Symptom-Based Disease Detection System In Bengali Using Convolution Neural Network

Enam Biswas
Department of Computer Science and Engineering
East West University, Dhaka, Bangladesh
Email: mrnorman.enam@gmail.com

Amit Kumar Das
Department of Computer Science and Engineering
East West University, Dhaka, Bangladesh
Email: amit.csedu@gmail.com

***Abstract*— Natural language processing (NLP) and automatic detection of the disease have become popular in the recent era. Several research work show disease detection system in several languages. We present a disease detection system from the clinical text which is in Bengali language consisting of a numerous set of diacritic character, at a sentence-level classification. The clinical dataset consisting of Bengali text which is generally user interpreted symptom for the most common disease. Also, our approach represents the NLP methodology for Bengali language processing and classification of disease using several types of neural networks with hyper-parameter tuning and word vectorization. The aim of the research is the initial detection of disease from the user's voice to text data, in our case Bengali. So, a speech recognition system developed in the Bengali language is used to feed the disease detection model and finalizing the output with the model-detected disease.**

Keywords—Neural network, Language model, Disease detection, Text classification.

I. INTRODUCTION

Medical diagnosis and detection of the disease have been first recorded in ancient Egypt (2630-2611 BC) [1]. From the very origin to modern ages numerous additions have brought a change in the disease detection system. Humans' thought of expression and sense of diagnosis of disease has been improved with generations. Whenever we visit a doctor in case of disease-related issue, initially we describe our problems (initial symptom by us, or patient) in our natural language. Our doctors or health professionals spend time detecting the critical issue (symptoms of disease) and conclude providing medication or solution for minimizing the disease. For a developing country like Bangladesh where the literacy rate is considerably low [2], people tend to visit nonprofessional medical specialists a lot. For example – "a patient having heart-related issues." So he should visit a Cardiologists. But over here they tend to visit an Allergists/Immunologists or other medical specialists. So, if we provide a system by which they could know what their disease is from their interpretation of symptom, both time and money can be saved.

Here comes our idea – using the machine learning core to build a disease detection model that will allow detection of disease from symptoms described in Bengali. Also, several disease detection types of research are available in other languages such as English, Spanish, etc. as these languages are highly developed considering numerous work on speech recognition system and natural language processing. Bengali is one of the most popular languages (ranked 7th worldwide), as one-sixth of words' population speak Bengali [3][4]. So the primary focus is to work on the Bengali language to develop the human-computer interaction.

This research shows the core and basics of applying NLP on Bengali by representing the disease classification from plain text applying the concept of the neural network which is used widely for classification problems and NLP. The neural network is achieving excellent results, such as sentence modeling and other NLP tasks. The first model is an artificial neural network (ANN) model; which in our case a basic neural network to classify the disease from symptom data, word embedding with bag-of-words (BOW), with colossal size matrix on input layer results in a comparatively slower model. The other model is based on a convolutional neural network (CNN) maintaining the basics of sentence classification as implemented on English [5]. The CNN model consists of one convolution layer on top of word embedding (Word2Vec: continuous bag-of-words (CBOW), skip-gram model and fastText) trained on Bengali disease symptom dataset. The validity of the dataset is ensured by crowdsourcing on our Bengali data.

Also, the accuracy of detection of disease is shown in two cases – one, from the user plain data and the other one is voice data corresponding to text data. Test data from user input as text format has overall a better accuracy than input for test data from a proposed speech recognition in Bengali by Jahirul, Masiath & Rakibul [6]. The contribution of the research work follows –

- The outcome of the research work is made open source so that doctors and health professionals can contribute to the system by enriching it with information and core purpose of helping people can be served.

- It can be implemented in other systems.

- A revised concept for Bengali stemming and also sentence level classification for Bengali.

The paper consisting of sections as follow – Section II demonstrates some related work on NLP in Bengali language and some disease detection system. Section III containing information about the data used for the system. Description of the data processing and Bengali sentence architecture is also described in Section III, to gain a proper understanding of the language. Later, ANN and CNN both models are represented in Section IV. Experimental set up for word embedding was described in section V. Section VI contains the procedure for testing and evaluating the proposed model. Finally, Section VII includes the research summary.

II. RELATED WORK

There is a numerous amount of research on sentence level classification of tasks in several languages. But in case of conducting classification on health-related data is contemplated as a particular case. Because of neumerous diacritic character, the structure of Bengali being more complicated than another well-developed language like English, each specific task behind the classification method is clarified. Because in the case of research on NLP, the Bengali language is still under development. Also, for classification on

978-1-7281-1557-3/19/$31.00 ©2019 IEEE

record events of patients and long stablishing diseases (example – diabetes) Support Vector Machines (SVM) and Latent Dirichlet Allocation (LDA) serves a great purpose [7][8].

For text classification research, making a group of similar words representing semantically by the projection of words in vector space [9]. To learn Word2vec representation, Mikolov et al. introduce a great approach [10], which was also used on medical text research in the English language [11]. Well known classification features like BOW, n-grams and their TF-IDF has been used with ConvNets by Zhang & LeCun in English [12]. To gain a fixed size representation of Bengali sentence, some of the embeddings is implemented. Our work is the first such kind of research which is performed on Bengali and also in general human interpreted symptom of the disease.

III. Dataset & Data Processing

The dataset that has been used to classify disease from general symptom was scrapped from medical websites and data stores. Our main aim of the research is to allow user input their symptom in natural language that they express while describing to a doctor about their problems. Such example is shown in Table I. We do not expect the user to know the medical term of their disease symptom and most of the cases the medical condition don't have any meaning in Bengali as we tend to memorize them in English and thus globally.

TABLE I. Symptom Demonstration

	A symptom of a max Acute Sinusitis patient	
	English	***Bengali***
Patient to Doctor	There is a massive amount of pain behind my eyes. The pain also seems to be on my forehead. Also, I am suffering from fever, cold for several days. The mucus from the nose is yellow and sometimes deep red. There is a drainage of mucus from throat.	চোখের [chokher] পেছনে [pechone] প্রচণ্ড [prochondo] ব্যথা [betha]। কপালেও [kopaleo] ব্যথা [betha] মনে [mone] হয় [hoy]। আর [ar] হচ্ছে [hocche] আমি [ami] অনেক [onek] দিন [din] ধরে [dhore] জ্বর [jor] আর [ar] ঠাণ্ডাতে [thandate] ভুগছি [vugchi]। নাকের [naker] সর্দি [shordi] হলুদ [holud] এবং [ebong] কোন [kono] সময় [somoy] গাঢ় [garo] লাল [lal]। গলা [gola] দিয়েও [diyeo] কফ [kof] বের [ber] হয় [hoy]।

A. Dataset

The kind of information we looked for are sentences with symptoms in the form of conversational language. For such kind of text, websites like MedicineNet[1], Mayo Clinic[2], WebMD[3], CDC[4], and Healthline[5] consists of various data. First, to collect data, we have chosen a set of most common diseases and their symptoms. Few data example of symptom for one disease is demonstrated in Table II. In our dataset overall, 59 diseases are included consisting of symptoms over 1500.

[1]MedicineNet is a medical website that provides detailed information about diseases, conditions, medications and general health. *https://www.medicinenet.com/*

[2]Mayo Clinic is a nonprofit academic medical center. *https://www.mayoclinic.org/*

[3]WebMD provides valuable health information, tools for managing your health, and support to those who seek information. *https://www.webmd.com/*

[4]CDC is one of the major operating components of the Department of Health and Human Services. *https://www.cdc.gov/*

[5]Healthline Media, Inc. is a privately owned provider of health information headquartered in San Francisco. *https://www.healthline.com/*

[6]Dr. Md. Shafiqur Rahman, Medical Officer in at Bangladesh University of Engineering & Technology (BUET), Dhaka.

[7]Dr. Altaf Hossain, M.B.B.S. & a Medicine doctor currently living in Dhaka, Bangladesh.

B. Test Data

To validate our model and gain confirmation of detection of disease, we didn't use the traditional way of splitting a portion of training data. We have collected the test data in voice and text form from patients of Dr. Md. Shafiqur Rahman[6]. Before the collection of data, patients were notified about the data collection and the reason behind it. Also for the data collection, personal information like – name, age, etc. is eliminated. For using the data, proper participants consent have been taken. Collected data from patients' demonstration of their symptoms of the disease to the doctor is in both text and audio clip format.

C. Data Processing

Initially, the data was collected in the English language. Later it was converted to Bengali using Google Translate API. However, the output for each entry from the API was not correct in most of the cases. Things noticed – Google translate API tend to perform well by translating one word at a time. But the accuracy of the API went down when we give a whole sentence, in case of Bengali. Most of the time translated the sentence from Bengali doesn't make any sense, even if the individual word translation of the sentence. So, later all the translated data was revised and processed manually. A demonstration of such example is provided in Table III.

TABLE II. Example of an Incorrectly Translated Symptom

English text	***Translated text (wrong)***	***Revised text***
Tiny red dots on your skin from broken blood vessels.	ভাঙা [vanga] রক্তবাহী [roktobahi] জাহাজ [jahaj] থেকে [theke] আপনার [apnar] ত্বকের [toker] ক্ষুদ্র [khudro] লাল [laal] বিন্দু [bindu]।	ছেড়া [chera] রক্তনালী [roktonali] থেকে [theke] আপনার [apnar] চামড়ায় [chambray] ছোট [choto] লাল [laal] দাগ [dag]।

This kind of translation almost changed the context of the sentence. Also, the spelling of words and several meaning and synonyms have been added to the dataset during revision.

D. Stemming

The processing of data before going into training has been looked upon carefully. As well as the test data. So, to allow our machine to learn something, we have gone through data cleaning to teach the core words to our machine. Else in case of word prefix and suffix we might gain worse performance. So here comes a concept called stemming. Several stemming algorithms are developed for numerous languages. But hardly any Bengali stemmer is found which is open sourced or already developed.

Word in the Bengali language generally shows two types of inflection – verbal and nominal inflection. In some cases, pronominal and adjective inflection is observed, but hardy occur in describing thoughts [13].

a) Verb and Noun Inflection: In Bengali language verb inflection only occurs as a suffix. As, Verb formation (*verb-roots + verb-suffix;* Table IV). To detect the root of the verb, we have to understand the types of verbs. Bengali verbs are of two types – Finite and Non-finite. In the case of finite verbs, the occurrence of inflection are seen because of change in tense, a variation of persons and relation or honor (intimate, familiar, formal) [13]. Verb roots can be categorized into three groups: roots with only one complex

Bengali character, roots with two complex Bengali character and roots with three complex Bengali characters *(consonants + vowel marks;* Table III).

TABLE III. TYPES OF VOWEL MARK IN BENGALI

Types of vowel marks							
Vowel Marks	অ [ô]	ই [i]	উ [u]	ঋ [r/ri]	আ [a]	ঈ [ī/ee]	ঊ [ū/oo]
		ি	ু	ৃ	া	ী	ূ
Complex Vowel Marks	এ [e]		ও [o]		ঐ [oi]		ঔ [ou]
	ে		ো		ৈ		ৌ

TABLE IV. CATEGORY OF VERB ROOTS

Verb Roots	
Category 1	হ [ha], খা [kha], দি [di], শু [shu], etc.
Category 2	কর [kor], কহ [koh], উঠ [uth], ফিরা [fira], দৌড়া [doura], etc.
Category 3	চটকা [cotka], বিগড়া [bigra], ছোবলা [chobla], etc.

This shows an example of building up of a category two verb-root. In Bengali, noun inflections are seen due to nominative, objective, genitive and locative causes which may vary from singular to plural [13]. Noun inflection (Table V) only happens at the end of the core noun.

TABLE V. LIST OF NOUN SUFFIX

Noun	Suffix
Singular	রা [ra], টা [ta], টি [ti], খানা [khana], খানি [khani], etc.
Plural	এরা [era], গুলি [guli], গুলো [gulo], দের [der], etc.

b) Stemming procedure for Verb & Noun Inflection: In case of verbal stemming, steps of stemming must be followed serially. Two types of inflections – independent inflection (containing minimum of length one e.g., ই [i], ছ [ch], ক [k/ka], etc. or two inflections suffix, e.g., লা [la], লো [lo], টা [ta], etc.) and combined inflection (combination of two or more independent inflections) are seen generally. From comparative study and few fixes with rules of [13] and related article on word formation in Bengali, we have proposed a basic algorithm (providing some fixes of [13]). Figure 1 contains some demonstration of steps of stemming by the proposed set of rules. In case of verb stemming set of rules should be followed strictly.

Observation of noun inflection shows that it generally happens independently [13]. As the inflection of the noun occurs at a limited number, the set of rules are also simple, and infections are easy to eliminate. Figure 2 contains some examples and flow of a process for Bengali noun stemming. But in the case of noun stemming it is not mandatory to follow all the rules one after another. It serves the purpose of our symptom-based disease dataset. Also during the data cleaning procedure, unnecessary characters were eliminated.

E. Data Validation

To ensure data quality (correctness and usefulness), the concept of crowdsourcing is used for dataset validation. Participation of the group of people for this task was voluntary, and participants are well notified about the purpose of crowdsourcing. Group of people participated – Doctor (2), Graduate student (3), Undergraduate student (6) and student of college (3). The data with symptoms in English and Bengali translation was given to all the participants, and they were asked to give it a rating. For data correctness and quality the average rating from the group of doctors is 9.25 out of 10. For checking the quality of the translation, data were given between the groups of students. The average translation rating is 8.86 out of 10. Also, participants were asked to fix the problem that they have observed during the validation process.

IV. MODEL DESCRIPTION

To conduct a comparative study with processed core data (stemmed data), we have experimented on the dataset for classification with and without word vectorization in the form of naïve approach to study proved well know classification approach that performed quite well in languages like English. Here, we will be discussing our ANN and CNN based model from experimental setup along with the optimization.

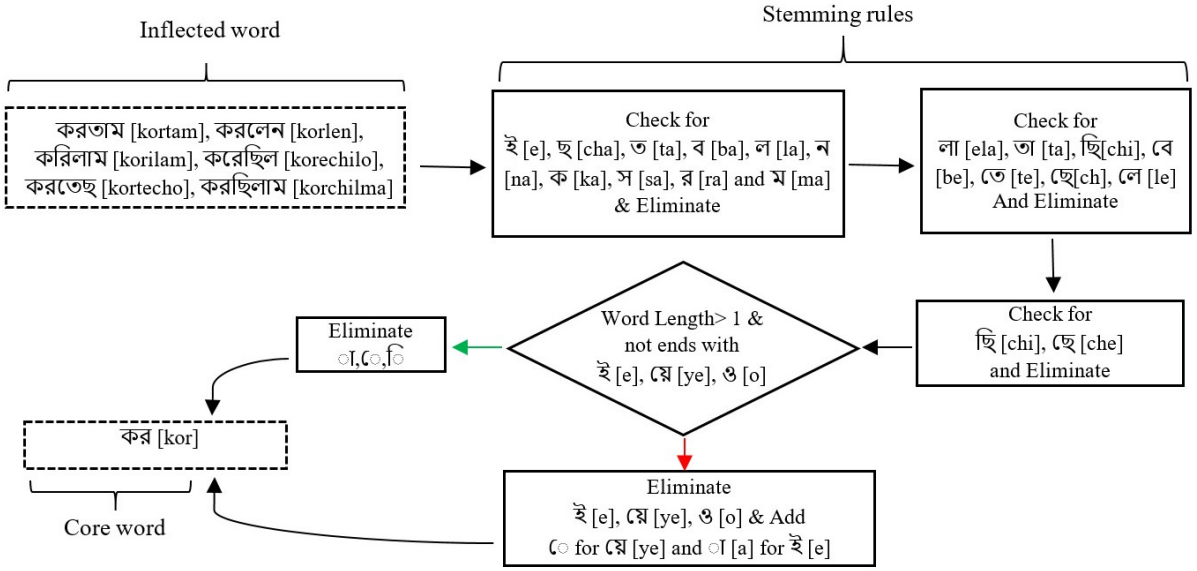

Fig. 1. Rules of Verb Inflection Elimination Rules.

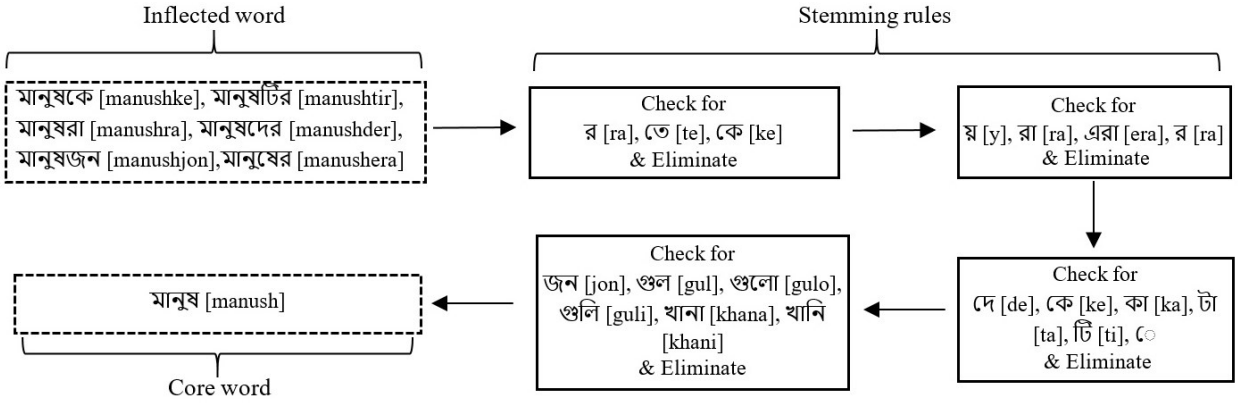

Fig. 2. Rules of Noun Inflection Elimination.

A. The architecture of the ANN System

The ANN-based model was a simple model, based on the neural network. For this system, we have experimented classification of Bengali in several approaches: the basic model and word vectored model (BOW). Structure of the system –

1) Processing: From the unique word in the corpus, each training sentence is reduced to binary array. This concept is similar to the basics of bag-of-word. Some problems occurred, e.g., the binary array gets significantly bigger with greater amount of data as well as the computation gets slower. Word vectorization seemed to reduce this problem.

2) Model: Our core function of ANN is a 3-layer neural network. Where, first one is the input layer, and the last one is the output layer, having one hidden layer in between. To normalize values and its derivative to measure the error rate, we have experimented with several activation functions (Sigmoid, Tanh & ReLU). *Tanh* seemed to perform faster and accurate in our case.

3) Parameter and Training: After having some experiment with the learning rate we have set our learning rate as 0.001 and a dropout of 0.5, means randomly selected node will be dropping out 50% weight at each weight update cycle. We have conducted 50000 epochs of training.

B. The architecture of CNN System

This system uses a simple convolution neural network with single-channel architecture, based on the study and experiment [5]. Word vector that was already trained by us on the dataset is used over the one layer of convolution. The word vectors are kept static during the process.

1) Formulation of CNN: For *i-th* word in a sentence let $v_i \in R^k$ be the *k*-dimensional word vector. For representing a sentence with length *l* –

$$v_{i:n} = v_i \oplus v_{i+1} \oplus \dots \oplus v_n \quad (1)$$

So, $v_{i:n}$ is referring as a concatenation ($\oplus$) of words v_i, v_{i+1}, … , v_n. A filter $x \in R^k$ is applied to a window of h words in order to introduce a new feature to operation convolution. From window of words $v_{i:\, i+h-1}$ generation of new window follows as –

$$m_i = f\,(w \cdot v_{i:\, i+h-1} + b) \quad (2)$$

Where f represents a non-linear function and b as bias term, $b \in R$. A feature map is produced by applying this filter to each possible word window. If we consider M (where $M \in R_{n\text{-}h+1}$) as a feature map, then we can represent it as –

$$M = [\, m_1, m_2, \dots, m_{n-h+1} \,] \quad (3)$$

Each feature was taken from the maximum value corresponding to the particular filter M by applying max-over-time pooling operation, Collobert et al., 2011. This process will extract one feature from one filter. Also, we are allowing the model to get multiple features from changing window size through multiple filters. Lastly, the probability for classification comes from a fully connected softmax layer from the calculation of features.

2) Parameters and Training: For this model we use: Convolution filters of 3, 4 and 5 [5]. They are applying a dropout rate of 0.25 for a batch size 50. First, the trained model was saved, and the accurate measurement is conducted separately for each specific input; is provided in Section V.

V. Word Embedding System and Findings

For dimensionality reduction and contextual similarity, word embedding is used to represent words into a corresponding vector of real numbers. Word vector concept BOW is used with ANN model. Having some performance issue and drawback other techniques are used with CNN model. It is popular to use an unsupervised neural language model to initialize word vectors (*word2vec, fastText* etc.*)*.

We have implemented *word2vec* model with corresponding CBOW and skip-gram model for Bengali. Also, *fastText* has been implemented to gain a comparison. Our dataset contains 6680 words (stemmed) in total and 1441 unique stemmed words. For our small dataset, we have tried to keep the dimensionality of embedding vector low, 20 for both of these systems and we looked for 5 number of context word. The result we have got is somehow not relatable in some cases. Also for our small dataset, the cosine similarity between neighboring words seemed high. Few terms are mostly seen in the dataset, example - *"pain"*, *"skin"*, *"nose"* and *"fever"* etc. (Bengali form: *"*ব্যথা*" [betha]*, *"*চামড়া*" [chamra]*, *"*নাক*" [nak]* and *"*জ্বর*" [jor]*). The most similar word related to the key term *"pain"* depending on cosine similarity, are shown on Table VI.

TABLE VI. MOST SIMILAR WORDS

Model	Context Words (Top 4, left to right)			
	Bengali (stemmed)			
word2vec CBOW	কর [kor]	মধ্য [moddho]	য [ja]	সঙ্গ [shongo]
word2vec skip-gram	কর [kor]	মধ্য [moddho]	য [ja]	রক্তপাত [roktopat]
fastText	রক্তপাত [roktopat]	স্বাভাবিক [shavabik]	রক্ত [rokto]	অস্বাভাব [oshavab]

For term "*pain"*, the detected context words seem almost similar except য *[ja]*, detected with word2vec models. Overall fastText performed better with an average score of 0.988 (word2vec CBOW – 0.708 and word2vec skip-gram 0.748) in this particular case. Score of these test cases is demonstrated in Figure 3. Overall, word2vec with skip-gram scores better, used in the main CNN model.

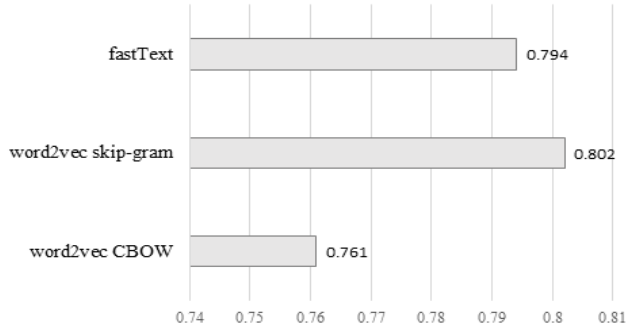

Fig. 3. Word embedding scores.

VI. ANALYSIS

We have used 20 test cases to analyze our systems. In case of analyzing with voice data, we have used a Bengali speech recognition system [6]. The Bengali speech recognition system is in alpha state and having a WER (word error rate) of 0.37. Also, it tends to perform better without any noise. For our test setup, both 20 cases of voice and text data are used. We have analyzed the ANN system with (ANN-stemmed) and without (ANN-non-stemmed) stemming of dataset to show the importance of word stemming in Bengali and how the performance degrades with word inflected data (demonstrated on Table VII). While testing our model the accuracy was examined by the statistical analysis of detection of a disease according to the doctors[6, 7], compared to our system and efficiency from the system.

TABLE VII. SYSTEM ACCURACY

System	Text Accuracy	Speech Accuracy	Disease Detection (out of 20)
ANN-non-stemmed	68.32%	57.45%	14
ANN-stemmed	74.56%	62.36%	16
CNN	81.88%	63.77%	17

Our observation – the CNN system worked better overall with an accuracy of 81.88%. By the implementation of word stemming, the ANN-stemmed system outperformed the non-stemmed system by 6.24%. This difference might not seem much because of the amount of data. Although the performance of voice interconnected system seem pretty low, as the text output from the system is incorrect in some cases. Amount of data used in our system, SMV (support vector machine) might perform better. But, we have tried to conclude with neural networks model considering the amount of data will be much higher. Also, the trained vector models were only applied to the CNN model, as theoretically it performs better than ANN [12].

VII. CONCLUSION

In this paper, we introduce an extensive study on Bengali text level classification. Also, it represents a novel approach for symptom-based disease classification. The performance of neural network systems on top of word embedding models even with the small amount of data proves the efficiency of the system. Also, the multilayer convolutional network creates more features while training, thus semantic representations are learned in order to text comparison. We could gain a much better result with greater amount of conversational data. Also the *word2vec* and *fastText* model could perform better so well the whole system. The basics can also be used to gather knowledge for the similar kind of task in other languages.

In future, our target is to implement our technique with larger scale and fine gained set of medical classification. Relating to computer vision literature, our convolutional networks will perform better with a larger dataset. We aim to check our hypothesis with more patient-doctor conversational data.

REFERENCES

[1] Medical diagnosis, "*https://en.wikipedia.org/wiki/Medical_diagnosis*" accessed on March 9, 2019.

[2] M. R. Ullah, M. A. R. Bhuiyan and A. K. Das, "IHEMHA: Interactive healthcare system design with emotion computing and medical history analysis," 2017 6th International Conference on Informatics, Electronics and Vision, 2017.

[3] R. A. Tuhin, B. K. Paul, F. Nawrine, M. Akter and A. K. Das, " An Automated System of Sentiment Analysis from Bangla Text using Supervised Learning Techniques," 2019 4th International Conference on Computer and Communication Systems (ICCCS), Singapore, 2019.

[4] A. K. Das, T. Adhikary, M. A. Razzaque, M. Alrubaian, M. M. Hassan, Z. Uddin, and B. Song, "Big media healthcare data processing in cloud: a collaborative resource management perspective," Cluster Computing, June 2017.

[5] Y. Kim, "Convolutional Neural Networks for Sentence Classification", 2014 Conference on EMNLP, October 2014.

[6] J. Islam, M. Mubassira, M. R. Islam and A. K. Das, "A Speech Recognition System for Bengali Language using Recurrent Neural Network," (ICCCS), Singapore, 2019.

[7] Marafino, Ben J., J. M. Davies, N. S. Bardach, M. L. Dean and R. A. Dudley. "N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit." JAMIA 2014.

[8] Wang, Lipo, F. Chu and W. Xie. "Accurate Cancer Classification Using Expressions of Very Few Genes." IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2007.

[9] G. S. Luis, G. H. J. Manuel, "CNN text classification model using Word2Vec", June 2017.

[10] Mikolov, Tomas, I. Sutskever, K. Chen, G. S. Corrado and J. Dean. "Distributed Representations of Words and Phrases and their Compositionality." NIPS 2013.

[11] M. HUGHES, I. LI, S. KOTOULASa and T. SUZUMURA, "Medical Text Classification using Convolutional Neural Networks", Stud Health Technol Inform, 2017.

[12] A. Conneau, H. Schwenk, L. Barrault, Y. Lecun, "Very Deep Convolutional Networks for Text Classification", EACL 2017.

[13] M. R. Mahmud, M. Afrin, M. A. Razzaque, E. Miller and J. Iwashige, "A rule based Bengali stemmer," 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), New Delhi, 2014.