ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 5, Issue 5, May 2016

A Tag Mining Framework for Health Seekers using Deep Learning

Miss. Kalyanee J. Harne, Prof. Krutika K. Chhajed

***Abstract*— Health is one of the increasing subjects used for assessing health condition among patients who suffer from specific ailment or diseases .An important problem of current Web search is that search queries are usually short and not enough for knowledge inferring, and thus are not good enough for specifying the precise user need. There are many online and offline methods to get the information requested by the health seeker. Here the sparse deep learning algorithm is used as the data mining technique. Deep learning is a branch of machine learning based on a set of algorithms that attempt to model high-level abstractions in data by using model architectures, with complex structures. The proposed scheme uses question-answering, deep learning as inferring methods. Some attributes used are raw features, medical attributes etc. Deep learning refers to the depth wise analysis of the raw features as input nodes in one layer and hidden nodes in the next layers of the learning architecture. That is it learns the internal relations between the data collected and several layers. The input will be the raw data, the output will be predicted disease and the intermediate layers will be hidden from the data. An extensive learning structure is used to infer the diseases given by the queries of health seekers.**

***Index Terms*— Data Mining, Question Answering, Deep Learning**

I. INTRODUCTION

Recent studies have revealed that patients prefer online advice rather than embracing doctor's advice passively. This has been confirmed by a National Survey conducted by the Pew Research Center in Jan 2013.The survey results reported that one in three American adults have gone online to figure out their medical conditions in the past 12 months from the time of report. A number of community-based healthcare services have turned up, including Health Tap, HaoDF and WebMD. They are disseminating personalized health knowledge and connecting patients with doctors worldwide via question answering [2], [3]. These forums are very attractive to both professionals and health seekers. Data mining brings a set of tools and techniques that can be applied to the processed data to discover hidden patterns that provide healthcare professionals an additional source of knowledge for making decisions. The decision rests with health care professionals.

Community Question answering (cQA) services have gained a lot of popularity over the last few years. They allow people with diverse backgrounds to share their knowledge and experiences. Community based Question Answering (CQA) services are defined as dedicated platforms for users to respond to other users questions, resulting in the building of a community where users share and interactively give answers to questions (Liu et al., 2008).Users with diverse backgrounds do not necessarily share the same vocabulary. Take Health-Tap as an example, which is a question answering site for participants to ask and answer health-related questions. The questions are written by patients in narrative language. The same question may be described in substantially different ways by two individual health seekers. Recently, some sites have encouraged experts to annotate the medical records with medical concepts. However, the tags used often vary wildly and medical concepts may not be medical terminologies. For example, "heart attack" and "myocardial disorder" are employed by different experts to refer to the same medical diagnosis [7].The work aims deep learning of the possible disease given by the questions of health seekers [4]. Deep Learning is a relatively recently developed set of generative machine learning techniques that autonomously generate high-level representations from raw data sources, and using these representations can perform typical machine learning tasks such as classification, regression and clustering. The prime intention of deep learning comprises of two key components. The first globally mines the discriminant medical attributes from raw features. The raw features serve as input nodes in one layer and hidden nodes in the subsequent layer, respectively. The second learns the inter-relations between these two layers via pre-training. With incremental and alternative repeating of these two components, the scheme builds a sparsely connected deep learning architecture with three hidden layers. Deep learning scheme is applied to infer the possible diseases using Question and Answer data values.

II. LITERATURE SURVEY

Many people spend longer time online to explore health information. One survey in [5] shows that 59% of U.S. adults have explored the internet as a diagnostic tool in 2012. Another survey in [6] reports that the average consumer spends close to 52 hours annually online to find wellness knowledge, while only visits the doctors three times per year in 2013.

In 2009, David Barbella [8] proposed a system where the Support vector machines are a valuable and useful tool for making classifications. But their black-box nature means that they lack the natural explanatory value that many other classifiers possess. In the first, we report the support vectors most touching in the final classification for a particular test location. Next we determine which features of that test location would need to be changed in order to be placed on the separating surface between the two classifications.

In 2011, M. Shouman [9] established the availability of large amounts of medical data that leads to the need for powerful data analysis tools to extract useful knowledge for finding a particular need towards health. Here as a result, researchers have been investigating the result of hybridizing more than one technique showing enhanced results in the diagnosis of a disease (heart disease).It is done by the motivation by the world-wide increasing mortality of heart disease patients each year and the availability of huge amounts of data, researchers are using data mining techniques in the diagnosis of disease especially heart disease.

In 2011, Liqiang Nie, Meng Wang, Zheng-Jun Zha, Guangda Li and Tat-Seng [10] proposed a method to generate queries from QA pairs for multimedia search and perform query-dependent re-ranking for image and video data obtained from search engines by analyzing visual features. In this paper, author presents a plan that can advance content answers with picture and video data. Given an inquiry and the group contributed answer, author's methodology can figure out which sort of media data should be included, and afterward naturally gathers information from Internet to improve the printed answer.

In 2012, Tom Chao Zhou, Michael R. Lyu, Irwin King[11] introduced the Community based Question and Answering (CQA) administrations have conveyed clients to another time of information spread by permitting clients to make inquiries and to answer other clients questions. Author likewise perform a systematical correlation on how distinctive sorts of components add to the last results what's more; demonstrate that question-client relationship components play a key part in enhancing the general execution. There is a serious gap between the existing open questions and potential answerers. To bridge the gap, they present a new approach to Question Routing, which aims at routing open questions to suitable CQA users who may answer these questions.

In 2013, F. Wang [12] proposed a temporal knowledge representation and learning framework to perform large scale temporal signature mining of longitudinal heterogeneous event data occurrences. Novel stochastic optimization architecture performs large-scale incremental learning of group-specific temporal event signatures. It evaluates the framework on synthetic data and on an electronic health record dataset and its manipulation. This architecture enables the representation, extraction, and mining of high- order latent event structure and relationships within single and multiple event sequences.

In 2013, L. Nie, M. Wang, Y. Gao, Z.-J. Zha, and T.-S. Chua [13] introduced three practices. For a question, retrieve question answer pair from the available question answering sites database dynamically and select an answer medium to enrich the textual answer. By processing a large set of QA pairs and adding them to a pool, this approach can enable a novel multimedia question answering (MMQA) approach as users can find multimedia answers by matching their questions with those in the pool. In this paper, author proposes a plan that can enhance printed answers in cQA with fitting media information. This methodology naturally decides which kind of media data ought to be included for a printed answer.

In 2013, Matthias Galle [14] proposed that the n-gram representations of reports may enhance over a basic bag-of-word representation by unwinding the independence suspicion of word and presenting setting. Author showed new representations that maintain a strategic distance from pitfalls.

In 2014, L. Nie, Y.-L. Zhao, X.Wang, J. Shen, and T.-S. Chua [15] presents a novel plan that can exhaustively learn unmistakable labels for every question. Around 40% of the questions in the developing social-oriented question answering discussions have at generally one physically marked label, which is brought on by incomprehensive question understanding or casual labeling practices.

In 2014, L. Nie et al., [16] have presented their effort to inhibit the terminology gap among health inquirers and providers which deferred the cross system inter-operability. Local mining and global learning methods are conjointly exploited. Local mining targets to locally code the health registers by mining the medical concepts from discrete record and then mapping them to lexicons based on the exterior legitimate terminologies. Health provider released sources by utilizing either isolated or loosely connects rule-based and machine learning approaches.

Liqiang Nie [1] established a system in which disease is inferred from Health-Related Questions via Sparse Deep Learning technique is an efficient technique to identify diseases, or to monitor the health status. As mentioned earlier, vocabulary gap, incomplete information, inter-dependent medical attributes and limited ground truth have greatly hindered the performance of classic shallow machine learning approaches. Most of them can benefit from labeled data; unlabeled data supervised and unsupervised data, which ensures fair comparison. This technique mainly focused on sparse deep learning technique where each layer is incrementally added based on the users need. SVM is implemented here as a classifying tool.

III. SYSTEM ARCHITECTURE

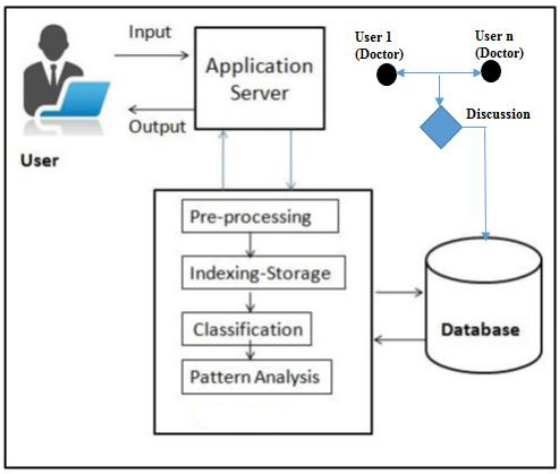

Fig. (a) System Architecture

1) In this architecture the four main components are the Health Seeker, Application Server, Doctors in community and the Dataset.
2) Health Seeker is the person who gives input in the form of Query to the Application server, Application Server is responsible for taking input from user and passing it to further block for its processing and in return it also gives back result to the health seeker.
3) Query's passed is processed first and different pre-processing operations are applied on the given query and finally divided in a form of Tokens using String-Tokenizer.
4) This Tokens are the passed in further block for Indexing Storage. Where we mine each and every data related to the tokens we got from our Data Set. Each Data is assigned with an Index value. All the Relevant Data we got is then passed further for its Classification.
5) In Classification Block the data are divided into different classes. First step in text classification is transforming text which is in string format into format suitable for learning algorithm.
6) In Pattern analysis Boyer-Moore-Horspool algorithm is used. Layer by layer elimination of entries is done by using Deep learning algorithm. Recursive Process is done here in classification and pattern analysis block till the result is not obtained by Eliminating the Entries from Generalized data to Specified data i.e., from huge data to a smaller data and hence the result is obtained here.
6) Finally, the Result obtained through Classification and Pattern analysis is then passed to user.

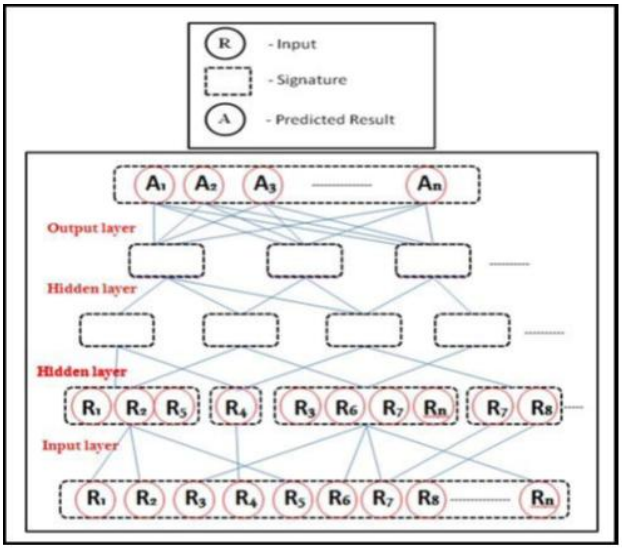

Fig (b) Hidden Layer Architecture

*a) Input Layer***:** Input Layer is used to make classifiers more efficient by reducing the amount of data to be analyzed as well as identifying relevant features to be considered in classification process. Ideally, pre-processing stage will refine features, which are input into a classification / learning process. The input in original form is not suitable for learning. They are transformed into format which matches into input of deep learning algorithm input. For this pre-processing on input text is carried out. Each word will correspond to one dimension and identical words to same dimension.

b) Hidden Layer: The hidden layer is where the system stores its internal abstract representation of the processed data. Formally, a one-hidden-layer is a function $f : R^D \rightarrow R^L$, where D is the size of input vector x and L is the size of the output vector $f(x)$, such that, in matrix notation:

$$f(x) = G(b^{(2)} + W^{(2)}(s(b^{(1)} + W^{(1)}x))),$$

With bias vectors $b^{(1)}, b^{(2)}$; weight matrices $W^{(1)}, W^{(2)}$ and activation functions G and s.

$W^{(1)} \in R^{D \times D_h}$ is the weight matrix connecting the input vector to the hidden layer. So each hidden layer going further into the network is a nonlinear combination of the layers below it, because of all the combining and recombining of the outputs from all the previous units in combination with their activation functions.

c) Extraction from Datasets: A collection of related sets of information that is composed of separate elements and can be manipulated as a unit by computer. In dataset, we use hyper plane. A hyper plane is a function like the equation for separation of data,

$y = mx + b$.

In fact, it uses to separates your data into the different classes.

When user fires any query then the system accept that request and compare with the collected dataset. The dataset are nothing but the raw data.

VI. PROPOSED METHODOLOGY

- *Boyer-Moore-Horspool algorithm is used for string pattern matching.*

Steps:
1. Take a user input.
2. Convert or place it in array i.e. Q1.
3. Fetch the data from database
4. Place that in another array i.e. A1.
5. Start matching from the beginning of the substring
6. If find a mismatch Next part will take.

```
Start at beginning of string
Start at beginning of match
While not at the end of the string:
   if match_position is 0:
      jump ahead m characters
      Look at character, jump back based on table 1
      if match the first character:
         advance match position
      advance string position
   else if I match:
      if I reached the end of the match:
        FOUND MATCH - return
      else:
        advance string position and match position.
   else:
      pos1 = table1 [character I failed to match]
      pos2 = table2 [how far into the match I am]
```

if pos1 < pos2:
jump back pos1 in string
set match position at beginning
else
set match position to pos2
FAILED TO MATCH

- *Algorithm Proposed Approach i.e. Deep Learning.*

1. Start
2. Take query input from the user
3. Divide the data in different tokens.
   Input= {i1, i2, i3…….in}
   Tokens= {i1}, {i2}…… {in}
   Dataset= {d1……dn}
4. Layer specification
   While (Layer)
   {

   Result R= {in} U {d1……..dn}

   }
5. Formally, a one-hidden-layer is a function $f : R^D \rightarrow R^L$, where D is the size of input vector x and L is the size of the output vector $f(x)$, such that, in matrix notation:

$$f(x) = G(b^{(2)} + W^{(2)}(s(b^{(1)} + W^{(1)}x))),$$

6. With bias vectors $b^{(1)}, b^{(2)}$; weight matrices $W^{(1)}, W^{(2)}$ and activation functions G and s.

   $W^{(1)} \in R^{D \times D_h}$ is the weight matrix connecting the input vector to the hidden layer

7. Finally Different layer processing the result R given to user.

V. RESULT ANALYSIS

a) Analysis with respect to various numbers of hidden layers:

Number of Layers	Performance on Dataset of Base Paper	Performance on Dataset of Proposed System
Structure with one hidden layer	89.00	89.10
Structure with two hidden layers	93.13	93.15
Structure with three hidden layers	98.21	98.43

Table 1: Comparison between the various numbers of Hidden Layer

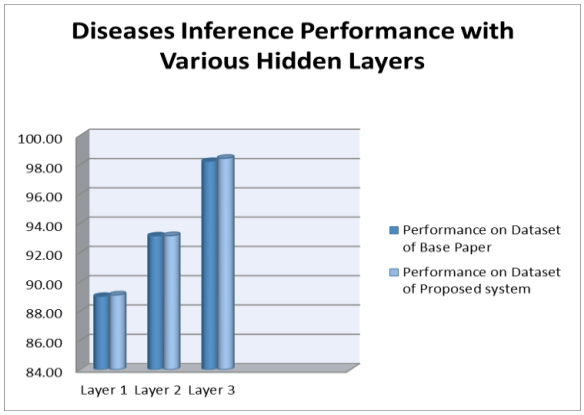

Figure 1: Graph showing performance with various numbers of Hidden Layers.

b) Analysis with respect to number of symptoms:

Number Of Symptoms	Project Result Accuracy
Symptom 1	30.00%
Symptoms 2	37.00%
Symptoms 3	70.05%
Symptoms 4	82.38%
Symptoms 5	88.39%

Table 2:Result analysis according to Number of symptoms

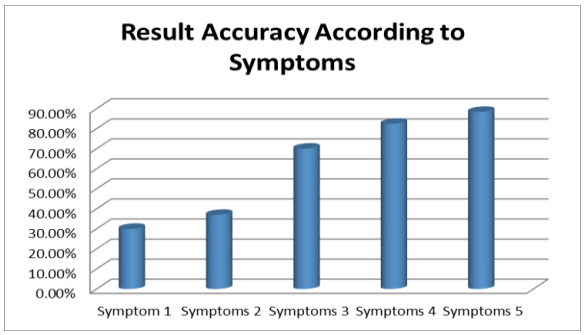

Figure 2: Graph showing Result accuracy according to symptoms

c) Analysis of time required for query processing and execution:

Number of Symptoms	Project Result Time(ms)
Symptom 1	2.40
Symptoms 2	3.00
Symptoms 3	3.10
Symptoms 4	3.80
Symptoms 5	4.40

Table 3: Time (ms) required for Query Processing and execution

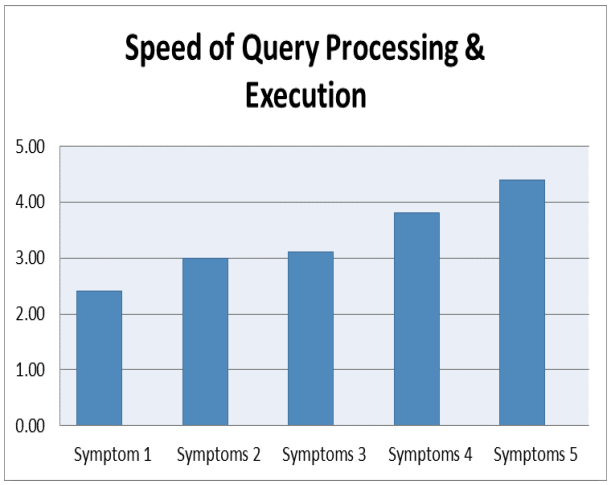

Figure 3: Graph showing Time (ms) required for Query Processing and execution

VI. CONCLUSION

Deep learning methods have recently made notable advances in the tasks of classification and representation learning. In this work we demonstrate our results (and feasible parameter ranges) in application of deep learning methods to structural and functional information. Thus the proposed system involves an efficient machine learning approach for mining health related data. The hidden layers between the input and output layers are incrementally increased based on the accuracy. Present study reveals the community-based health services. This proposed system presents a sparsely connected deep learning scheme that is able to infer the possible diseases given the questions of health seekers. In future it can be used in clinic for diseases differencing according to users symptoms.

REFERENCES

[1] Liqiang Nie, Meng Wang, Luming Zhang, Shuicheng Yan,Bo Zhang and Tat-Seng Chua,“ Disease Inference from Health-Related Questions via Sparse Deep Learning, ”IEEE Transactions on Knowledge and Data Engineering May 2014.

[2] L. Nye, M. Akbar, T. Li, and T.-S. Chua, “A joint local-global approach for medical terminology assignment,” in Proc. Int. ACM SIGIR Conf., 2014.

[3] L. Nye, T. Li, M. Akbar, and T.-S. Chua, “Wincher: Comprehensive vertical search for healthcare domain,” in Proc. Int. ACM SIGIR Conf., 2014, pp. 1245–1246.

[4] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013.

[5] S. Fox and M. Duggan, “Health online 2013,” Pew Research Center, Survey, 2013.

[6] “Online health research eclipsing patient-doctor conversations,” Makovsky Health and Kelton, Survey, 2013.

[7] Leroy, Gondy, and Hsinchun Chen. "Meeting medical terminology needs-the ontology-enhanced Medical Concept Mapper." Information Technology in Biomedicine, IEEE Transactions on 5.4 (2001): 261-270.

[8] D Barbella, S Benzaid, JM Christensen, B Jackson, XV Qin, DR Musicant DMIN, “ Understanding Support Vector Machine Classifications via a Recommender System Like Approach.” .pp. 305-311,2009

[9] M.Shouman, T. Turner, and R. Stocker, “Using decision tree for diagnosing heart disease patients,” in Proceedings of the Australasian Data Mining Conference, 2011

[10] L. Nie, M. Wang, Z. Zha, G. Li, and T.-S. Chua, “Multi-media answering: Enriching text qa with media informa-tion,” in Proceedings of the International ACM SIGIR Conference, 2011.

[11] T. C. Zhou, M. R. Lyu, and I. King, “A classification-based approach to question routing in community question answering,” in The International World Wide Web Conference, 2012

[12] F. Wang, N. Lee, J. Hu, J. Sun, S. Ebadollah , and A. Laine, “A framework for mining signatures from event sequences and its applications in healthcare data,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013.

[13] L. Nie, M. Wang, Y. Gao, Z.-J. Zha, and T.-S. Chua, “Beyond text qa: Multimedia answer generation by harvesting web information,” Multimedia, IEEE Transac-tions on, 2013

[14] M. Gall´e, “The bag-of-repeats representation of docu-ments,” in Proceedings of the International ACM SIGIR Conference, 2013.

[15] L. Nie, Y.-L. Zhao, X.Wang, J. Shen, and T.-S. Chua, “Learning to recommend descriptive tags for questions in social forums”, Acm Transactions on Information System, 2014.

[16] L. Nie, Y.-L. Zhao, M. Akbari, J. Shen, and T.-S. Chua, “Bridging the vocabulary gap between health seekers and healthcare knowledge,” IEEE Trans. Knowl.Data Eng., vol. 27, no. 2, pp. 396–409, Jun. 2014

Kalyanee J. Harne received the B.E. degrees in Computer Science & Engineering from Sipna College of Engineering & Technology in 2014. Now she is pursuing ME (CSE) from P.R.Pote (Patil) College of Engineering &Management, Amravati.

Krutika K. Chhajed received the B.E. degrees in Information Technology from Prof. Ram Meghe Institute of Technology & Research in 2005 and completed ME (CSE) from Prof. Ram Meghe Institute of Technology & Research, Amravati**.**