEMENTATION

Amazon Comprehend Medical is a natural language processing which makes easy usage of machine learning for fragmenting fitting medical information from unstructured text. Simple API call is enough to access Amazon Comprehend Medical, no knowledge of machine learning, no complex rules, no models to be trained. To extract compound medical information from unstructured text, Amazon Comprehend Medical is used. Using Amazon Comprehend Medical can quickly and precisely collect the data of patient, such as medical condition, medication, dosage, strength and frequency from a variety of sources like doctor's notes, clinical trial reports and patient health records. Figure 3 shows how the information extracted through the unstructured medical report. In Amazon Comprehend there are no servers to furnish because it is fully managed so no usage of machine learning models to develop, train or deploy. Can make use of extricated medical information and their bond to build applications for use cases like clinical decision support, revenue cycle management (medical coding), and clinical trial management. One of the chief ways to upgrade patient care and accelerate clinical research is by grasping and inspecting the insights and relationships that are trapped in free form medical text, including hospital admission notes and a patient's medical history. This is attained by writing and controlling a group of customized instructions for natural language processing software, which are complex to build, time draining to maintain and frangible.

Machine learning can update all that with models that can firmly understand the medical information in unstructured data, recognizing understandable relationships and improve over time. Amazon Comprehend Medical is HIPAA (Health Insurance Portability and Accountability Act) desirable and can rapidly recognize Protected Health Information (PHI), such as name, age and medical record number, can also be used to create applications that assured process, maintain and transmit PHI. The kibana tool is used to represent the PHI, anatomy, medications, medical condition, test treatment procedure in the form of dash board.

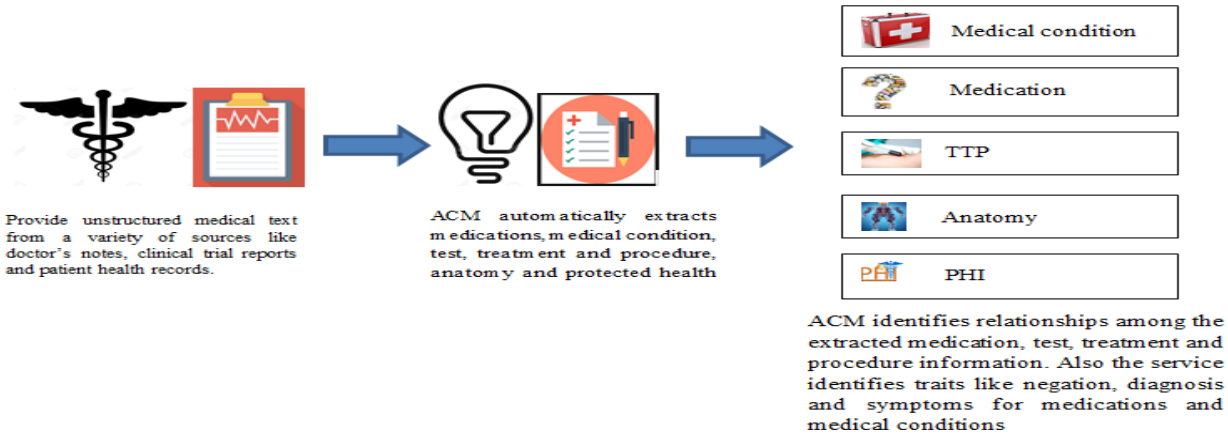

Figure 3. ACM Process for Extracting Medical Terms

CONCLUSION

To build a smart medical report reader, capable of sorting, analyzing and predicting the unstructured clinical text data. Converting unformed report into formed report representing the connection between the words extracted from report ie; json files. The report is an input for a machine and that machine convert any file to image file ie; png, jpeg etc… the content of file is analyzed then the meaning full words are extracted from it. Based on the extracted data the next steps are analyzed. After analyzing medical terms in report the machine will suggest a next steps such as about doctors, hospitals, treatment, stages etc.. This reduces the risk of emergency cases and emergency patients left without having appointments. Medical history of patients is well maintained in structured form. To reduce the time gap of appointments of emergency patients fast and efficient test analyzing machine is required to represent data in structure form, so that doctor would go through structured data and schedule next appointment in case of emergency. In future this Application can be enhanced and used across all verticals like X-Rays, CT scan, MRI scan using IP. System to predict the health information based on the food habits and lifestyle. To work on other disease reports and report the doctor. Future enhancement also helps to report patient about next scheduled appointments.

REFERENCES

1. Zanibbi, R., Blostein, D., Cordy, J.R.: A Survey of Table Recognition. "Models, Observations, Transformations, and Inferences. International Journal on Document Analysis and Recognition", 2004; 7: 1-16
2. Rosmayati, M., Abdul Razak, H., Zulaiha, A.O., Noor Maizura, M.N. "Ontological based for Supporting Multi Criteria Decision Making" In, Desheng Wen, Zhou, J. (eds.): 2010 2nd IEEE International Conference on Information Management and Engineering, IEEE Press, Chengdu, China. 2010; 1: 214-217
3. Bao, J.P., J.Y. Shen, H.Y. Liu, X.D. Liu. "A fast document copy detection model. Soft Computing - A Fusion of Foundations, Methodologies and Applications",2006; 10(1): 41 – 46.
4. G. A. Miller, "Word Net: A lexical database for English," *Commun. ACM*, 1995; 38: 39–41.
5. Neill, C.J., G. Shanmuganthan. "A Web – enabled plagiarism detection tool. IT Professional", 2004; 6(5): 19 – 23.
6. K. J. Ottenstein, "An algorithmic approach to the detection and prevention of plagiarism," *SIGCSE Bull.*,1997; 8(4): 30–41.
7. S. Pfeiffer, R. Lienhart, S. Fischer, and W. Effelsberg, "Abstracting Digital Moves Automatically" Journal on Visual Communications and Image Representation, 1996; 7(4): 345-353.
8. K.V. Mardia and T.J. Hainsworth, "A Spatial Thresholding Method for Image Segmentation," IEEE Transa. on Pattern Analysis and Machine Intelligence,1988; 10: 919-927.
9. M. Christel, T. Kanade, M. Mauldin, R. Reddy, M. Sirbu, S. Stevens and H. Wactlar, "Informedia Digital Video Library," Comm. of the ACM,1995; 38(4): 57-58.
10. Jiang Wu, Shao Lin Qu, Qing Zhuo, and Wen Yuan Wang, "Automatic text detection in complex color image", Machine Learning and Cybernetics, 2002. Proceedings. 2002 International Conference on, Nov. 2002; 3(4-5): 1167 – 1171.
11. L. Zhou and D. Zhang, "NLPIR: a theoretical framework for applying natural language processing to information retrieval," J. Am. Soc. Inf. Sci. Technol, 2003; 54(2): 115–123.
12. F. Sebastiani. Machine learning in automated text categorization. ACM Comput. Surv.,March 2002; 34(1): 1–47.