Automated Disease Inference For Community Health Care Using Graph Induced Neural Network

M.Dhanamalar[1], Dr. K. Kavitha[2]

[1]Research Scholar, Mother Teresa Women's University, Kodaikanal
[2]Assistant Professor, Mother Teresa Women's University, Kodaikanal
[1]dhanamca03@gmail.com, [2]kavitha.urc@gmail.com

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 28 April 2021

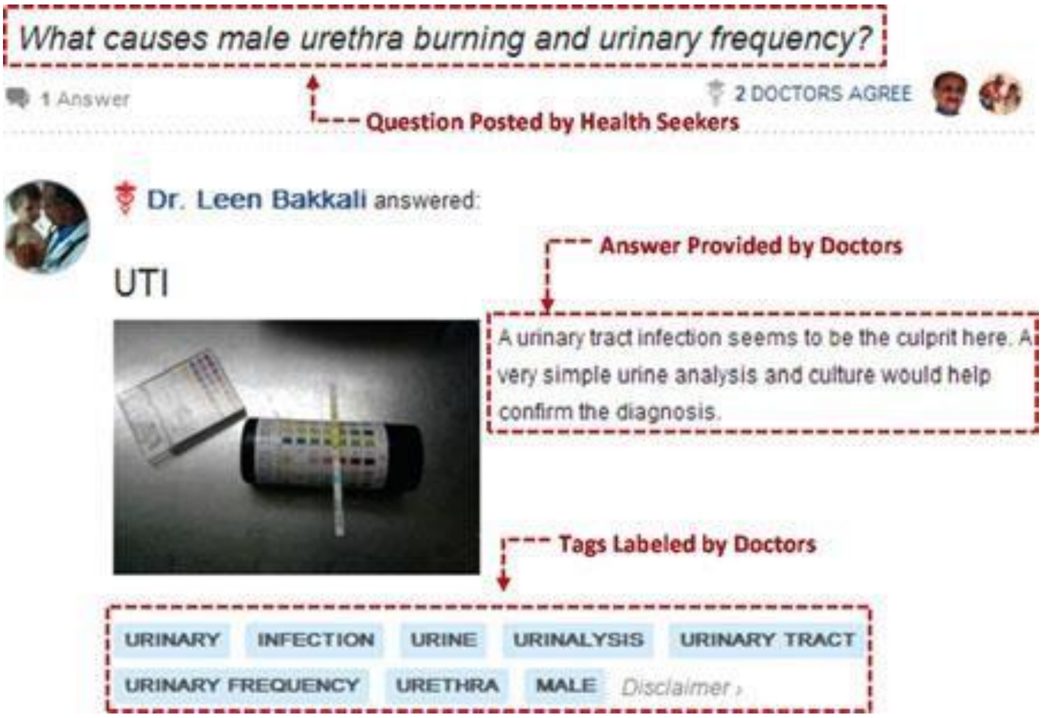
Abstract:

One of the main problems in the area of community health care is to identify the disease based on the symptoms and lack of medical vocabulary among the users. There are limited number of doctors and medical practitioners available in these forums. The lack of medical professionals requires an automated solution, but this solution must be capable of bridging the vo- cabulary gap between the users and the medical terms. Therefore a simple medical information retrieval and file sequencing might not be the solution, we need a system capable of processing natural language and convert the necessary terms to a suitable medical term thus leading to accurate disease inference. There are many excellent text mining algorithms that are available such as GRU and LSTM , these algorithms work excellently on health records and documents with accurate description of diseases. The community data is very vast and more relatable to the data which will actually be used in day to day usage of the algorithm. Graph based algorithms have been widely used fr natural language processing, with neural network's capability or retention of good information and elimination of noise can actually be of a great use.

1. Introduction

The age of internet has revolutionized the ways in which many tasks are done, tasks which earlier seemed to have a need of high level of human intellect is now available for next to no price on internet. One such task s medical ad vices long gone are the days when a patient with very little symptoms of the disease go to doctor for consultation, not only this changes the amount spent by people on average on medical consultation, doctors can also be relived of repetitive and busy schedules. But how reliable are these websites or bots on internet when it comes to inferring the health issue given medical condition in not so correct terminologies. There are many health care forums which serve as a repository for health issues, users often post their conditions with best possible words as they can, which is often not enough to infer the condition with out a visual inspection and many other information that a medical practitioner would need to provide the correct inference of disease. Most of the bots use a sequential data storage for memory and expect a near close description of disease, otherwise they fail to give accurate predictions. Even if we train a bot with LSTM and medical terms or health records the relationship exhibited in the model will be of no use when we try to apply the questions asked in community based health forums or bots. Text mining techniques used for health records may not necessarily provide similar results when applied over a Health care forum. The issue with the health forum data is, the questions asked are very vague in some cases, although using certain background information and further analysis a medical practitioner might have answered the question, we cannot guarantee the same when we think of a model trained over traditional methods to have that level of context and repeat the task that a doctor may apply. Our research is dedicated to bridge this gap in vocabulary between the user and the medical terms.

1.1 Motivation



What causes male urethra burning and urinary frequency?

Dr. Leen Bakkali answered:

UTI

A urinary tract infection seems to be the culprit here. A very simple urine analysis and culture would help confirm the diagnosis.

URINARY | INFECTION | URINE | URINALYSIS | URINARY TRACT | URINARY FREQUENCY | URETHRA | MALE

Figure 1.1: Example of a QA pair from community-based health services[1]

The biggest stumbling block of automatic health system is disease inference. According to our user study on 5;000 questions that health seekers frequently ask for:

1. supplemental cues of their diag-nosed diseases. 2.preventive information of their concerned diseases. 3.possible diseases of their manifested signals.

The first two problems requires the exact disease names and side effects which can happen for a given treatment of medication. This problem of identify- ing the exact disease can be solved by finding the match of the question from the archived data of the community websites for health care, this be- comes a specific type of data science problem of information organization and retrieval. For the third type of problem we need more details such as the physical and mental conditions of the users during the period and demo- graphic information, with these information only the medical experts would able to provide the right information, these data can also be extracted from the websites and thus, if disease inference is made possible with these data then the third type of problem can also be converted to the first two types thus providing a common solution to all the three problems in an automated fashion. Health seekers usually explain their issues with small questions which is very less for term detection over any algorithm without a strong context backing techniques or feature elimination. This is what was aimed by our research to make the solution more durable. Another important reason for emphasis on the context is the lack of medical vocabulary which is very troublesome when the terms from which the inference has to happen is also very less, as mentioned earlier the question posted by the user is usually very small and insufficient with unusual vocabulary can just make the things worse for a text mining algorithm like RNNs where context mining is just backed by eliminations of least important feature and context builds up wit such elimination can not be completely relied upon.

2. Literature Survey

Bridging the gap between what online health seekers with unusual signs req- uisite and what busy human doctors with biased proficiency can offer is of greater importance. Online health care forums have been assisting well being condition monitoring, illness modelling and validation of medical treatment by medical text mining. Precisely and competently concluding the diseases is non-trivial, especially for community-based health services due to the lexis gap, inadequate information, interrelated medical concepts, and incomplete high quality training samples.Large collections of electronic clinical records today provide us with a vast source of information on medical practice. How- ever, the utilization of those data for exploratory analysis to support clinical decisions is still limited. Extracting useful patterns from such data is partic- ularly challenging because it is longitudinal, sparse and heterogeneous.

S.Doan[2010] et al. Proposed that Several machine learning based systems have been developed and showed good performance in the challenge. Those sys- tems often involve two steps:

1. recognition of medication related entities; and
2. determination of the relation

between a medication name and its modifiers (e.g., dosage). A few machine learning algorithms including Conditional Random Field (CRF) and Maxi- mum Entropy have been applied to the Named Entity Recognition (NER) task at the first step. In this study, we developed a Support Vector Machine (SVM) based method to recognize medication related entities. In addition, we systematically investigated various types of features for NER in clinical text. Evaluation on 268 manually annotated discharge summaries from i2b2 challenge showed that the SVM-based NER system achieved the best F-score of 90.05 % (93.20 % Precision, 87.12 % Recall), when semantic features gen- erated from a rule-based system were included.

T. D. Wang et al.[2008] proposed a methodology to provide an interface to search health records and databases . This method of search mainly looks for absolute similarity in the patterns in cause and effect for the diseases. The interface created by the researcher had the capability to filter , align records based on the query over the database , say we want to fetch all the records which has a particular term or all the records belonging to particular category or sub category can be queried and presented on the tool. When we display the history of patients with a particular terms we see that some terms are co-occurring and there are some after effects of the treatment which remain the same for the disease over a group of demographics. The analysis with this tool presented much interesting results about health care records which would rather be difficult to analyse over a hard paper records where querying and fetching results with huge vlmes of the files is nearly impossible.

A. Khosla et al.[2010] proposed a method in feature selection with support vec- tor machines for identifying the diseases. The feature selection becomes very important when we are working on medical data , we must not ignore any attribute just because the attribute is inconsistent in the health records at the same time we cant kee a feature which adds no value but is common in a particular disease inference, to solve this problem the authors of this work proposed feature selection method called margin based censored regression, combined this method with SVM they produced a great disease classifica- tion solution which had substantially beaten the then best Cox model and produced results way above in modern AUC curves and concordance index.

N. Limsopatham et al.[2013] in his research stated that the complex medical ter-minologies are very challenging to be identified in the medical records, such as the words such as tumour,cancr, and neoplasma which are all terms which may be similar and attribute to a particular type of diseases , traditional methods may find it very difficult to identify the right terms and group them in these cases. The works prior to this research involved in using information retrieval techniques such as bag of words which often would expect the exact words to be considered in the part of the information of the knowledge base created.

S. Ghumbre et al.[2011] created an expert system using SVM and radial ba-sis function network to perform diagnosis of heart attacks. The system is trained with medical data mostly of the symptoms to determine the type of heart disease the person may have based on symptoms and data from the patient’s health record. The results of this system is the studied against the radial basis function with orthogonal least square to same data sets . This study provide clear view on how machine learning models can be used on medical symptoms data .This study was the first of its kind to provide a ML based solution for automated disease inference.

2.1 Outcome of Literature Review

The literature review gives us a clear picture that the researches conducted in the field of health care text mining is mainly related to the electronic health records and other offline resources which are well structured but when it comes to the online resources these methodologies which are mentioned in the literature review often fail mainly due to following reasons. From the perspective o Many researches were conducted to get data from the user using the queries and questionnaire that were answered by the users , although these researches were confined to find the apt retrieval means for the medical data. There are existing work in the field on online answering for the health related queries except for the artificial intelligence apps with information retrieval and standard answering from stored solutions.

2.2 Problem Statement

To propose a system that overcomes the failure of online health QA based websites due to user's lack of vocabulary by automating the answers to the user's queries with natural language processing.

2.3 Research Objectives

To automate the answers that are provided by authorised doctors on the websites for health related problems using sparsed deep learning neural net- works on the signatures that are mined from the question and answer data set collected from the healthtap api and other health blogging websites using beautifulsoup api

The objectives for automated disease inference are the following:

- The model must be able to mine the signature words from the questions or queries that are provided by he users on th websites.
- The proposed system is expected to overcome the lack of proper vo- cabulary in the user's query that fails to provide the doctors accurate medical term for the illness.
- to develop a sparsed neural network with the first layer taking the raw input features from the QA datasets and three hidden layer to get the signatures using the graphs that are built over the terms used in the raw input features.
- The above structure of one hidden layer of signature is considered as input features for the other three hidden layer which uses the sigmoid function for pretraining and for further optimization we have also pro- posed to autotune the layers.

3. Methodology

3.1 Local Mining

The questions posted by a user group depends mainly on his level of educa- tion, kind of medical knowledge he is having. Sothe same question can be asked in different ways, and the kind of medical concept extraction highly depends on this. So, in order to meet this we will be normalizing the medical concepts obtained by the questions of user group. The first step is to extract the noun phrases out of the questions. The regular expression used in this process can be seen in 3.1. The next step is to extract the noun phrases which are relevant to medical corpus. The next step is to pass these medical noun phrases to SNOMED Internation Browser which will return us the set of medical concept indexes. The medical concepts can be obtained out of these indexes with the help of vetmed tool where one is passing these index numbers obtained from SNOMED and getting the medical concepts associ- ated with it. The one pair which maximizes the google search count will be considered as a medical concept associated with a medical noun phrase. The entire process along with a expression is depicted in Figure 3.1.

$$(Adjective|Noun)\ *(NounPreposition)?(Adjective|Noun)\ *Noun \tag{3.1}$$

3.2 Signature Mining

The data sets that we are trying to process contains question and answers . these questions are asked by the users and are having terms which are 8

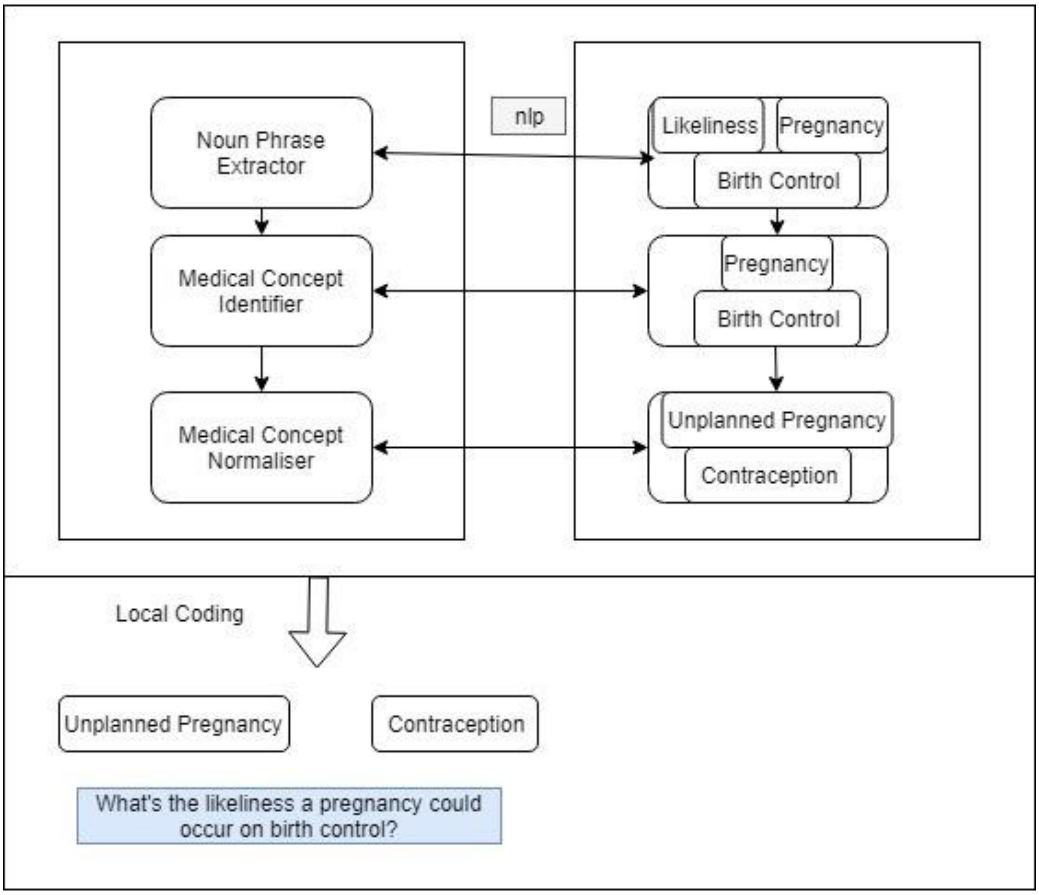


Figure 3.1: Local Mining

redundant in many disease definitions thus makes it difficult for a fuzzy system to make a decision on what disease type does the question belongs. One such example could be headache and nausea, i this case with just these terms we cannot infer if it is migraine or brain tumour. The first step in signature mining is to create a graph with medical terms extracted from the question and answer pairs as vertices and edge between them if they co-occur on global dictionary. The edge is determined by the terms, say headache and fatigue are two words if they co-occur more then a fixed number of times then we form the edge between them. The co-occurrence of the terms will fix the edge weight for the vertices forming edges. While implementing in the code side we made sure the extracted terms are matched in regular English dictionary and snomed medical dictonary and we validated whether the term is medical or non-medical terms. The second step of signature mining is to find the densest sub-graphs using k-clique algorithm, these sub-graphs will be the medical signatures.

3.2.1 K-Clique

The co-occurrence is graph is first constructed with the steps provided in signature mining. For the given co-occurrence graph we check for each vertex one by one and identify the edge if exist. If an edge is found after the traversal, we define a k value, which means the edge must be between k vertices then only we consider the sub=graph created by the edge. The identified sub-graph must also meet another condition that the graph must be complete graph. As discussed in signature mining technique we eliminate the edges which doesn't have the weight value with threshold, in the resultant graphs from k-cliques the words of these vertices co-occur in the root on both questions and answers thus converting inaccurate vocabulary of the user to a meaningful medical signature. The process of signature mining doesn't end here, we then use the mined signatures to normalize the signatures with accurate medical term, does eliminating the inaccuracy in training for the disease classification in total. The results of k-cliques are stored in separate files which will constitute the neurons of hidden layer. The input neurons will be mapped to these grouped signatures and anyways the signatures are stored in files with the filenames as the disease names which are the classes for the predictions.

3.3 Sparse Neural Network

Sparse neural network is a regular neural network but the hidden layers are not completely connected to the input layer here instead the results from the sub-graphs are mapped to the input layer and constitute the hidden layer. The input layer takes the normalized medical terms in each neuron the corresponding word vectors of the normalized medical terms are generated by binary vectors in which the bit in the word vector is turned 1 if the word is present in the bag and if not it has 0.The hidden layer is built in similar way but instead of random numbers we use the files which contains the signatures of the medical concepts these word vectors in the hidden layer neuron has the same text file name as that of the input layer files thus it becomes easy to map the layers and are multiplied by synapse values which are trained over and over to get a accurate model. The finally the hidden layer is strongly connected to the out layer which means each neuron the hidden layer is connected to the neuron in the output layer. We have used a sigmoid activation function when we train we have the actual class which is again a vector where one bit is turned 1 to represent the class able when we apply the sigmoid function over the hidden layer and the synapse product value with the learning rate we get a vector if this vector doesn't match the out neuron which is connected to hidden layer then we consider that as error and over motive in training is to reduce this error thereby increasing our model's accuracy. The architecture is depicted in Figure 3.2.

3.4 Long Short Term Memory Networks

We created a simple LSTM architecture using keras library which has one input, lstm and output layer, similar to the previous approach we pass the in- put matrix with the neurons having the input vectors but here we randomize the hidden layer instead of using the signatures that we mined. LSTM model is known for saving the context and training the model with stored context in the further iterations t helps the model to not just train the values over the synapse in one direction but all the possible combinations of the input. In the figure we can see the four time stamps of the LSTM model where at each of the time stamp the previous input value or vector in our case is added with the hidden layer vector to give us one possible input and rains the network similarly at each time stamp the input layer is kept on changing unless the model learns all the permutations and combinations of the input here instead of using a sigmod function the LSTM network uses coupled tanh function as the activation functions. We shall see how this model preforms against the Sparsely connected neural networks model. The architecture is depicted in Figure 3.3.

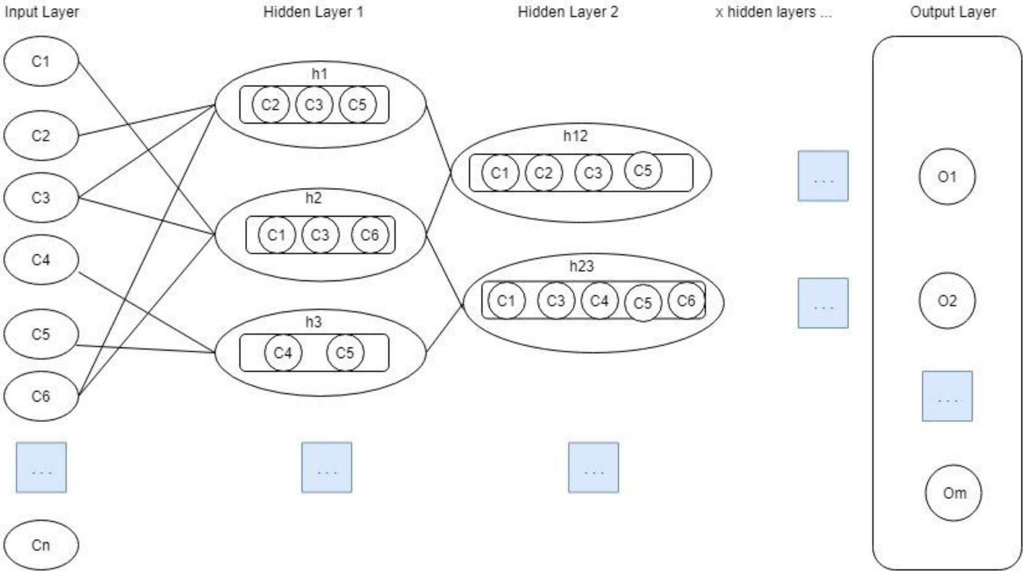


Figure 3.2: Proposed Methodology

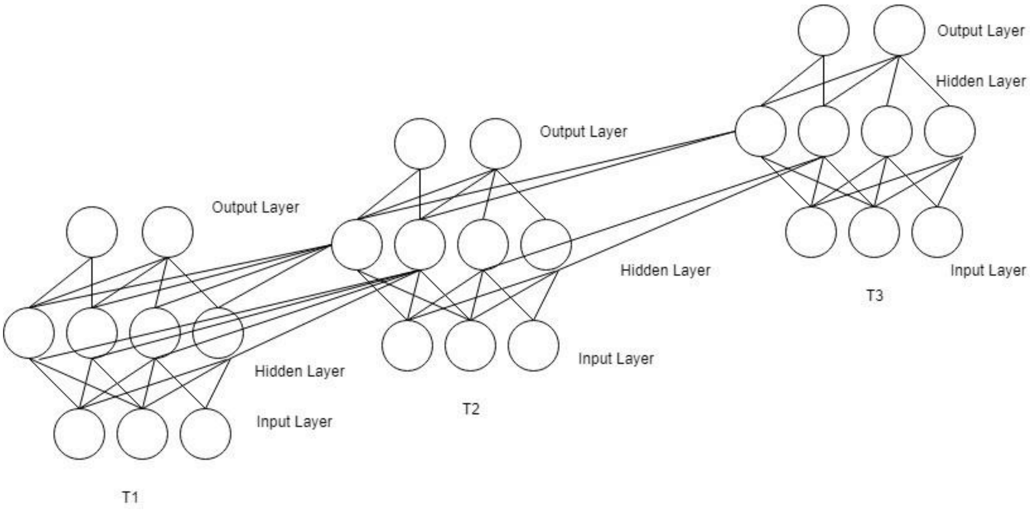


Figure 3.3: LSTM Architecture

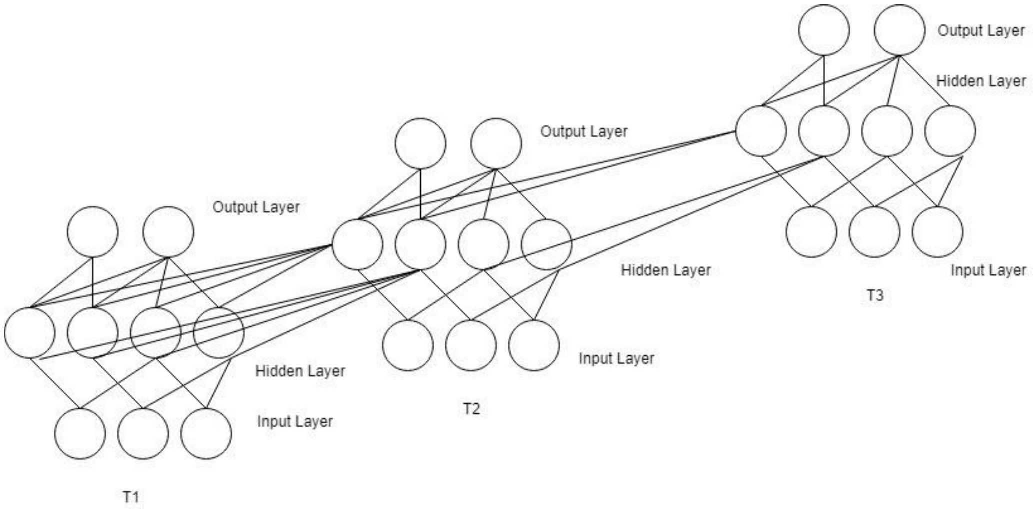


Figure 3.4: Sparsed RNN Architecture

4. Experimental Results

Dataset

Data collection was done by web scarpping health care forums and primary health cared data description sites , some examples of the sites include health- tap, webMD and Medline plus. These sites are used by registered users and medical professionals who provide solutions for the users medical query. These question answer pairs are categorised by the sites based on the dis- ease inferred by the medical experts, thus the data extracted has the disease type as its class which is the file name. The questions belonging to same disease groups are put in a text file and are further processed for the sig- natures. We considered the disease concepts from everyone healthy as our knowledge base. Each disease type in every one healthy has Sign/Symptoms, Drugs, Treatments, Attributes, Further Tests, Lifestyle Changes, etc. We searched every disease concept in everyone healthy in our json file, if there is any match we created a text file with name as that of disease concept and first line will be question and second line will have answer. When we extend this for all disease concepts, we got the odd numbered lines with the question and even numbered lines with the answers.

The next step is to obtain the noun phrases out of the question and answers, which are extracted in the previous step. The regular expression for the noun phrases can be seen in (1). We used nltk library to extract the noun phrases which are matching the regular expression.

$$(Adjective|Noun) * (Noun \quad Preposition)?(Adjective|Noun) * Noun \quad (4.1)$$

This list of noun phrases will contain medical as well as non-medical fea- tures. In order to remove the non-medical features, we used metamap tool. metamap tool was available in the form of web page which returns type of medical term such as Disease or Syndrome, Symptom, etc. We ran an au- tomation script to get the medical features using selenium.

Next step is to normalize the medical features in order to handle the vocab- ulary gaps. This can be with the help of SNOMED CT and vetmed tool. SNOMED CT will return the index of medical concepts which we need to search in vetmed in order to get the medical concept associated with a medical feature. One medical feature may associated with many of the SNOMED CT indexes. In order to get unique medical concept associated with each of the medical features we used google distance. We calculated google dis- tance between medical features and each of the medical concepts obtained from vetmed. The one which gives maximum google distance value associ- ated with medical feature and medical concept obtained from vetmed will be considered for bridging the vocabulary gap. We used pymedtermino python module to obtain the index files for medical features in SNOMED CT. For the medical concept extraction out of the indexes from the SNOMED CT, we used automation for vetmed website.

4.1.1 Subgraph Mining

Once the medical term are separated from the noun phrases we created sep- arate files with these medical terms based on the disease types. The medical noun phrases will be used to form co-occurrence graphs by taking bigrams and checking how often they occur together in the question pairs if the fre-quency of their co-occurrence is grater then 5 ,then they were stored in a dictionary with the name of the disease type that they belonged to. This dictionary was used to build the graph with medical terms as the vertices and an edge between them with their frequency as the weights of the graph.Once the graph was built the next step was to get the k-cliques in the generated co-occurrence graph where the k value ranged from 2-5.For the first k i.e for 2 we got the cliques as the pairs of medical terms which had a minimum limit of the weight sum.As explained in the previous section we found the sub- graphs in the co-occurrence graphs and if the frequency or the edge weight sum is more the threshold value then sub-graph was accepted and stored in a csv file with name of file as the disease type which they belonged to. The step is repeated by adding the increasing the k value which means increasing the number of terms upto 5 terms together. The first set of sub-graphs gen- erated with k=2 to 5 were used as input to the next set and another set of sub-graphs were generated. The sub-graphs generated at 3 sets became part of the 3 hidden layers that we used in neural network architecture.

4.1.2 Training

Once the medical terms are normalized using the procedure given above we use the normalized medical features as the input for the sparse neural net- works. We created the proposed model using python scripts where the training function takes input matrix as a 2 dimensional matrix where the normalized medical terms are represented as word vectors and the weights for these neu- rons are randomly assigned using built-in functions. The size of thismatrix will be C x M where C is the number of normalized terms and the M is the length of the bag of words that we have initialized at the beginning. The word vector has the bit 0 or 1 depending on whether the input normalized term is present or not in the bag of words.We created function for calculating sig- moid values for given synapse values and input vector. The other parameter that we need to pass is the output matrix of length N x N where again N is number of classes , each row in the matrix represents which class the given input sample belongs. We will then have to import the hidden layer matrices where each of them has the vector representation of the signatures that we mined using the technique explained in the previous section. Then we created a function that would add the previous hidden layer values to the input layer vector and the resultant would act as the new input layer for the next time frame or inner iterations as given in our program. Once all the layers are ini- tialized we assign the learning rate range which we used as 0.01 to 0.1 and trained over 10x10000 iterations once the delta which is the average error is minimized the training stops and stores the synapse values in a json file which could be used to process a user query by activating a sigmoid function in order to determine the disease type that the query is given.

$$E(v,h) = -\sum_{i \in visible} a_i v_i - \sum_{j \in hidden} b_i h_i - \sum_{i,j} v_i h_j w_i j \text{..................}(1)$$

$$p(v) = \frac{1}{Z} \sum_h e^{-E(v,h)} \text{ , where } Z = \sum_{v,h} e^{-E(v,h)} \text{..................}(2)$$

$$p(v) = \frac{\sum_h e^{-E(v,h)}}{\sum_{v,h} e^{-E(v,h)}} \text{..................}(3)$$

$$\Delta w_{ij} = \alpha(< v_i h_j >_{data} - < v_i h_j >_{model}) \text{..................}(4)$$

The table given below gives us an intuition about how accurate he models are with respect to a n-fold cross validation which we have used in our program. We can infer that the sparse models perform much better than the strongly connected neural network models. We can justify this trend by saying that the model trained may not be able to get results for inputs which are in different orders for the case of simple neural networks and SVM which led us to use the LSTM model which can store the context of the inputs for further training. But when we use the data which has the similar features to predict the disease type it becomes quite difficult to get an accurate result which is avoided by sparsely connecting the input and the hidden layers , which means that the grouping of normalized terms or signatures as the hidden layer nodes, which in turn doesn’t constitute the complete input layer connections can help reducing the contexts where similar features of disease for different diseases can be avoided. For example if I give query consisting of headache, vomiting and nausea I might end up with migraine, brain tumour and some other disease with equal probability when we used the strong con- nected neural networks and LSTM. But when you process the same query in a sparse model it would generate a signature with the three features occurring together and thus uniquely displaying migraine as my disease type, this is one scenario in which a sparse method may give a better context identifi- cation than that of the LSTM networks. With combined benefits of context learning and sparse connections the model that was built in our work gave much better inference then the other models. Thus finally with proper fine tuning the model created by our work can lead to accurate results.

The feature selection methods have always been a forte of restricted Boltz- mann Machine which train the neural network over the hidden layer to gen- erate the input layer thus providing the sets of weights which will knock off the irrelevant neurons in hidden layers thus providing us a context which can help in better the prediction, although this works well over less data, when more datasets are to be trained the accuracy is deteriorated since the probability values are no very different for the neurons with correct and incorrect contexts. The complete results can be summarized in the graph given below.

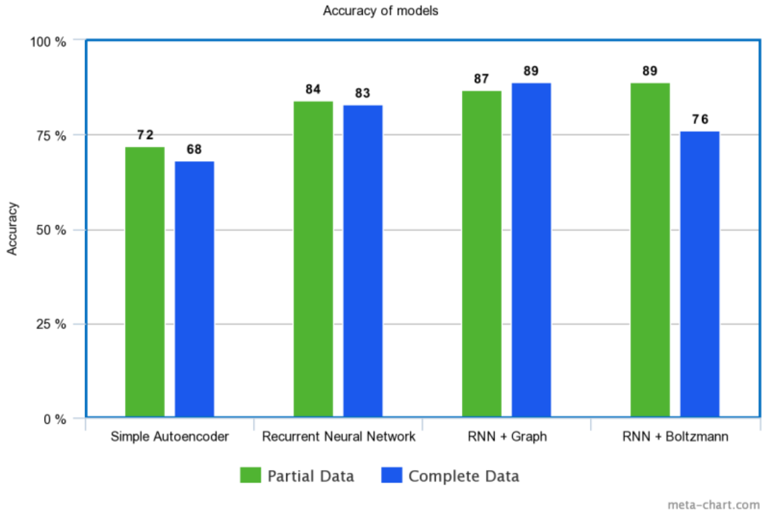


Figure 4.1: Graphical comparison of the accuracy

Table 4.1: Comparison of accuracy of the models

Algorithm / technique	Accuracy on partial data	Accuracy on complete data
Simple Autoencoder	72.88%	68%
Recurrent Neural net-Work	83.88%	82.78%
Recurrent Neural network + Graph mining	87.88%	88.28%
RNN + Boltzmann Machine	89.58%	76%

5. **Conclusion and Future Work**

The research's main objective was to meet the vocabulary gap between the user and medical expert, which the sparse neural network successfully ac- complished. The algorithm built in this research , helped analyze the vague description of the disease and normalize the same to predict exact disease type of the medical conditions of the user.The results of the algorithm were compared with the deep learning algorithm such as LSTM and shallow learn- ing algorithm such as SVM. The algorithm produced accuracy similar to the deep learning algorithms over medical records with accurate inputs , thus suggesting that sparse neural network may be as good as the deep learning networks for the electronic health records. One other method to produce the results of sparsely connected neural network is to use a optimizer such as restricted Boltzmann machine for removing neurons which dont provide enough contribution for the predictions. A responsive system wherein fur- ther response could be asked to the user after inferring on primary query to get better and accurate results like, for example when i give symptoms for a disease like migraine it can further ask me whether symptoms related to migraine are faced by me which could also include a module on image based classification wherein the user may be prompted to either take a picture of affected part in order to give a better validated results, which could be future work associated with this project work.

6. Bibliography

[1]S. Doan and H. Xu, "Recognizing medication related entities in hospital discharge summaries using support vector machine," in Proc. Int. Conf. Comput. Linguistics, 2010, pp. 259–266.

[2]T. D. Wang, C. Plaisant, A. J. Quinn, R. Stanchak, S. Murphy, and
B. Shneiderman, "Aligning temporal data by sentinel events: Discover- ing patterns in electronic health records," in Proc. SIGCHI Conf. Human Factors Comput. Syst., 2008, pp. 457–466.

[3]M. Shouman, T. Turner, and R. Stocker, "Using decision tree for diagnos- ing heart disease patients," in Proc. 9th Australasian Data Mining Conf., 2011, pp. 23–30.

[4]A. Khosla, Y. Cao, C. C.-Y. Lin, H.-K. Chiu, J. Hu, and H. Lee, "An integrated machine learning approach to stroke prediction," in Proc. 16th ACM SIGKDD Conf. Knowl. Discovery Data Mining, 2010, pp. 183–192.

[5]N. Limsopatham, C. Macdonald, and I. Ounis, "Learning to combine rep- resentations for medical records search," in Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2013, pp. 833–836.

[6]S. Ghumbre, C. Patil, and A. Ghatol, "Heart disease diagnosis using support vector machine," in Proc. Int. Conf. Comput. Sci. Inf. Technol., 2011, pp. 84–88.

[7]S. Fox and M. Duggan, "Health online 2013," Pew Research Center, Sur- vey, 2013.

[8]"Online health research eclipsing patient-doctor conversations," Makovsky Health and Kelton, Survey, 2013.

[9]T. C. Zhou, M. R. Lyu, and I. King, "A classification-based approach to question routing in community question answering," in Proc. 21st Int. World Wide Web Conf., 2012, pp. 783–790.

[10]D. A. Davis, N. V. Chawla, N. Blumm, N. Christakis, and A.-L. Barabasi, "Predicting individual disease risk based on medical history," in Proc. 13th Int. Conf. Inf. Knowl. Manage., 2008, pp. 769–778.

[11]L. Nie, M. Wang, Z. Zha, G. Li, and T.-S. Chua, "Multimedia answering: Enriching text qa with media information," in Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2011, 695–704.

[12]L. Nie, M. Wang, Y. Gao, Z.-J. Zha, and T.-S. Chua, "Beyond text qa: Multimedia answer generation

by harvesting web information," IEEE Trans. Multimedia, vol. 15, no. 2, pp. 426–441, Feb. 2013.

[13]L. Nie, Y.-L. Zhao, X. Wang, J. Shen, and T.-S. Chua, "Learning to recommend descriptive tags for questions in social forums," ACM Trans. Inf. Syst., vol. 32, no. 1, p. 5, 2014.

[14]D. Zhang and W. S. Lee, "Extracting key-substring-group features for text classification," in Proc. 12th ACM SIGKDD Conf. Knowl. Discovery Data Mining, 2006, pp. 474–483.

[15]M. Galle, "The bag-of-repeats representation of documents," in Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2013, pp. 1053–1056.

[16]Nie, Liqiang, 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, 2015.