ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2018.0232

AN APPROACH FOR AUTO-GENERATING SOLUTION TO USER-GENERATED MEDICAL CONTENT USING DEEP LEARNING TECHNIQUES

Faraz Bagwan and Leena Deshpande
Department of Computer Engineering, Vishwakarma Institute of Information Technology, India

Abstract

One of many things humans are obsessive about is health. Presently, when faced with a health-related issue one goes to the web first, to find closure to his/her problem. The community Question Answering (cQA) forum allows people to pose their query and/or discuss it. Due to alike or unique nature of the health query it may go unanswered. Many a time the answers provided are ill-founded, leaving the user discontent. This indicates that the process is dependent on supplementary users or experts, in relation to their ability and/or the time taken to answer the question. Hence, the need to create an answer predictor which provides instant and better-quality result. We, therefore propose a novel scheme where deep learning is used to produce appropriate answer to the given health query. Both historical data i.e. cQA and general medical data are used to form a powerful Knowledge Base (KB), to assist the health predictor.

Keywords:

Community Question Answering, Deep Learning, Health- Related Issue

1. INTRODUCTION

The lives of people have always been a busy one, more so in current technological age. Health has become a secondary concern which is tended only when it's symptomatically visible. With pollution and population, an environment susceptible of allergies, viral fevers, immuno-compromising agents, etc., is in existence and proliferating. This has led to the increase in the number of patients. Additionally, in many less developed countries, healthcare is a business opportunity. Dearth of knowledge in patients about one's symptoms give doctors leeway to trick them. Also, inflation has made most people economically challenged and one thinks twice for diagnosis and tries to outmaneuver the disease. This has changed the approach of patients where instead of visiting the doctor first they research it on the web, mainly health based cQA forums. The cQA forums [1][2][6][12][22] considered here are strictly health based like healthtap, iCliniq, etc., wherein users with health queries acts as patients. These queries are answered by experts i.e. doctors. The problem with such forums is the lack of available answers to any given medical/health-related question. Statistically, an expert in a health cQA forum answers thousands of queries yearly of which many would be similar thus exhausting the expert which may lead to wrong or no answers. The internet is full of forums allowing people to ask these questions, from which some of them retrieves great answers. However, browsing these questions to locate one similar to your own, and also finding one with a satisfying answer is a hard and time-consuming task, and posting the question yourself is even worse. By automating the process of generating answers to these questions, and thereby creating a kind of "digital health responder" one could solve this problem. There are many challenges to use cQA data and general medical data. The unstructured nature of the data consists of noise and redundancies and needs to be preprocessed. Also, the language used might have semantic and syntactic complications thereby creating a lexical gap [21]. For instance, the question, "what can I not eat to reduce fat?", is similar to, "what can't I eat to reduce fat?", and to, "What foods can make me fat?", but has a completely different answer than the question, "what can I eat to reduce fat?". Training the model to distinguish such minutiae is hard.

In this research work, we explore the utilization of Long Short Term Memory (LSTM) based Recurrent Neural Networks to the Short Text Conversation issue [7] [8]. Particularly the assignment of creating answers to medicinal inquiries is examined. Data for training is assembled from online administrations containing client produced health inquiries with related answers created by experts. The proposed models are trained and evaluated using data originating from different online services like WebMD, HealthTap, and iCliniq, from which an optimal dataset is determined. The models proposed to tackle the assignment are for the most part roused by models used in Neural Machine Translation but also contains extensions based on Transfer learning and Multi-task learning, are all in view of the Encoder Decoder system having an encoder figure an inert vector portrayal of an inquiry from which the condition of a decoder is instated and used to create an answer and are altogether prepared end-to-end. One model architecture containing a decoder RNN with two "modes", controlled by a binary input to the decoder, are proposed. One mode being a "language-model" mode where the decoder is trained on general medical/health-related text, and another being a "answer generating" mode where the decoder learns to generate answers to given encoded questions. Another model architecture handling the answer generation task and the task of classifying question categories as a combined task is proposed, where the network classifies the question category from the final state of the encoder and feeds the predicted class as an extra input to the decoder. In the end, a novel language model, based on RNN is prepared using general medical content and used to help the proposed models amid derivation by combining their probabilities at each time-venture in an endeavor on enhancing the produced answer quality.

2. RELATED WORK

The cQA forums can be categorized into two sorts [6] [12]. They are general and specific. Our work is domain-specific i.e. health, hence we have immensely reviewed the body of work done in the health domain. Apart from QA based data we also reviewed some work which provided some insights into predictive diagnosis.

Xu [10] made use of SVM in combination with a rule-based system to extract and classify Named Entities (NE) and its relations (NER) using the hospital discharge summaries.

Lee et al. [23][24] proposed a method to effectively learn the hospital health records and extract signatures with respect to time in order to correctly find the relationships between single and/or multiple events. A temporal-event matrix was used to find meaningful relations among data samples using approximation methods.

Sondhi et al. [18] proposed a framework called Sympgraph which is used to extract relations between a sample of patients' records consisting of their symptoms and the time they occurred. The graph is constructed using vector space model where distance is calculated to find similarity between symptoms and time. It also allows classification, analysis and extraction of relationships from patients' temporal-symptoms graph data.

Ghumbre et al. [19] proposed a method to predict heart disease by learning 10 years of patients' data bearing heart problems. They used Decision tree and SVM respectively for their experiment.

Liu [15] proposed a method to effectively perform classification of diseases using semantic text mining. Shortcomings of the traditional information retrieval classification method [17] were discussed where the bag-of-words method seemed insufficient for proper classification. Also, tests were performed on health data with semantic features and without it, using SVM and Naive Bayes, and semantic text learning provided better results.

Huber [13] proposed a deep learning method i.e. a sparse auto-encoder framework of neural network to provide better cancer diagnosis and prediction from gene data. The process is basically divided into 2 steps i.e. feature learning and classification learning.

Li [8] proposed a Neural Network Encoder Decoder framework to auto-generate appropriate responses to a post. Weibo forum was used for data. The system outperformed traditional IR methods and Statistical Machine Translation (SMT) methods.

Khosla et al. [25] proposed a feature selection model in combination with SVM to predict strokes in patients. The method outperformed the current Cox proportional hazards model in an evaluation from ROC curves and concordance index [9].

Zhou et al. [16] proposed multitask based regression approach to predict the progression of Alzheimer's disease in patients. In order to do so, a time based regularizer was used named lasso. It ensured difference in each progression and the time pattern. The ADNI database was used for the same.

Latha et al. [12] proposed an information retrieval based approach for ranking answers in the Yahoo question answering community to improve solution seekers experience.

Basterrech et al. [20] proposed the involvement of GPUs i.e. CUDA, for enhancing the computational ability of prediction models i.e. random neural networks. In comparison with a traditional c program, as a case study, it proved better.

Nie et al. [6] proposed a model i.e. plane, to better predict the answers and rank them based on its data in accordance with the question. The prediction model is trained using KNN algorithm. The ranking is performed using a pairwise learning technique like SVM.

Saradha, [26] proposed an artificial intelligence based human computer interaction system which answered user's question via feature extraction and speech synthesis.

3. RESEARCH DESIGN

The proposed framework depends on RNN's Encoder Decoder model. Although Feed Forward (FF) Neural Network [4] is an intense model, it has its impediments. One being that it is a memory less model, i.e. it has no data about past input sources. RNN has a temporal understanding of its input which is critical to medical inferences [11]. For instance, the sentence, "cause can lead brain exposure damage", makes no sense but the sentence, "Lead exposure can cause brain damage", makes sense as it has order or understanding of time. In a RNN, each layer l won't just pass its output to the following layer l+1, however will likewise pass it to itself, henceforth a "repetitive association", making it a network of recurrent nodes.

The Fig.1, depicts an RNN based Encoder-Decoder framework for the HealthQA system. It basically can be divided into 4 parts: Query Processing, Answer Inquisition, Knowledge Base and Answer Generation.

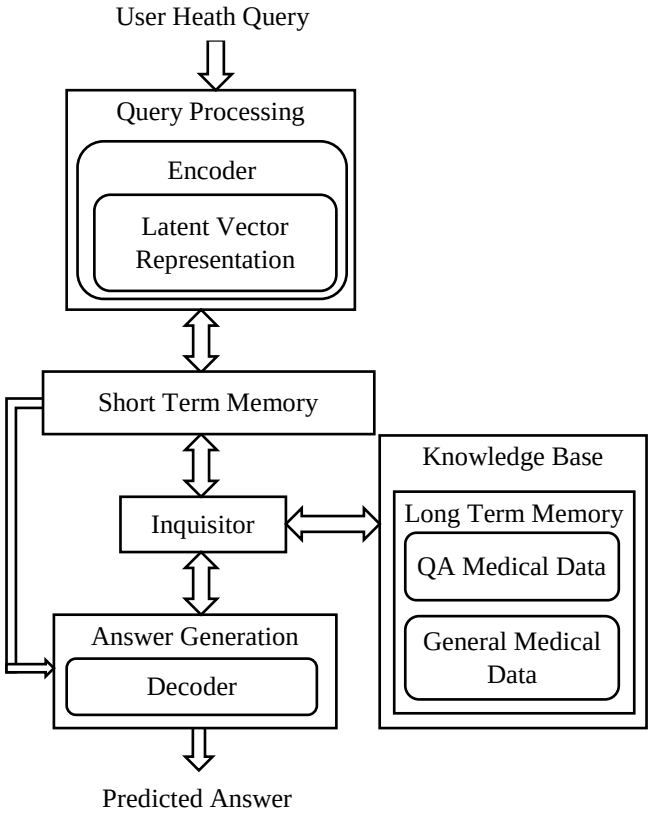

Fig.1. RNN based Architecture with Encoder-Decoder Framework using LSTM/BLSTM models

- Query Processing: The query Q is treated as a sequence used for mapping the answer in the answer generation process. The RNN requires the Natural Language Query in a certain format for processing, hence, it encodes it into a Latent Vector Representation denoted as E_Q. The "Encoder"

eliminates trivial words, processes it into a vector value and stores it in the Short Term Memory for future reference.

- Answer Inquisition: The "Inquisitor" inquires for the related Answer using the vector E_Q from Knowledge Base. It initially carries out a search for similar terms in KB to build a set of possible answers denoted as C_Q. The likelihood between C_Q and E_Q is evaluated further. The answer with the most probability-score, P_C, is provided for further processing.
- Knowledge Base: Initially, health-based QA data and general medical data is scrapped from multiple sources. It is then preprocessed to form a Knowledge Base. This KB is huge and is required for each query, hence, after preprocessing it is stored in the Long Term Memory.
- Answer Generation: The answer generation is the process of forming a sequence of appropriate words from the amalgamation of knowledge and vocabulary. The answer with the most relevance i.e. P_C is processed along with E_Q to form a solution with natural language abilities denoted as *A*. This process is conducted by the "Decoder".

4. EXPERIMENTS

4.1 DATA

Availability of valid, standard dataset was sparse. For this research data was obtained from 3 text corpora:

The Q/A dataset: The question/answer (Q/A) pairs are scraped from WebMD Answers, a website containing categorized questions asked by the users and their associated answers given by experts. The following is an example of a Q/A pair:

Q: "I have a wart on my cheek it has been there for only two weeks. Is there a homeopathic cure I can try?"

A: "Hi you may find you answer on www.earthclinic.com, if not it would be a good starting place. Good luck"

Each question can have multiple answers from specialists or simply from other users, and each answer therefore make up a question/answer pair with the given question. To only obtain good quality answers, the answers coming from a user and not a specialist are filtered based on the users' number of posts, votes, or followers which is available through the page. An answer from a user is considered valid if one of the following criteria are met:

- The user has written more than 50 answers.
- The user has more than 10 votes.
- The user has more than 5 followers.
- Other users have marked the answer as useful.

The extended Q/A dataset: To extend the Q/A dataset some additional websites were used to obtain Q/A pairs, for instance, eHealth, HealthTap, iCliniq, QuestionDoctors. This resulted in, a total of 166,804 Q/A pairs with a total of 66 million characters and thereby an increase in number of observations in turn increasing the diversity of the data, since they originate from different websites.

The PubMed dataset: The dataset contains a mix of text from papers with medicine/health-related topics, and text from Wikipedia articles. Abstract and body text is used from a total of 53,541 papers. The Wikipedia articles mentioned are from within the parent category "Health and fitness portal" and its subcategories, which resulted in a total of 7,077 articles.

4.2 EXPERIMENT METHOD

The general "Encoder-Decoder" framework is used for the experiment. The models will be operating on a character level, by learning character embeddings, and thereby using a small "alphabet" of characters instead of a (usually very big) "vocabulary" of words which is often used in this kind of models. The different models used are,

LSTM: Using regular LSTM cells in encoder and decoder in Fig.2.

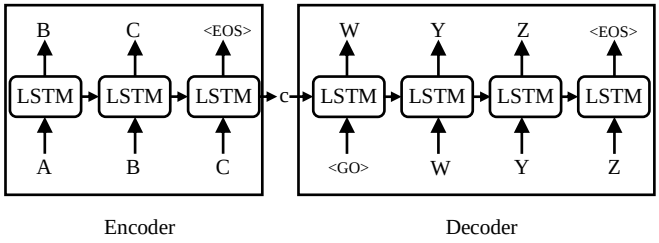

Fig.2. Graphical representation of the LSTM model

BLSTM: Using Bidirectional LSTM cells in the encoder in Fig.3.

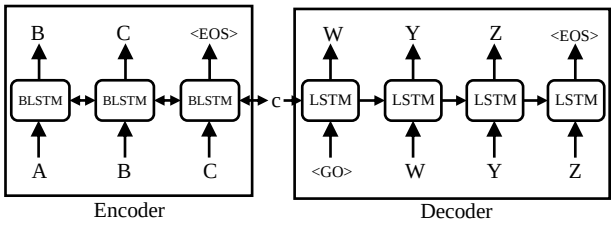

Fig.3. Graphical representation of the BLSTM model.

The prediction or inference phase refers to the generation of an answer candidate from a given question and is the main job of the decoder. The standard procedure is to encode the question using the encoder and set the initial state of the decoder to the final state of the encoder. By feeding the decoder with the <*GO*> token the decoding begins. At each time-step the decoder computes the probabilities of each token in the alphabet, from a given input token, and selects the token with highest probability as its prediction hence using it as input to the next time-step. This is repeated until the <*EOS*> (End of Statement) is predicted. This is a greedy approach since taking the one with highest probability at each timestep doesn't guarantee the most probable sequence to be chosen.

The training of the networks was performed using mini-batches of size 128 with the Adam algorithm. Data was split into training (70%), validation (10%), and test (20%). The networks were all regularized using early stopping based on validation set and gradient norms.

4.3 EXPERIMENT RESULTS

From Table.1, Table.2, Table.3, and Table.4, some examples of predictions can be seen. These predictions were performed on truncated Q/A pairs.

Table.1. Q/A test sample prediction #1

Question	Can a former heart attack (19 yrs. ago) cause an abnormally on ekg when taking a stress test?
Answer	Yes, an old heart attack may still show up on an electrocardiogram, especially when having a stress
Predicted Answer	Yes, but it is not a real time to treat any antibiotic that is related to the area of the time. An

Table.2. Q/A test sample prediction #2

Question	What should i do if i suspect an overdose of humalog kwikpen?
Answer	If overdose is suspected, contact your local poison control center or emergency room immediately. Us
Predicted Answer	If overdose is suspected, contact your local poison control center or emergency room immediately. Us

Table.3. Q/A test sample prediction #3

Question	Is losing weight common while on on the depo provera shot?
Answer	No. It is not common. Actually, a bit of weight.
Predicted Answer	No. It is not normal. It is not a really important to make sure that you are taking and will also

The Table.4, depicts an example of contradicting Q/A test sample prediction.

Table.4. Q/A test sample prediction #4

Question	Can nuclear stress test cause cancer?
Answer	No. It will not cause cancer. I had one myself.
Predicted Answer	Yes. According to the process can be caused by a problem.

Learned template answers: From inspecting the test predictions, it is found that the model has learned a few common answers, which seems to be template answers used by the professionals at WebMD. In Table.5, an example of similar questions being answered with the same common answer, which has been learned by the model, is evident.

Table.5. Q/A test sample prediction #5

Question1	What should I do if I suspect an overdose of agesic?
Question2	What should I do if I suspect an overdose of bidil?
Question3	What should I do if I suspect an overdose of verv?
Question4	What should I do if I suspect an overdose of cal-g?
Question5	What should I do if I suspect an overdose of copd?
Question6	What should I do if I suspect an overdose of desyrel?
.	.
.	.
.	.
Question30	What should I do if I suspect an overdose of albuterol?
Common Answer	If overdose is suspected contact your local poison control center or emergency room immediately. US residents can call the US national poison hotline at 1-800-222-1222. Canada residents can call a provincial poison control center.
Predicted Answer	If overdose is suspected contact your local poison control center or emergency room immediately. US residents can call the US national poison hotline at 1-800-222-1222. Canada residents can call a provincial poison control center.

4.4 RESULT EVALUATION

The Bilingual Evaluation Understudy (BLEU) [3] is the metric used for the assessment of the model generated answer with the already existing reference answers. A score of 1.0 is returned when the match is flawless, while an imperfect result is shown using 0.0. Given a candidate sentence *C* and a set of reference sentences *R*, the BLEU score is calculated as follows:

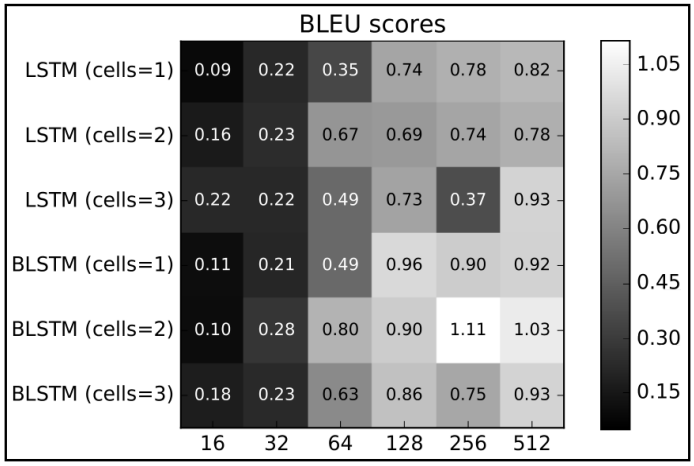

Fig.4. BLEU scores for models with number of cells and state sizes

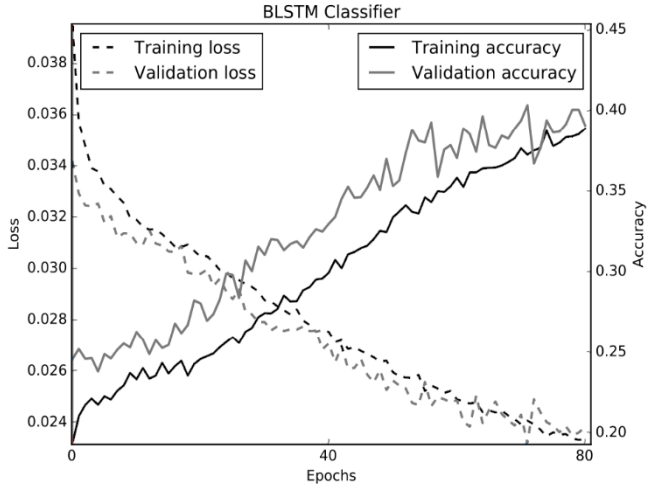

Fig.5. Training and validation loss and accuracy obtained using the BLSTM model

For each n-gram in *C*, the maximum occurrence is counted from the corresponding *R*. These two counts are summed up and a modified precision is given as, the total number of maximum relevant n-gram occurrences in *R* divided by the total number of n-gram occurrences in the candidate. The LSTM and BLSTM model were trained with different number of cells (number of

recurrent layers), state sizes, character embedding sizes and without the use of dropout in the Q/A dataset. Finally, they were evaluated using the BLEU score. The resulting BLEU scores can be seen in Fig.4. It seems that the BLSTM model is the better model. Furthermore, higher BLEU score seems to be achieved when increasing the state size of the model. However, when not applying any dropout during training it seems to have a sweet spot of 1.11 at 256 units and when 2 cells are used.

From the Cross-Validation plot it can be seen in Fig.5, that after only 80 epochs, a validation accuracy of approximately 40% and rising should be possible for the multi-task network. Predicting the most frequent class every time yields an accuracy of approximately 25%, hence 40% and increasing is better than the baseline. Similarly, the loss for both training and validation data is observed to be decreasing which is always a positive sign for the model.

5. CONCLUSION

From this research study it can be concluded that training an end-to-end neural network using the Encoder-Decoder framework for generating answers to medical questions is a very difficult task, but possible. A final BLEU score of 1.11 was obtained using a bidirectional recurrent neural network and it has been shown that it is possible to learn certain fixed template answers to specific type of questions, and, also in some way learn responses to specific types of questions requiring more information from the user. Furthermore, a large fraction of the generated answers are correctly spelled and are understandable, hence, an evidence that the model have learned some understanding of syntax and semantics. Since, the set of valid answers to a given question can be very big and diverse compared to the source, the BLEU score might not be a valid performance metric for this task. One could easily come up with a completely valid answer to a question, which would yield a BLEU score of 0, hence, another evaluation criteria is needed. In [8], [22], the predicted answers are evaluated manually by a set of humans who determine whether the given answers are valid or not. Due to lack of available resources this was not a possible solution, hence the use of the BLEU score. Finally, it was found possible to construct a model classifying the question category and generating answers simultaneously, however, the obtained accuracy slightly decreased when the network had to predict sequences. This might have been due to lack of parameters in the model, investigating this has been left for future work.

REFERENCES

[1] X.Y. Peng, Y. Chen and Z.W. Huang, "A Chinese Question Answering System using Web Service on Restricted Domain", *Proceedings of International Conference on Artificial Intelligence and Computational Intelligence*, pp. 350-353, 2010.

[2] H. Zhang, L. Zhu, S. Xu and W. Li, "Xml-Based Document Retrieval in Chinese Diseases Question Answering System", *Mobile, Ubiquitous, and Intelligent Computing*, Vol. 274, pp. 211-217. 2014.

[3] Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation", *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311-318, 2002.

[4] Kazuma Hashimoto et al., "A Joint Many-Task Model: Growing a Neural Network for Multiple NLP Tasks", *Proceedings of International Conference on Empirical Methods in Natural Language Processing*, pp. 1-7, 2017.

[5] Trevor Hastie, Robert Tibshirani and Jerome Friedman, "*The Elements of Statistical Learning: Data Mining, Inference, and Prediction*", Springer, 2001.

[6] Liqiang Nie, Xiaochi Wei, Dongxiang Zhang, Xiang Wang, Zhipeng Gao and Yi Yang. "Data-Driven Answer Selection in Community QA Systems", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 29, No. 6, pp. 413-421, 2017.

[7] Zongcheng Ji, Zhengdong Lu and Hang Li, "An Information Retrieval Approach to Short Text Conversation", *Proceedings of International Conference on Computation and Language*, pp. 23-27, 2014.

[8] Lifeng Shang, Zhengdong Lu and Hang Li, "Neural Responding Machine for Short-Text Conversation", *Proceedings of 53rd Annual Meeting of the Association for Computational Linguistics*, pp. 1577-1586, 2015.

[9] D.A. Davis, N.V. Chawla, N. Blumm, N. Christakis and A.L. Barabasi, "Predicting Individual Disease Risk based on Medical History", *Proceedings of International Conference on Information and Knowledge Management*, pp. 773-778, 2008.

[10] S. Doan and H. Xu, "Recognizing Medication Related Entities in Hospital Discharge Summaries using Support Vector Machine", *Proceedings of International Conference on Computational Linguistics*, pp. 330-337, 2010.

[11] I. Batal, L. Sacchi, R. Bellazzi and M. Hauskrecht, "A Temporal Abstraction Framework for Classifying Clinical Temporal Data", *Proceedings of American Medical Informatics Association*, pp. 227-234, 2008.

[12] K. Latha and R. Rajaram, "Improvisation of Seeker Satisfaction in Yahoo! Community Question Answering Portal", *ICTACT Journal on Soft Computing*, Vol. 1, No. 3 pp. 152-162, 2011.

[13] R. Fakoor, F. Ladhak, A. Nazi and M. Huber, "Using Deep Learning to Enhance Cancer Diagnosis and Classification", *Proceedings of International Conference on Machine Learning*, pp. 291-297, 2013.

[14] M. Shouman, T. Turner and R. Stocker, "Using Decision Tree for Diagnosing Heart Disease Patients", *Proceedings of 9th Australasian Data Mining Conference*, pp. 336-343, 2011.

[15] Y. Zhang and B. Liu, "Semantic Text Classification of Disease Reporting", *Proceedings of the International ACM SIGIR Conference*, pp. 43-49, 2007.

[16] J. Zhou, L. Yuan, J. Liu and J. Ye, "A Multi-Task Learning Formulation for Predicting Disease Progression", *Proceedings of ACM International Conference on Knowledge Discovery and Data Mining*, pp. 881-885, 2011.

[17] B. Koopman, P. Bruza, L. Sitbon and M. Lawley, "Evaluating Medical Information Retrieval", *Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 123-128, 2011.

[18] P. Sondhi, J. Sun, H. Tong and C. Zhai, "Sympgraph: A Framework for Mining Clinical Notes through Symptom Relation Graphs", *Proceedings of International ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 21-24, 2012.

[19] S. Ghumbre, C. Patil and A. Ghatol, "Heart Disease Diagnosis using Support Vector Machine", *Proceedings of International Conference on Computer Science and Information Technology*, pp. 12-16, 2011.

[20] Sebastian Basterrech, Jan Janousek and Vaclav Snasel, "A Performance Study of Random Neural Network as Supervised Learning Tool using CUDA", *Journal of Internet Technology*, Vol. 17, No. 4, pp. 771-778, 2016.

[21] S.H. Yang, S.P. Crain and H. Zha, "Bridging the Language Gap: Topic Adaptation for Documents with Different Technicality", *Proceedings of 14th International Conference on Artificial Intelligence and Statistics*, pp. 91-95, 2011.

[22] Alan Ritter, Colin Cherry and William B. Dolan, "Data-Driven Response Generation in Social Media", *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pp. 141-146, 2011.

[23] F. Wang, N. Lee, J. Hu, J. Sun, S. Ebadollahi 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*, Vol. 35, No. 2, pp. 272-285, 2013.

[24] F. Wang, N. Lee, J. Hu, J. Sun and S. Ebadollahi, "Towards Heterogeneous Temporal Clinical Event Pattern Discovery: A Convolutional Approach", *Proceedings of International Conference on Knowledge Discovery and Data Mining*, pp. 178-183, 2012.

[25] A. Khosla, Y. Cao, C. C.Y. Lin, H.K. Chiu, J. Hu and H. Lee," An Integrated Machine Learning Approach to Stroke Prediction", *Proceedings of International Conference on Knowledge Discovery and Data Mining*, pp. 30-34, 2010.

[26] K. Karpagam and A. Saradha, "An Intelligent Conversation Agent for Health Care Domain", *ICTACT Journal on Soft Computing*, Vol. 4, No. 3, pp. 772-776, 2014.