International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-4, February 2020

Disease Inference from Health-Related Questions using Sparse Neural Network and LSTM

M. Dhanamalar, K. Kavitha

ABSTRACT: ***There are many relationship and context retrieval related problems that are resolved using medical records with Deep learning-based techniques, but when it comes to community data such as health care forums where the quality of data cannot be met as health records due to gap in medical vocabulary it becomes impossible to provide an accurate solution, our research involves in providing a solution to this problem. The graph algorithms have always provided the best solutions to map and normalize in NLP domain, we have used the same to find the normalized medical terms for out user questions in the health care forums. Instead of training the actual data from the forums with the LSTM, we create medical signatures of the words coming together to form a context in the medical dictionary. We consider the words used in the dataset as vertices and find dense subgraphs to uniquely identify the condition with medical dictionary data. In simple words, we aim to build a system to convert the vague descriptions of the disease to match the accurate medical term from the medical dictionary such as snomedCT. We used the words which co-occur to define our relations which will, in turn, provide us with a solution to bridge the gap of medical vocabulary. The mappings of normalized terms are foundations to build the hidden layer of our neural networks, instead of constructing a direct connection between the input neurons to all hidden neurons we connect only the subgraph results thus improving our accuracy to a better level then existing methodologies.***

Keywords: Machine Learning; Recurrent Neural Networks; Sparse learning; NLP; Sub Graph Mining; Health care; Bioinformatics;

I. INTRODUCTION

Introduction In this generation of time the cost of medical services has risen to large heights, with the increasing population the human medical advisers have got to do a tremendous job to meet the demands of treating the diseases. This problem paved way to online forums related to medical queries to rise, wherein the users will post the symptoms of the problems they face in the body, given the questions on the forums there various medical experts in the forums who are registered based on their authority. Even this requires doctors to manually provide medical solutions. These type of forums also sometimes may not be useful because the users may not be able to convey the medical problem they are facing due to lack of correct vocabulary.

According to a survey almost there are 70 patients to be attended by a doctor in a day this number is only for the urban region spanning of rural most of them don't even have a proper medical practitioner.

Revised Manuscript Received on February 28, 2020.
***** Correspondence Author
Ms. M. Dhanamalar*, Assistant Professor, Department of Computer Science, Kristu Jayanti College, Bengaluru, India.
Dr. Kavitha. K, Assistant Professor, Mother Teresa Women's University, Kodaikanal, Tamil Nadu, India.

The current prevailing online health resources can be roughly categorized into two categories.

One is the reputable portals run by official sectors, renowned organizations, or other professional health providers. They are disseminating up-to-date health information by releasing the most accurate, well-structured, and formally presented health knowledge on various topics. While the other is the forums where the user asks a question and multiple solutions some relevant and others often misleading The problems gave us the inspiration to provide an automated medical inference system which would learn from the question-answer pair which is collected from various health-related websites which have the label of disease assigned by authorized doctors.

1.1 Motivation

The biggest stumbling block of the automatic health system is disease inference. According to our user study on 5000 questions that health seekers frequently ask for:

1. Plausible inference for their diagnoses.
2. Method to prevent the condition of a diagnosed disease.
3. possible diseases with the signal of their medical conditions.

The former two genres usually involve the exact disease names and expected sub-topics or sub-problems of the given diseases, such as the side effects of specific medications, and treatments. They can be automatically and precisely answered by either directly matching the questions in the archived repositories or syntactic information extraction from the structured health portals. The third genre conveys parts of the health seekers' demographic information, physical and mental symptoms, as well as medical histories, in which they do not know what conditions they might have and expect the doctors to offer them some forts of online diagnosis. If the diseases are correctly inferred, these questions are naturally transferred to the first genre.

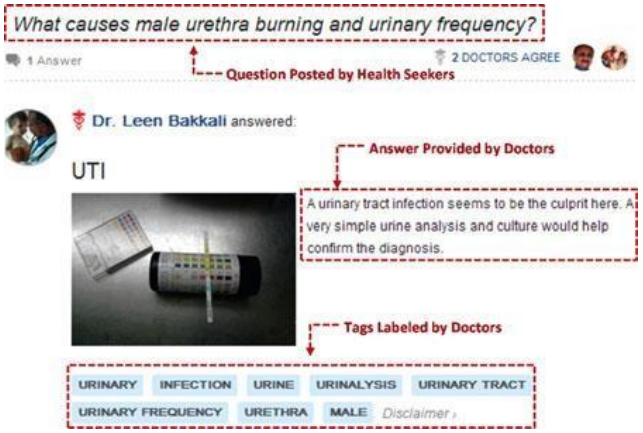

Figure 1. Example-Question Answer pair from Community based health services[1]

Hence a robust disease inference approach is the key to break the barrier of automatic wellness systems. However, little research has been dedicated to disease inference in community-based health services. Disease inference is different from topics or tags assignment to short questions an example to which is given in Figure 1, where topics or tags are direct summarizations of given data instances and they may explicitly appear in the questions.

While disease inference is a reasoning consequence based on the given question, this task is nontrivial due to the following reasons. First, the vocabulary gap between diverse health seekers makes the data more inconsistent, as compared to other formats of health data. For example, "shortness of breath" and "breathless" were used by different health seekers to refer to the same semantic "dyspnea". Second, health seekers describe their problems in short questions, containing 14:5 terms per question on average.

The incompleteness hinders the effective similarity estimation based on shared contexts. Third, medical attributes such as age, gender and symptoms, are highly correlated and do not unusually appear as compact patterns to signal the health problems. For example, "tight chest", "wheezing", and "dyspnea" frequently co-occur to the hint of "asthma". In addition, it is expensive to construct the ground truth for various diseases. These factors limit the disease inference performance that can be obtained by general shallow learning methods. Shallow learning methods refer to the most modern learning algorithms such as decision trees and support vector machine (SVM), where the output of a learning scheme is directly followed by a classifier as if the system has only one layer.

This paper aims to build a disease inference scheme that is able to automatically infer the possible diseases of the given questions in community-based health services. We first analyze and categorize the information needs of health seekers. As a byproduct, we differentiate questions of this kind that require disease inference from other kinds. It is worth emphasizing that large-scale data often leads to an explosion of feature space in the lights of n-gram representations, especially for the community-generated inconsistent data.

To avoid this problem, we utilize medical terminologies to represent our data. Our scheme builds a novel deep learning model, comprising two components. The first globally mines the latent medical signatures. They are compact patterns of inter-dependent medical terminologies or raw features, which can infer the incomplete information. The raw features and signatures respectively serve as input nodes in one layer and hidden nodes in the subsequent layer. The second learns the inter-relations between these two layers via pre-training. Following that, the hidden nodes are viewed as raw features for more abstract sig-nature mining.

With incremental and alternative repeating of these two components, our scheme builds a sparsely connected deep learning architecture with three hidden layers. The detailed literature survey carried out can be summarized in the table below.

II. PROPOSED METHODOLOGY LOCAL MINING

The main hurdle in translating the user query's vocabulary to original medical concepts is a repetition of the same medical terms in multiple diseases, to overcome this we have proposed the method called as signature mining. Once the medical terminologies in the question-answer pairs are extracted they are to be normalized, since training the raw medical terminologies against the disease types may not yield the best possible results in which case multiple diseases may have the same terminologies like for example both brain tumour and migraine may have common medical terms such as severe headaches, passing out and nausea in which case the model may not be able to get the best results.

Table 1. Literature Survey

Name of the Paper	Merits	Demerits
Recognizing medication-related entities in hospital discharge summaries using SVM S.Doan1 2010	Application of various Machine Learning algorithms includes CRF and SVM on established F-score of 90% in i2b2 Challenge	i2b2 challenge mainly constituted of HER datasets, but when the same algorithm applied over online data they failed to bare well.
Aligning temporal data by sentinel events: Discovering patterns in EHR T. D. Wang2 2008	Mainly the target of the paper was to establish rank, align the health records based on co-occurrence which led to a good foundation other than Machine Learning	It is failed when the visualization tool couldn't establish relations when deployed on regular datasets other than EHR.
Using decision tree for diagnosing heart disease patients M. Shouman[3] 2011	Introduced Data mining techniques like decision tree based on gini indes, information gain and gain	Decision tree found to be inefficient in case of forming patterns out of signatures for disease inference.

An integrated machine learning approach to stroke prediction A. Khosla[4] 2010	Introduced margin based censored classifiers.	This kind of feature extraction didn't yield proposed results with SVM's for online data.
Learning to combine representations for medical records search N.Limsopatha[5] 2013	Proposed a novel approach for retrieving resources then the traditional methods like Bag of words and Bag of concepts.	The Information Retrieval can be ineffective in case of improper vocabulary or terms used in queries.
Heart disease diagnosis using support vector machine S. Ghumbre[6] 2011	Introduced radial based functions and orthogonal least square with automated learning methods.	RBF and OLS requires proper relationship extractions which isn't available in case of our dataset, it may be ineffective.

To over come this issue we introduce the concept of signature mining in which the word vectors formed by the grouping of co-occurring terms could uniquely identify a disease type thereby improving the model's accuracy. We have the medical terms that are used in questions which we will normalize and store the normalized medical terms in separate files. The extracted medical terms are then used to plot graph over the question-answer pairs in which we take the medical terminologies as the vertices and if the medical terms say M1, M2 says occur together in a question then the two vertices will have an edge between them and this would happen when the co-occurrence value is at least the minimum threshold value which we considered as 5 which means that the pair of medical terminologies which we are considering- ing must occur at least 5 times together so that can form an edge in our co-occurrence graph. With this method, we take all the possible combinations of medical terminologies co-occurrence and build our final co-occurrence graph. Once this graph is constructed we used something called a k-clique algorithm to find the densest subgraphs among these edges. The resulting k-clique will be our final signature.

k-clique

For a give co-occurrence graph we will traverse one by one each vertex and for k valued adjacent vertices we check if there is an edge and finally if there is an edge then the resultant graph is checked for if it is a complete graph or not. For the first step the 2-cliques can be simply generated by getting the vertices enumerated, in our case, we will consider the vertex names as our medical terms and if the weight of the edges forming the cliques is again checked if grater then the threshold which in our case would be 10. So from a natural language perspective what are these cliques or the n-gram words is these words may co-occur in the question thus they constitute our signatures. For the 3 and 4-cliques, we consider the fact that the edges exist between all the vertices participating that means that they are complete graphs, if they are not complete graphs then they are discarded. These subgraphs are then stored as separate files. Once the subgraphs are mined completely then the medical terminologies participating in the k-cliques are replaced by the corresponding normalized terms, this is the final result which will constitute our hidden layer's neurons. This procedure is repeated by taking the k-cliques as our input again we mine further k-cliques our this input(i.e previous k-clique results) to get the corresponding intermediate hidden layers.

III. IMPLEMENTATION

We scrapped question and answer data in the form of a json file which had three fundamental tags i.e. question, answer and the name of the disease, from healthTap, WebMD and eHealth forums. We preprocessed it to remove words with uni- code format and once it was preprocessed we removed unnecessary URLs and pushed the main attributes to a text file. The next step was to get the noun phrases from the questions in the text file, we have done this using Natural Language Processing library NLTK in python. The noun phrases included the terms which represented a disease so, for each word in this noun-phrase, we found the synonym from International Medical Dictionary SNOMED using a library called pymedtermino. After getting the results from this SNOMED we redirected them to Metamap Tool.
We used Web API of Metamap tool to collect information from this web API we used automation and web scrapping combined hybrid model.

Automation was done using Selenium and Web Scrapping was done using Beautiful Soup. So finally we got normalized medical features. For these normalised medical features we found the co-occurrence by taking a bigram against the question and answer pair with a line by the line search function.

Once the co-occurrence was found we created a co-occurrence graph for the same. Over this graph, we are applying a greedy algorithm to get dense subgraphs which will be our signatures for training the Neural network to provide a disease inference.

We needed normalised medical features for the neural network. That can be obtained from everyonehealthy.com website. We got the frequency of each symptom of each disease with Pubmed and LDC dictionary, and if found that frequency in Pubmed dictionary is more then we will add it to our normalised feature set. we have considered a small number of medical concepts for training which contains 871 data entries which belong to 9 classes.

We mentioned earlier that the files are stored with the class as their file names which also acts as output and each entry in the file is a normalised medical feature, these are the input to our neural network. Each word in the text files is stemmed using Lancaster stemmer of nltk library in python, after which we are left with 805 unique words which represent the normalized medical features. The is a binary vector of size 805(bag of total words) which takes the value 1 if the word in the input entry is present in the bag and 0 otherwise.

We created a simple autoencoder with one hidden layer which takes random values and has a dimension of 910 neurons and a sigmoid function as the activation function. We trained the neural network for 30000 iterations to reduce the mean error in detecting the disease. After training we stored the weights in js on file and when we get an input string it will convert the input string into a vector of 0's and 1's of size 805, applying sigmoid function with weights stored and the given input string vector we receive the vector representation of the class and finally a disease concept with the highest probability is returned.

IV. EXPERIMENTAL RESULTS

In Figure 2, we have shown the scrapped data sample from health tap which is of JASON format. From this file, we have removed some unnecessary tags and have considered only questions and answers and parsed the same to a text file which is shown in Figure 3.

After getting the text file we then collected the noun phrases from both the question and answers using NLTK library in python which is shown in Figure 4. Of these noun phrases, each word was searched against a medical dictionary called SNOMED CT using Pymedtermino the results of the search can be seen in Figure 5.

The returned objects from SNOMED were then searched over METAMAP and those results are shown in Figure 6. These results are the normalized medical attributes we then create a co-occurrence graph between the normalized terms and the QA pairs, an example of the same is shown in Figure 7.

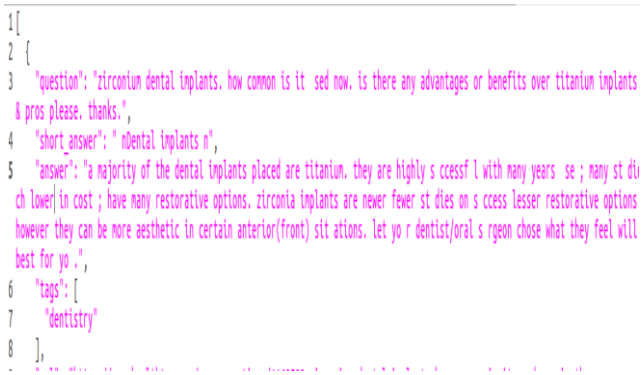

Figure 2. Question-Answer pair from healthTap

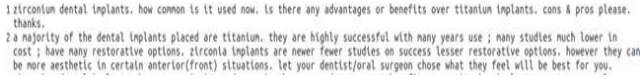

Figure 3. Preprocessed Question-Answer Pairs from healthTap

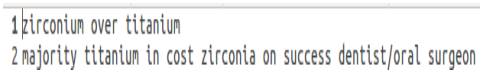

Figure 4. Noun- phrases using NLP

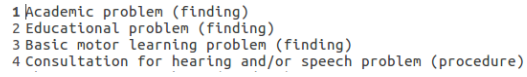

Figure 5. Medical Features From SNOMED International Browser

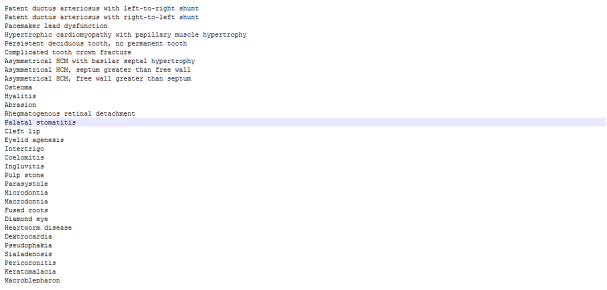

Figure 6. MetaMap results for SNOMED data.

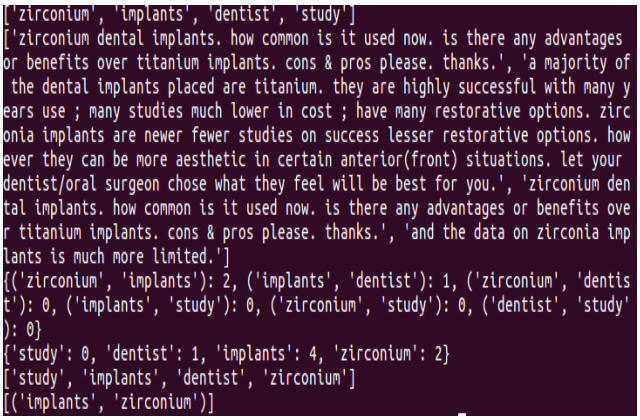

Figure 7. Overlapping subgraphs

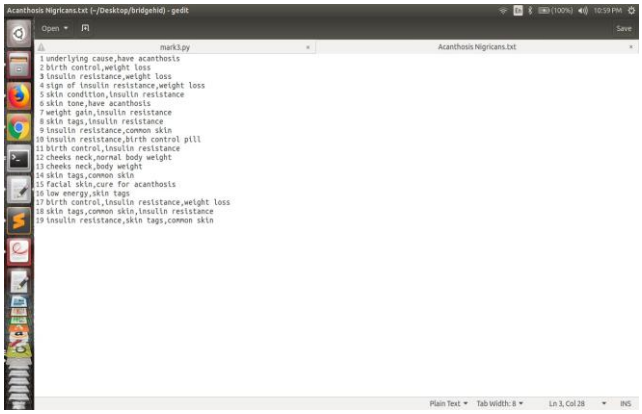

Figure 8. Subgraphs Formed for Pair of Words

As shown in Figure 9. we have considered a small number of medical concepts for training which contains 871 data entries which belong to 9 classes.

We mentioned earlier that the files are stored with the class as their file names which also acts as output and each entry in the file is a normalized medical feature, these are the input to our neural network. Each word in the text files is stemmed using lancester stemmer of nltk library in python, after which we are left with 805 unique words which represent the normalized medical features.

The is a binary vector of size 805(bag of total words) which takes the value 1 if the word in the input entry is present in the bag and 0 otherwise. We created a simple auto encoder with one hidden layer which takes random values and has a dimension of 910 neurons and a sigmoid function as activation function. We trained the neural network for 30000 iterations to reduce the mean error in detecting the disease. After training we stored the weights in json file and when we get an input string as shown in Figure 10.

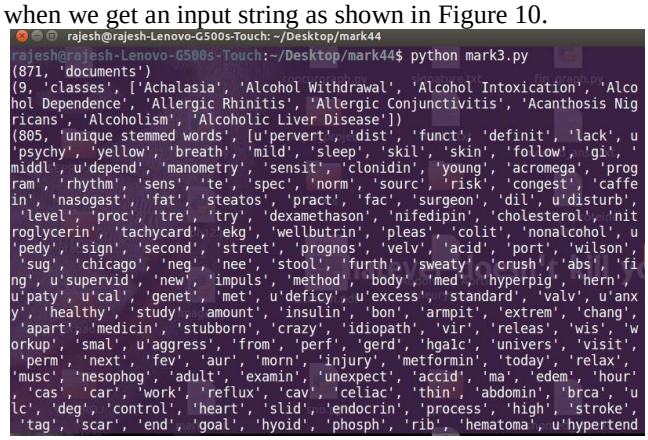

Figure 9. Stemmed Features of Normalized Medical Features

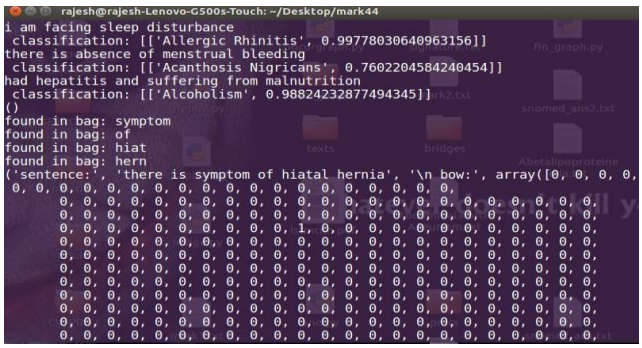

Figure 10. Classification of Input Queries

It will convert the input string into a vector of 0's and 1's of size 805, applying sigmoid function with weights stored and the given input string vector we receive the vector representation of the class and finally, a disease concept with the highest probability is returned as shown in Figure 11. In Figure 12 we can see that for input string "there is a symptom of hiatal hernia" we see that our code was able to get the words symptom, hiat and hernia in the bag thus corresponding bits (in the vector of 805 bits) will be turned 1s and this vector is taken as x in the sigmoid function with weighs from json file the result of this would return us a vector of size 9 bits whichever bit has 1 in the vector that bit represents the class in which the input string belongs which can be illustrated with Figure 10,11,12. In Figure 13 shows the classification of general neural network with one hidden layer. Figure 14 and 15 shows the automation of meta map. Figure 16 shows the normalized medical features against the features present in the dataset. Figure 17 and 18 shows the model accuracy for sparse neural network and LSTM respectively.

Figure 11. 0-1 Mapping of Stemmed Features

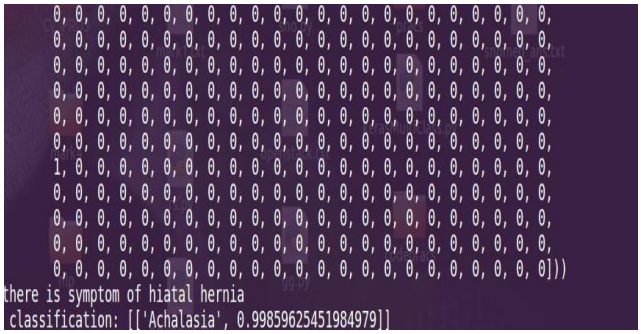

Figure 12. Classification of Queries and 0/1 mapping of each input feature.

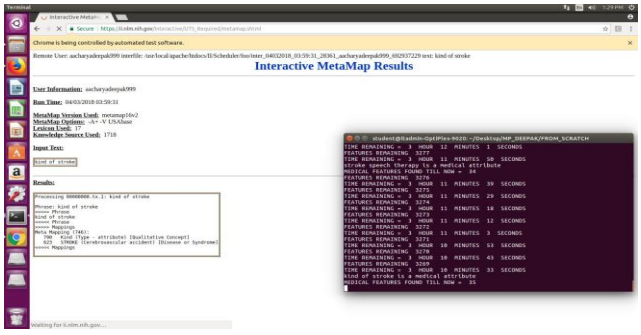

Figure 13. Automation of Metamap which results in a medical attribute.

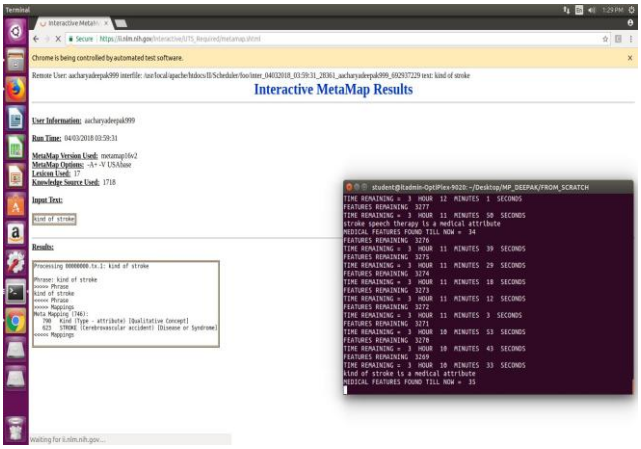

Figure 14. Automation of Metamap which results in a medical attribute

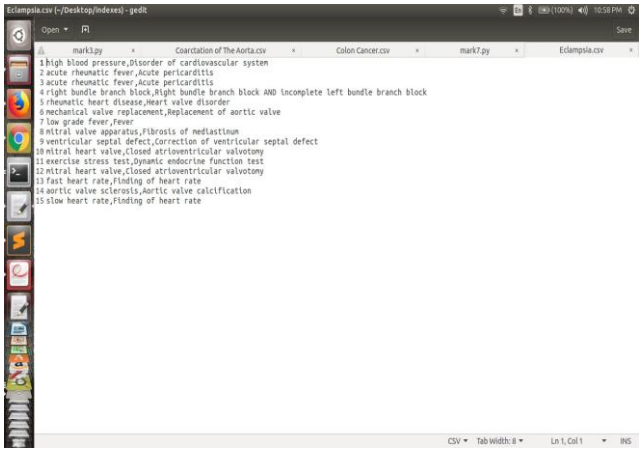

Figure 15. Normalized Features after Metamap

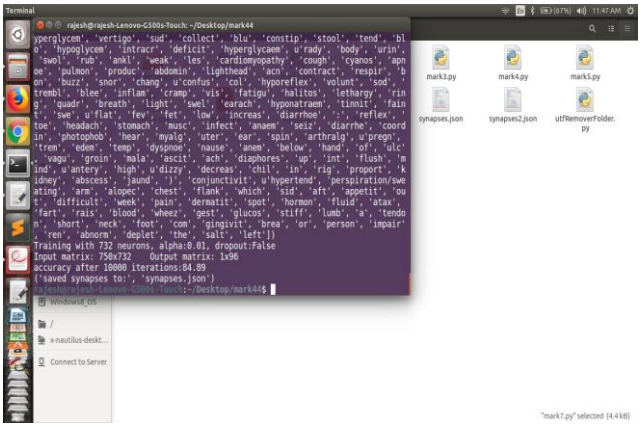

Figure 16. Sparse Neural Network Model Accuracy

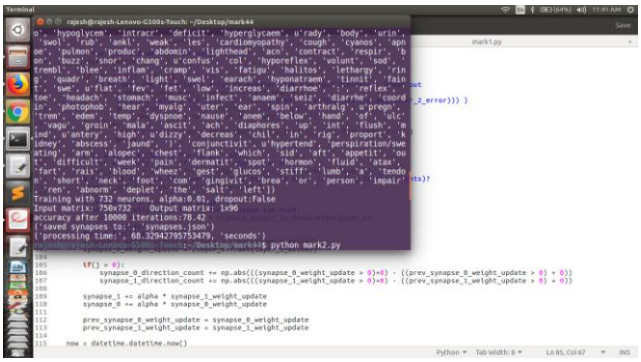

Figure 17. LSTM Model Accuracy

V. CONCLUSION

The problem of learning the features of irrelevant context arising it community data cannot be resolved with LSTM but with our methodology we have achieved the construction of language model to fetch best results out of forum data and feed to LSTM to automate the inference instead of a medical expert answering the questions with the knowledge it acquires by training over millions of QA pairs. While the LSTM has its own limitations, on the other hand, the sparse network's powerful context is of no use in case of the huge dataset when the retention rate has to be of the adequate amount. The vocabulary gap is bridged by the signature mining and the recurrent property of proposed neural network will provide a better lasting context which would solve the problem of limited context retention in the regular network used by the previous research work. The sparse connections in the network will limit the hidden layers nodes from providing inefficient weights to the context of less importance and thus improves the efficiency in learning the required knowledge.

REFERENCES

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: Discovering 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 diagnosing 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 representations for the 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, Survey, 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.

AUTHORS PROFILE

Ms.M.Dhanamalar, Research Scholar Mother Teresa Women's University, Kodaikanal. She is currently working as Assistant professor, Computer Science Department in Kristu Jayanti College, Bengaluru.

Dr.Kavitha. K, Assistant Professor, Mother Teresa Women's University, Kodaikanal. She has published more than fifty papers in International Journals and National/International Conferences. She has published two book chapters. Her research interest are Data Mining, Cloud Computing, Data Science.