International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 09, September-2015

Automated Healthcare System using Text Mining: A Survey

Ashwin P. Nikam, Abhishek A. Patil, Rohan N. Panchal, Ajinkya K. Ghodekar
Dept of Computer Engineering
NBN Sinhgad School of Engineering
Pune, India

***Abstract*—Statistics from the World Health Organization (WHO) have recently shown that the disease and mortality rates of the entire world population greatly depend on the quality of healthcare access. Access to proper healthcare in each and every part of the world is a major hurdle we have yet to overcome. The attention and focus towards providing reliable healthcare access to even the most remote areas of the world is increasing day by day and has become a major issue. This paper presents a feasible solution in the form of automated medical diagnosis and treatment of diseases for those masses that are deprived the access to experienced professional healthcare. It focuses more on the diagnosis part and the challenges which should be overcome while achieving this feat. Text mining approach has been used in this solution for realizing an Automated Healthcare System which would be ubiquitous, thus providing proper and reliable healthcare.**

Keywords—Medical Diagnosis, Information Retrieval, Machine Learning, Text Mining

I. INTRODUCTION

The problem of medical diagnosis has been attracting substantial amount of attention over a significant period of over four decades. However, the problem hasn't been solved yet to the desired level. A large section of the world population is still deprived of access to proper healthcare. Proper healthcare may refer to prevention and cure of diseases or medical conditions wherein, prevention may consist of the monitoring part and cure may consist of diagnosing the disease. This paper focuses more towards the diagnosis part. The ultimate aim is to provide reliable medical diagnosis to underprivileged people without access to experienced and reliable physicians. There have been some developments in this direction. MYCIN, a popular solution to this approach was developed in 1970 as an expert system used for diagnosing blood clotting diseases although it had around 600 rules. We discuss a solution in this paper which requires processing of thousands of documents as vectors with more than a thousand dimensions. Focus is more towards personalization of big data, interpreting stable patterns and building augmented intelligence with unstructured data. This paper extends these ideas and uses text mining techniques for solving a critical problem impacting a major chunk of population – that of medical diagnosis.

This paper discusses a feasible solution for medical diagnosis which is a) based on past diagnosis b) general and adequate enough to be used by masses c) considers the problem from a text mining perspective.

The goal of this paper is quite ambitious and unique and uses Information Retrieval/ Text Mining approach to reach this goal.

The paper is organized categorically for better understanding of the entire concept. Section 2 contains the literature survey. Section 3 discusses the vision of automated medical diagnosis which helps in the understanding of this concept. Section 4 explains how such automated medical diagnosis can be practically realized. Section 5 explains all the possible hurdles and challenges which need to be overcome and Section 6 concludes the paper with future directions.

II. RELATED WORK

Significant amount of research and work has been dedicated to developing projects similar to the one being discussed in this paper. Some projects which can be related to our vision and contribute in realizing our objectives have been cited in this literature survey. Scheuermann et al. present a framework in [1] to represent ontologies of the diseases and diagnoses and relationships between them. They use the ontology approach to help with medical diagnosis. The other category of papers consists of work related to the machine learning realm. One such is [2] where three different Machine Learning algorithms are used to predict the onset of type II diabetes and compare the results. In [3], the authors apply machine learning (SVM-Adaboost) to predict addiction to smoking based on brain features as an alternative to the traditional techniques in addiction related neurobiology. Authors of [4] show how linear machine learning algorithms such as Logistic Regression takes just minutes to do predictions, while other powerful methods like Support Vector Machines (SVM)and Random Forests take hours or even days. They use it to predict mortality of men diagnosed with prostate cancer.

Healthcare delivery via teleconsulting, though requiring manual expertise for diagnosis is an important step in realizing our vision. In [5], the authors describe their working model of providing teleconsulting to the underprivileged. The authors of [6] use a naive Bayes-based machine learning system to associate phrases in clinical notes with medical concepts. Annotation techniques such as this can help our discussion because our text mining approach relies on clinical notes made by different physicians who can possibly use varying phrases to describe the same concept. So, a mapping between the phrases in the notes and medical concepts can result in better accuracy in mining the notes.

III. AUTOMATED HEALTHCARE

The field of automated healthcare consists of a broad scheme of things and a lot of projects have been developed related to this field. As shown in Figure 1. [7], Automated Healthcare mostly involves people using wearables who can communicate with machines in the cloud using their handheld devices [8]. Such machines in the cloud have enormous computing power and handheld devices like a Smartphone can be used to communicate with them. Any light weight protocol such as Bluetooth can be used to communicate between the wearable and the handheld device while Wi-Fi or data networks like 3G/4G can be used for communication between the handheld device and the cloud.

The machines in cloud have enormous computing power which can use the data obtained from the wearables to learn about the person's current condition and diagnose any abnormalities in his health. [8]

Apart from the information obtained from the wearables, valuable insights and further information about the conditions of a person may be obtained through a Q & A session which may consist of voice or text and may include pictures and videos. If the information received from the wearables is unclear or not cogent then such Q& A sessions can be initiated. Processing of the information obtained from the images and videos and relating it with the information in the discharge sheets is a complicated process which is still in the idea stage. In the long run we can also contemplate and consider the possibility of appointing a physician for routinely checking the diagnosis decisions and information available in the corpus related to complex cases where knowledge isn't readily available in the corpus. [7]

Using a cloud infrastructure to store health related information is advantageous because the data can be aggregated in such ways that it can't be personally identifiable. Such data is used to perform analytics, gain some knowledge from the data and draw conclusions which can be applied to patients suffering from similar conditions [9]. Diagnosis can generally be referred to as identifying the symptoms and accurately predicting the disease which the patient is suffering from.

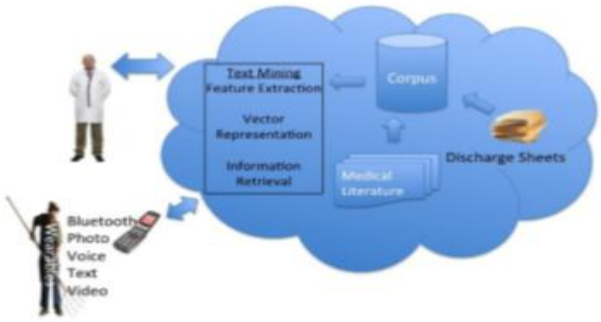

Figure 1: Automated Health Monitoring, Diagnosis and Regulation

IV. AUTOMATED DIAGNOSIS

A novice physician will generally depend on academic, book knowledge for diagnosing and further treatment while an experienced physician relies on his past experience and practical knowledge, more than book knowledge. Such knowledge or experience is gained by physicians when they deal with similar cases. All the details about the patients, their symptoms, diagnosis etc is stored on a discharge sheet in an accurate manner and such discharge sheets are maintained by the hospitals and clinics. Figure 2 from [7] shows a sample discharge sheet where all the personally identifiable information is removed. The idea is to use the expertise which is captured in the discharge sheet for diagnosis without any manual intervention. After all, a physician would use the same expertise anyway. The concept of text mining and information retrieval allows us to realize this idea.

Diagnosis: Allergic Bronchitis with Asthma
Case Summary: Patient 36 years male was admitted with complaints of breathlessness & cough for last 7 days. At the time of admission Pulse 126/min, BP 130/90 mmHg, RR 24/min, SpO2 94 with O2, Chest spasms wheezing+, ronchi++. Patient was investigated & treated conservatively with me /V antibiotics, me /V fluids, Nebulization & other supportive treatment. Now the patient is being discharged in satisfactory condition.
Treatment Advice:

- Tab. Augmentin 1gm 1tab. Twice daily
- Syp. Rapitus 2 TSF thrice daily
- Tab. Deriphyllin-R 150 mg 1 tab. Twice daily
- Forocort Rotacap 1 cap. Twice daily with Rotahaler

Figure 2: A Sample Discharge Sheet

The idea of text mining [10] is to identify a specific discharge sheet from a given set of discharge sheets. Text mining is used to match the symptoms of the patient to a discharge sheet with the exact symptoms or almost similar symptoms. Such a discharge sheet is used for diagnosis of that patient and further treatment. This diagnosis is returned back to the person from the computers in the cloud via the handheld device.

Text mining approach is more efficient than any other approach when it comes to automated medical diagnosis. To prove this hypothesis, text mining task can be performed on the currently available data set. The discharge sheets available are all in a single word document. Such document is first split into a corpus of documents where each document corresponds to one discharge sheet. As the corpus of discharge sheets available is quite less in size to perform text mining task, we can make the use of data sets available in public domain to check the feasibility of text mining approach. One such data set is the PMC full-text journal literature which has been made available as a part of the TREC CDS competition. We may chose the burns related literature from open access PMC and convert it into a corpus as well. Once the discharge sheets and the free access PMC has been converted into a corpus of documents we can perform text mining on this available data set efficiently. [7]

Text mining can be used to determine the relative frequencies of words in the corpus and generate a word cloud as shown in Figure 3. The bigger the size of the word, the more frequent it will be in the word cloud. Cluster analysis is another technique where relationships and similarities between various documents can be identified and depending upon such relationships, similar documents may be grouped together into one cluster. The next step is to apply text mining techniques for diagnosis problem. In the diagnosis problem, we have to find the discharge sheet which best matches the symptoms of the patient using k-nearest algorithm, once we have enough discharge sheets to do the processing.

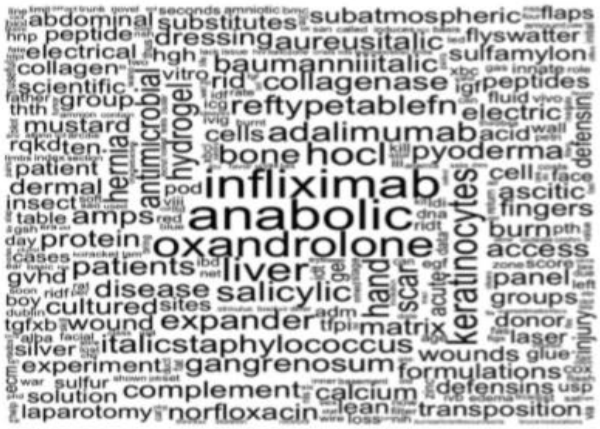

Figure 3: Word cloud from a PMC Data Set

The working of k-nearest algorithm consists of vectors which represent the documents. The vectors can be visualized as points in a multidimensional space called as the Vector Space Model. Each dimension is a word in the corpus. We can summarize this model by stating that each discharge sheet can be interpreted as a point in the vector space. We label each of these points with the diagnosis stated in the discharge sheet. The closer the points are to each other, more is the similarity between their symptoms. The symptoms of a patient can also be visualized as a document which can be plotted on the same vector space consisting of the discharge sheets as a point. Now, we just have to find the closest point to the symptoms document in the vector space. [11]

Visualization of points plotted in a multidimensional space is very difficult hence we use 'R' language functions to represent the documents in our corpus in two dimensions. This visualization helps in understanding how the points are grouped, their proximities to one another and other such properties. Text mining produces semantically correct results even though semantics wasn't used. [11]

The next step is the actual k-nearest algorithm where in the k nearest neighbors to the symptoms document are plotted. This is done by computing distance between symptoms document and the discharge sheet documents. Among all the neighbors the discharge sheet document with the minimum distance in the vector space is chosen. The diagnosis listed on this nearest neighbor to the symptoms document would be the diagnosis for given symptoms. The number of discharge sheets available is less for this approach hence it is currently a work in progress. [7]

V. CHALLENGES

The matter of medical diagnosis itself has been researched for more than four decades and finding a feasible solution is a formidable challenge. With addition of monitoring, it just multiplies the problem. The different types of challenges are given below.

A. Privacy and Security

An individual's medical data is very private and sensitive which cannot be compromised. It is necessary that the identity of the patients shouldn't be disclosed. For instance, if the system is hacked and a false diagnosis is sent to a person's handheld device, then the results could be fatal. [10]

B. Dataset

The collection of data is a major obstacle in the development of the system. Medical diagnosis is not possible without the massive data which may provide accurate result. Accumulation of big data may help in Machine Learning and Text Mining which is important for system development.
The integrity of data is one more issue that needs to be addressed. The doctors tend to be informal and incomplete while making reports. The Government can help by making it mandatory to fill the medical information in ways such that machine processing may become easier. [7]

C. Human Expertise

There will be need of human intervention when the system fails to give an accurate match. Manual intervention is also needed for routine checkup of the diagnosis detected by the system. Though some efforts mentioned in [5] give us some hope that manual intervention could be reduced.

D. Labeling

Each report should be labeled precisely with the diseases' name. Sometimes the doctors may use too many words or some other synonym for the equivalent term. Therefore there is need to map such diseases in a category of standardized term. Semantic Web techniques can be helpful in mapping the written diagnosis to standard term. [10]

E. Questionable quality of available reports

We have different techniques all over the world which are used for medical diagnosis of a particular disease. Depending upon the development of the country the diagnosis technology differs. Misdiagnosis and not capturing of symptoms is possible and relying on such report may create an ambiguity in the system.

F. Cost

The cost involved in realization of this technology is a major challenge. Wearables are still in the commercialization zone. Such cost of components which are to be used by the people for communication with the cloud machines may not make it easy for quick and rapid adoption of such ideas for underprivileged masses. [7]

VI. CONCLUSION AND FUTURE DIRECTIONS

We present our vision of how healthcare which is a vital function of life support can be made reliable with the use of machines. We have mainly focused our discussions on the automation of medical diagnosis, using techniques such as Text Mining and Information Retrieval. This will help us to move closer to the goal of providing ubiquitous medical diagnosis to the masses.

We plan to get more real world data to get a better feel of the problem. This in a way will change the approach of tackling the problem. We intend to use different datasets and classification techniques. We are also keen to try the n-gram model and use K-NN algorithm to examine the feasibility of automated model. [11]

Hadoop has been the foundation of big data storage and processing. It has been evolving at a rapid space which may be used to handle big data and eventually improve the results of the system. We may use semantic web kind of reasoning which might influence the existing approaches. A few more steps in our vision include the QA interface, speech recognition, reasoning and explanation. The cognitive computing programming paradigm could also be a useful angle of exploration. [12]

REFERENCES

[1] Scheuermann, R. H., Ceusters, W., & Smith, B. (2009). Toward an onto- logical treatment of disease and diagnosis. Summit on translational bioin- formatics, 2009, 116. Shadbolt, N., Hall, W., & Berners-Lee, T. (2006). The semantic web revis- ited. Intelligent Systems, IEEE, 21(3), 96-101.

[2] Sarwar, A., & Sharma, V. (2014). Comparative analysis of machine learning techniques in prognosis of type II diabetes. AI & society, 29(1), 123-129.

[3] Pariyadath, V., Stein, E. A., & Ross, T. J. (2014). Machine learning classification of resting state functional connectivity predicts smoking status. Name: Frontiers in Human Neuroscience, 8, 425.

[4] Ngufor, C., Wojtusiak, J., Hooker, A., Oz, T., & Hadley, J. (2014, March). Extreme Logistic Regression: A Large Scale Learning Algorithm with Application to Prostate Cancer Mortality Prediction. In The Twenty-Seventh International Flairs Conference.

[5] Sapounas, D., Jackson, K., & Ervin, D. (2011, October). International Consultants in Medicine: A Framework for Medical Expertise and Social Telemedicine Addressing Medical Disparities. In Global Humanitarian Technology Conference (GHTC), 2011 IEEE (pp. 208-211). IEEE.

[6] Gobbel, G. T., Reeves, R., Jayaramaraja, S., Giuse, D., Speroff, T., Brown, S. H., ... & Matheny, M. E. (2014). Development and evaluation of RapTAT: a machine learning system for concept mapping of phrases from medical narratives. Journal of biomedical informatics, 48, 54-65.

[7] Pendyala,V.S.; Yi Fang; Holiday, J. ; Zalzala, A. , A Text Mining Approach to Automated Healthcare for the Masses, IEEE Global Humanitarian Technology Conference (GHTC), Oct. 2014, pp. 28-35.

[8] Park, S., & Jayaram, S. (2003). Enhancing the quality of life through wearable technology. Engineering in Medicine and Biology Magazine, IEEE, 22(3), 41-48.

[9] Logothetis, D., Olston, C., Reed, B., Webb, K. C., & Yocum, K. (2010, June). Stateful bulk processing for incremental analytics. In Proceedings of the 1st ACM symposium on Cloud computing (pp. 51-62). ACM.

[10] Akilan, A. Text Mining: Challenges and Future directions. (2015) Electronics and Communication Systems (ICECS) 2nd International Conference.

[11] Fang Lu; Qingyuan Bai. Intelligent Systems and Knowledge Engineering (ISKE), (2010). International Conference.

[12] Amir, A., Datta, P., Risk, W. P., Cassidy, A. S., Kusnitz, J. A., Esser, S. K., ... & Modha, D. D. (2013, August). Cognitive computing programming paradigm: a coherent language for composing networks of neurosynaptic cores. In Neural Networks (IJCNN), The 2013 International Joint Conference on (pp. 1-10). IEEE.