International Conference on Innovation in Engineering and Technology (ICIET) 27-29 December, 2018

Disease Identification From Unstructured User Input

Fahim Faisal
Department of Computer Science and Engineering
Islamic University of Technology
Dhaka, Bangladesh
fahimfaisal@iut-dhaka.edu

Shafkat Ahmed Bhuiyan
Department of Computer Science and Engineering
Islamic University of Technology
Dhaka, Bangladesh
shafkatahmed@iut-dhaka.edu

Dr. Abu Raihan Mostofa Kamal
Professor
Department of Computer Science and Engineering
Islamic University of Technology
Dhaka, Bangladesh
raihanrcc@gmail.com

***Abstract*—The increasing number of Internet users leads to the rapid popularization of online searching for health related advice. Now a days, just in case of facing health problem, people tend to "go online" initially instead of consulting with a health professional. With the proliferation of online symptom checker sites and health forums, it is easy to gain knowledge regarding health condition supported by a number of given symptoms. Though existing symptom checkers provide instant sense of disease diagnosis, these question-answering and selection based systems lack in interactivity. Online health forum sites can also be underwhelming because of it's time intensive nature and reliability issues. In this scenario, this paper proposes an web based automated disease identification framework which takes unstructured textual data like health forum posts as input and provides a symptom-disease correlation based ranking of probable diseases as output considering all important factors. The proposed framework incorporates a lexicographic and semantic feature based two-phase state-of-the-art text classification system and a disease knowledge base based similarity measurement module to identify probable disease. We evaluate this framework varying the number of feature components and the result suggests that, significant accuracy and reliability is obtained over baseline systems by effective feature engineering at the same time of keeping up with increased user interactivity.**

***Index Terms*—Disease identification, Text Mining, Clinical Decision Support System.**

I. Introduction

The exponential growth of the Internet has triggered the age of information revolution. The rise of social networks, contemporary search engines and ubiquitous uninterrupted access through different kinds of devices have made information retrieval and sharing effortless than ever. Health care is a domain where access to the unprecedented magnitude of information is changing the way people interact. Nowadays many people use the Internet to self-diagnose their illness which includes both non-urgent symptoms and urgent symptoms like chest pain. In these cases, people use the Internet as their primary source of knowledge rather than consulting with an expert first. Though these high amounts of online resources are easily accessible, online searching can lead users to confusing and unreliable information, and people with imperative symptoms might not be directed to seek emergent care.

Using online health forums to search for medical advice and remedies is also a very common practice. Here people describe their problems and get guidance from health professionals and experienced people. Though access to these sites is free of health-related costs [1], these forum posts might have issues regarding trustworthiness, credibility and reliability [2]. As a result, fabricated posts or replies can drive the user in the wrong direction and gaining the wrong information might worsen the situation. Besides, following a specific forum thread might seem tiresome and time-consuming to a person who is in a critical emotional state. An intelligent online disease identification framework can be helpful users in these contexts which uses an intelligent decision support system to process inputs is like forum posts so that users can get the instant suggestion based on their queries without facing the above-mentioned problem.

Another popular medium of self-diagnosis is online symptom checker. There exists a large number of online symptom checker sites which provide rich disease databases and state-of-the-art decision support systems based on guided user input, long question and answer session and symptom-disease relation ([3],[4],[5],[6],[7],[8],[9]). These systems lack in adequate user interactivity as users cannot write or narrate in natural language as they can do in health forums or in time of consulting a doctor. So a feasible disease identification module or symptom checker should be capable of processing unguided patient's narrative in text format at least. The challenging task here is to extract relevant information from this emotionally biased, noisy and unstructured text where grammatical and spelling mistakes might occur frequently too.

Profound research work has been done on diagnosis using clinical texts and Electronic Health Record (EHR) data. Clinical text documents are domain specific with frequent use of clinical terminologies whereas, general users express their problems using non-medical terms [10]. It is unlikely that a user will use cardio(heart)-myo(muscle)-pathy(disease) related terms to express his heart-related problem. So a text input based disease identification system needs to map these non-medical texts with corresponding clinical terms. This type of systems can also be mentioned as a Clinical Decision Support System (CDSS). Most of the proposed and designed CDSS are doctor-centric. Less amount of research work has been done on patient-centric web-based CDSS which can help a patient to act without direct supervision from a professional [11].

Given these scenarios, this paper presents a noble disease identification framework that allows users to write their problem with minimum guidance. A two-phase text classifica-

978-1-5386-5229-9/18/$31.00 ©2018 IEEE

tion module is designed to perform entity extraction from user input followed by a disease-symptom correlation-based similarity measurement module which provides a ranking of probable diseases.

The remainder of this paper is as follows: in section II we briefly discuss the most relevant works regarding the above-mentioned scenario while different components of proposed disease identification framework are presented in section III. Experimental results are presented in section IV. In section V, we state the future scope of work in this domain and concludes the paper.

II. Related Work

Our work is related to two research areas: Information extraction from domain specific unstructured text and health forum based clinical decision support system.

Information retrieval using Natural Language Processing (NLP) is an active research field. Most of the research in this domain first performs rigorous cleaning, preprocessing and normalization tasks to convert noisy text into structured vector representation. Extensive research work has been done on various text representation techniques like Word2Vec: Continious Bag-of-Words model [12], Tf-Idf model. Available domain specific lexicons like: WordNet, UML are also used to improve feature performance. These representations are then used to perform entity recognition and relation extraction tasks. In [13] the authors perform "semantic relation extraction" using a two phase conditional random field (CRF) based classification to detect and characterize relation between disease-treatment and gene-disease entities. [14] identifies disease name by given symptom and extract relationship between disease and treatment by using a labeling algorithm for parsing sentences so that they can classify the information from database into cure, prevention and side-effect categories.

Various clinical decision support systems are proposed to help health-care professionals to prescribe efficiently. [15] is a symptom-medication relation based framework which uses Electronic Health Record (EHR) as the data source. [16] builds up a framework based on artificial neural network for the general practitioners to recommend them to prescribe the most trending drug to their patients. Besides, specific disease diagnosis systems like [17] mostly rely on EHR data. In [18] both structured EHR data and unstructured textual medical data are used to predict the state of Alzheimer's disease.

A large amount of research work has been published on disease identification using web[19] and social media data [10]. In [10] the authors propose a methodology to predict disease outbreak (flu trends) using a time-series twitter data mining. Trustworthiness and reliability of social media data are also big research issues. [2] uses a Markov Random Field-based model to judge a social network user's trustworthiness, the credibility of a medical statement and to predict the rare side-effect of the medicine discussed in a forum post.

Online symptom checker sites like webmd.com, mayoclinic.org take guided user input and mostly rely on symptom selection, Question & Answer session and word-database matching method. ([3], [8]) present Q&A session based on selected symptoms. Besides these features ([9], [6]) provide pictorial representation of human body for selection process. In [7], a long symptom list is given and based on the tabbed ones, user has to select important parameters like intensity, organ location, duration etc.[5], [4] use multiple symptom inputs and provide the exact database matching results. These existing solutions suffer from the following limitations.I None of these systems provide the scope of linguistic diversity. II Surfing over the long list and Q&A session is a time-consuming and tedious task for the user. III During selection process these systems use explicit medical terms which are hardly attainable by most of the users and thus, limits the scope of user interaction IV. Also, exact information matching from databases for multiple symptoms does not always take important demographic parameters into consideration. As a result, it is hardly optimal in terms of accuracy.

[20] is another study which targets to increase user interaction, accuracy and takes above limitations into consideration. Here, users can give a single symptom and related parameters as input in one line free text format. However, this framework does not provide the full scope of informal user input and misses out important parameters like past medical history and drug side effects.

III. Proposed Disease Identification Framework

In our suggested disease identification framework, an interactive interface provides space to give input in free English text form. There is a maximum word limit like other microblogging sites. Then the system performs preprocessing tasks to build the feature space for disease identification.

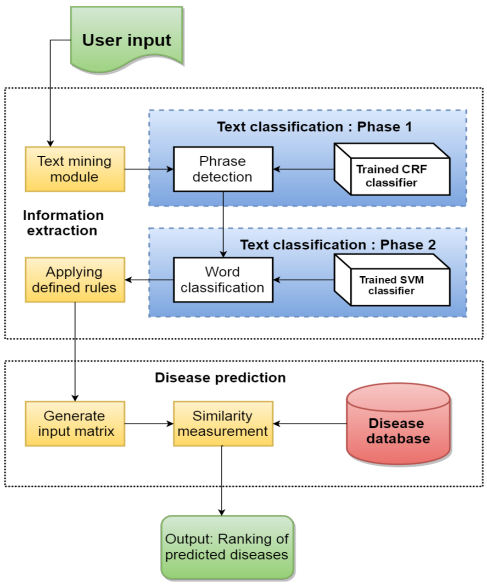

Fig. 1. System architecture

The system architecture is shown in fig. 1. There are two main modules reside in this architecture: (1) Information

extraction (2) Disease identification. Module 1 deals with text processing and entity extraction whereas disease identification part is done in module 2.

A. Information Extraction

At first, we analyzed posts from different sub-categories of a medical support forum: Dailystrength[21] to select necessary feature types. Based on our analysis we selected 6 types of properties: Symptom*(s)*, organ*(o)*, disease*(d)*, time*(t)*, medication*(m1)* and modifiers*(m2)* as primary class attributes.

We use python NLTK package for text preprocessing tasks. After spell correction, parts of speech tagging, tokenization five types of lexicon based and semantic based features are prepared:

a) Semantic bio-medical word tagging: We use Ontotag Semantic bio-medical tagger[1] as one of the main features. This is an online API service which provides bio-medical text tagging based on unified medical language[2] (UML) database. As for example, headache will be tagged as *Sign or Symptom* by this service.

Algorithm 1 Feature selection algorithm

```
1:  function classification_feature(input)
2:      pre_text ← text_preprocess(input);
3:      Let feature[1 ... number of words in pre_text ] be new array;
4:      for all sentence in pre_text do
5:          for all word in sentence do
6:              feature[word] ← feature_selection(word);
7:          end for
8:      end for
9:      return feature
10: end function

11: function feature_selection(word)
12:     Let feature[1 ... number of feature] be new array
13:     feature[word_feature] ← wordfeature(word);
14:     feature[uml] ← semantic_biomedical_tagger(word);
15:     dict_tagged ← dictionary_tagging(word) ;
16:     if dict_tagged ← null then
17:         synonym ← wordnet.semantic_similarity(word);
18:         dict_tagged ← dictionary_tagging(synonym);
19:         dictionary[word] ← dict_tagged;
20:     end if
21:     feature[dictionary] ← dict_tagged;
22:     return feature
23: end function

24: function dictionary_tagging(word)
25:     tagged_result ← dictionary.lookup(word);
26:     if dictionary.lookup(word)← null then
27:         tagged_result ← dictionary.lavenstain_similarity(word);
28:     end if
29:     return tagged_result
30: end function
```

b) Dynamic dictionary feature: We build a dictionary which contains non-medical terms with their corresponding class mapping *(s, o, d, t, m1, m2)*, technical term correlation, semantic similarity rating and synonym score. If any word with adequate feature score is not present in this dictionary, then our *feature selection algorithm* store this word with it's associated scores into the dictionary. This way, this dynamic dictionary automatically update itself (as shown in line no 19 of algorithm 1) and use this appended information in next searching.

c) Semantic word similarity: Our framework identifies pain, pains, painful as a word with same meaning. If word stemming and lemmatization part fails to find these type of variations, ***Levenshtein Distance*** [23] to measure the distance between two words.

d) Synonym mapping: Words unidentified by dictionary having relevant UML tags are identified by assigning synonym scores based on Word-net [24] synonym score.

e) Word feature: Additional word features include parts of speech tag, bi-gram, tri-gram, regular expression based replacers and word shape features.

1) Text classification module: Previously prepared feature vector is fed into a two phase text classification model:

a) Phase one: This phase is used to identify the word phrases by tagging each word using *Beginning-Inside-Outside (BIO)* format. We use conditional random field (CRF) [25] method to train a classification model based on dailystrength health forum dataset.

b) Phase two: This is the main part of our classification module. The state-of-the-art supervised classification methods are based on deep learning but in a limited computing environment, we need to think about the quality-efficiency trade-offs. Support vector machine [26] is a supervised learning method which provenly performs best from these quality-efficiency trade-off aspects. So we use SVM with linear kernel to tag all words according to the seven class labels: symptom, organ, disease, time, modifier, medication and others.

TABLE I
DEFINED RULES AND EXAMPLES

Defined rules	Examples
Modifier(color) + organ = symptom	Red skin
Modifier(direction) + organ = organ	Left shoulder
Modifier(direction) + symptom = symptom	Back pain
Lab test + modifier = test result	Blood test report is fine
Modifier(number) + 'years' + 'old' = age	15 years old (age = 15)
Modifier(negation + normal activity) = symptom	Unable to walk, painful breathing

2) Applying defined rules: After performing two phase word classification, we have all words tagged with relevant attribute names. Next predefined rules are applied to minimize the probability of missing important correlations in three steps:

a) Numerical mapping of modifiers: According to our system, intensity is measured in three range values: high(3), medium(2) and low(1). So high, sever and extreme these words will be mapped as high(3). In these way, age number and time values will be mapped in a age-group and time-range.

b) Matching predefined patterns: Tagged attributes are mapped following some predefined patterns as shown in table table I.

[1]https://console.s4.ontotext.com/

[2]https://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html

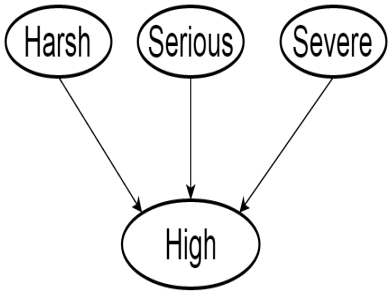

Fig. 2. Numerical mapping of modifier

c) Symptom mapping: Further symptom mapping is done in particular cases. As for example, symptom phrase *vision change* with negation word before that *(eg. severe pain but no change in vision)* will not be considered as a symptom.

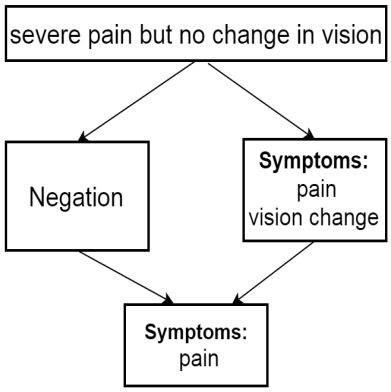

Fig. 3. Symptom mapping

A complete breakdown example of a sample input using our information extraction module is shown in fig. 4

B. Disease Identification

Numerically mapped entities are formatted in an input matrix. Then this matrix is matched against disease matrices from our disease database. This database is prepared based on recorded disease-symptom associations from expert sources [27],[7],[9] where each disease case integrates two components: Disease-document matrix and Disease-data matrix.

Disease-document matrix: It captures the behavior of unstructured input. Each disease object in our database has some predefined text cases. From these cases, our identification module prepares $tfidf$ [28] disease-document matrices. If in a disease database D, a disease object d has formatted text representation, then for each $d \in D$ our system calculates $tfidf(d)$.

$$tfidf(d) = \sum_{t \in q \cap d} tf_{t,d} \times log_{10} \frac{N}{df_t} \quad (1)$$

Free text format input:
Hello, I'm 24 years old. For some days I have fever with moderate shaking chills. I have also diarrhea and muscle pain.
Lower case conversion, replacing short form, spell correction:
hello, i am 24 years old. for some days i have fever with moderate shaking chills. i have also diarrhea and muscle pain.
Text preprocessing:
(be, 24, years, old) (some, days, have, fever, moderate, shake, chill)(have, diarrhea, muscle, pain)
BIO tagging:
(be-B, 24 -B, years -I, old-I) (some-B, days-I, have-B, fever-B, moderate-B, shake-I, chill-B)(have-B, diarrhea-B, muscle-B, pain-I)
Word tagging:
(be-B others, 24 –B modifier, years –I time , old-I modifier) (some-B modifier, days-I time, have-B others, fever-B symptom, moderate-B modifier, shake-I symptom, chill-B symptom)(have-B others, diarrhea-B disease, muscle-B organ, pain-I symptom)
Applying defined rules:
Sentence 1 - Age: 24 [rule: modifier (number) + time + modifier]
Sentence 2 - Symptom: fever, shake, chill Intensity: medium [synonym mapping]
Sentence 3 - Disease: diarrhea Organ: muscle Symptom: pain

Fig. 4. A sample input processing performed in information extraction module

$$here,$$
$$tf_{t,d} = number\ of\ occurrences\ of\ term\ t\ in\ disease\ case\ d$$
$$df_t = number\ of\ disease\ case\ containing\ term\ t$$
$$N = total\ number\ of\ disease\ cases$$

Disease-data matrix: Our disease identification module considers word ordering which is not performed in the bag-of-word representation of $tf–idf$ matrix. As for example, two text *"pain in muscle and weak feeling"*, *"weak muscle and pain feeling"* are same text in $tf–idf$ representation. So association between entities: *(pain, muscle),(weak,muscle)* is lost. To preserve this association, each disease object has a data-matrix component which is formed in a similar representation scheme like [20] where related entities are associated with each other. As for example, [29] and [30] are two diseases and their corresponding data-matrix representation $dm_{eye-redness}$ as $dm_{conjunctivitis}$ are as follows:

$$\begin{pmatrix} S[0] = redcolor & T[0] = * & I[0] = * & O[0] = eye \\ S[1] = headache & T[1] = * & I[1] = * & O[1] = * \\ S[2] = visionchange & T[2] = * & I[2] = * & O[2] = eye \\ S[3] = pain & T[3] = * & I[3] = high & O[3] = eye \\ S[4] = nausa & T[4] = * & I[4] = * & O[4] = * \\ S[5] = vomitimg & T[5] = * & I[5] = * & O[5] = * \\ Agegroupe = * & Medication = * & Relateddisease = * & Food = * \\ Gender = * & \times & \times & \times \end{pmatrix}$$

Fig. 5. Data-matrix $dm_{eye-redness}$

$$\begin{pmatrix} S[0]=grittyfeeling & T[0]=* & I[0]=* & O[0]=eye \\ S[1]=itchiness & T[1]=* & I[1]=* & O[1]=eye \\ S[2]=tear & T[2]=* & I[2]=high & O[2]=eye \\ S[3]=thickdischarge & T[3]=night & I[3]=* & O[3]=eye \\ Agegroupe=* & Medication=* & Relateddisease=* & Food=* \\ Gender=* & \times & \times & \times \end{pmatrix}$$

Fig. 6. Data-matrix $dm_{conjunctivitis}$

$$\begin{pmatrix} S[0]=tear & T[0]=<1week & I[0]=high & O[0]=eye \\ S[1]=itchiness & T[1]=* & I[1]=* & O[1]=* \\ S[2]=headache & T[2]=* & I[2]=* & O[2]=* \\ Agegroupe=2 & Medication=* & Relateddisease=* & Food=* \\ Gender=male & \times & \times & \times \end{pmatrix}$$

Fig. 7. Data-matrix representation of user query

1) Similarity measurement: From user query input u, Query-document matrix $tfidf(u)$ and query-data matrix d_{query} are created. Next, $cosine\ similarity$ [31] is calculated between these user input matrices and disease object components from database to identify probable diseases. Firstly, we can consider the case of similarity measurement between $tfidf(u)$ and $tfidf(d)$ for all $d \in D$. If corresponding vector representations of user input u and disease object d are $\vec{V}(u)$and $\vec{V}(d)$, then cosine similarity between these vectors can be calculated as follows:

$$Similarity_{cosine}(\vec{V}(u), \vec{V}(d)) = \frac{\vec{V}(u).\vec{V}(d)}{|\vec{V}(u)|.|\vec{V}(d)|} \ where\ d \in D \tag{2}$$

Next, similarity between data-matrices needs to be calculated. After extracting relevant attributes from user input u, we get the feature space f for query-data matrix dm_{query} generation. Here, $f = [s \cup t \cup i \cup o \cup \dots agegroup]$ where $s, t, i, o, \dots agegroup$ represents extracted symptom, numerically mapped and normalized values of time and intensity, organ and other attributes. This numerical mapping is done based on dictionary id, similarity score, synonym score and intensity rating of extracted attributes. Then min-max normalization scale these values to map in the query-data matrix dm_{query}. $Jaccard\ coefficient$ [31] is a similarity measurement method which is used to calculate similarity between dm_{query} and dm_d for all $d \in D$. $Jaccard\ coefficient$ between dm_{query} and dm_d is calculated as follows:

$$Coefficient_{jaccard}(dm_{query}, dm_d) = \frac{a}{a+b+c} \tag{3}$$

Where,
$d \in D$
$a =$ number of attributes $\in$ both objects.
$b =$ number of attributes $\in dm_{query}$ but not in dm_d.
$c =$ number of attributes $\in dm_d$ but not in dm_{query}.

For example, we can consider a user input – *"I'm 23 years old male. For three days I'm facing eye problem. A lot of tears with itchiness feeling. I have headache also though there is no change in vision."* The query-data matrix of this input dm_{query} can be visualized as follows:

Now if we calculate $jaccard\ coefficient$ of dm_{query} and disease-data matrices: $dm_{conjunctivitis}$ and $dm_{eye-redness}$, we get as follows:

TABLE II
JACCARD COEFFICIENT MEASUREMENT

Entity name	dm_{query}	$dm_{eye-redness}$	$dm_{conjunctivitis}$
RedColor		1	
Eye(red color)		1	
Headache	1	1	
Vision change		1	
Eye (vision change)		1	
Pain		1	
High (pain)		1	
Eye (pain)		1	
Nausa		1	
Vomiting		1	
Gritty feeling			1
Eye (gritty feeling)			1
Itchiness			1
Eye (itchiness)			1
Tear	1		1
Eye (tear)	1		1
High (tear)	1		1
Thick discharge			1
High (thick discharge)			1
0-10 days (tear)	1		
Age Groupe = 2	1		
Gender = male	1		

$$\begin{aligned} &Coefficient_{jaccard}(dm_{query}, dm_{eye-redness}) \\ &\qquad = \frac{1}{1+6+9} = 0.0625 \\ &Coefficient_{jaccard}(dm_{query}, dm_{conjunctivitis}) \\ &\qquad = \frac{3}{3+4+5} = 0.2307 \end{aligned} \tag{4}$$

From the calculated result, it is clear that $dm_{conjunctivitis}$ is more similar with dm_{query} than $dm_{eye-redness}$. So, according to data matrix similarity measurement, the user has more chance of having *conjunctivitis* than *eye-redness*.

Finally, identification module calculates a weighted sum of the calculated *jaccard coefficient* and *cosine similarity*.

$$\begin{aligned} Similarity(u,d) &= w_{cs} \times Coefficient_{jaccard}(u,d) \\ &+ (1-w_{cs}) \times Similarity_{cosine}(\vec{V}(u), \vec{V}(d)) \\ &where,\ w_{cs} < 0.5\ and\ d \in D \end{aligned} \tag{5}$$

Based on this similarity measurement, identification module suggests a list of most probable diseases articulating probability group for each output which can be high or low.

IV. Evaluation and Accuracy

A. Dataset & Baselines

We tagged 196 selected health forum posts from dailystrength [21] dataset containing 4915 words according to the BIO-format and class labels. Disease identification data for our experimental test cases were collected from online symptom checker site [5] .

In information extraction module, a naive baseline which classifies simply relaying on the class distribution of training data is used at first. Then,we perform our experiment using svm classifier with varying components. In identification module, we use disease-symptom associations recorded in [27], [9], [7] as ground truth values. In this setting, we split our data-set as 70-30 as training and testing data. Using the extracted information, we perform experiment to evaluate the performance of disease identification. These extracted information is given as manual input to online symptom checker [5].Then comparison is done with respect to ground truth values recorded from [27], [9], [7]. In our model, output is presented as a ranking of probable diseases in two clusters where one cluster presents diseases with high probability (H) and the other one is of low probability (L) disease. This is because, there can be many common symptoms result in different diseases. We perform total 10 experiments, each time based on different type of disease where significance of demographic information is not considered. Then, we perform same experiments considering demographic information.

B. Evaluation Metrics

As evaluation metrics, we use accuracy, precision, recall and f-score. In disease identification module, we compare obtained result with recorded ground truth. If for disease $D-n$, truth value and predicted value are same then it is 1, otherwise calculate it as 0. In this way, we can find out accuracy by computing the ratio of cumulative match factor and total number of disease.

$$\begin{aligned} Accuracy &= cm/N \\ Where,\ cm &= cumulative\ match\ factor \\ N &= total\ number\ of\ diseases \end{aligned} \tag{6}$$

C. Results and discussions

TABLE III
Results of phase one classification

Class	precision	recall	F1-score	support
B	.920	.979	.949	877
I	.778	.457	.575	138
O	.689	.69	.687	21

1) Setting I: Information Extraction: At first, we present the result of our implemented information extraction part. Classification report of phrase detection part is shown in table III. We obtained moderate precision and low recall in this phase.

TABLE IV
Accuracy comparison in setting I for varying components

Iteration	Naive baseline	SVM		
		Without dictionary feature	Without biomedical tagger	With all components
1	56.5	75.1	79.9	92.7
2	58.6	76.5	81.1	92.4
3	56.6	76.7	80.8	93.4
4	58.5	741	82.1	95.0
5	58.1	76.2	80.3	92.9

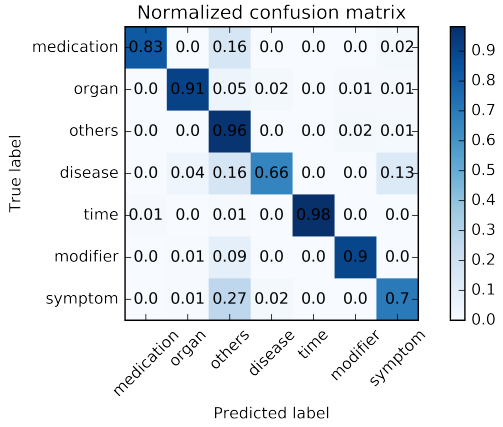

Fig. 8. Confusion Matrix for phase two classification

A detailed report of accuracy comparison of word classification part in setting I is presented in table IV. We perform 5-fold cross validation experiments and compare our result with baseline methods. At first, we use a simple naive baseline results in 57.66% average accuracy. Then we perform same experiments with varying components using SVM classifier with linear kernel. Including bio medical tagger feature results in 75.22% accuracy whereas, including dictionary feature shows 5.12% improvement. After incorporating all these features, we achieve an average accuracy improvement of 12.44%.

2) Setting II: Disease Identification: In table V, we presented the accuracy comparison of our identification module with online symptom checker. Here, we compared our result with the result obtained from existed web symptom checker site. In all 10 experiments, our system results in better identification. When we do not consider demographic information, it is 4.603% accuracy improvement whereas, considering demographic information results in slightly better performance with 5.783% accuracy improvement.

TABLE V
ACCURACY COMPARISON OF DISEASE IDENTIFICATION MODULE

Experiment no	Experiment type based on disease category [32],[33]	Symptom checker [5]	Our model	
			Without demographic information	With demographic information
1	Raspiratory Tract diseases: Nose and respiration disorder	64.34	71.54	71.54
2	Chronic diseases: Chrones's disease Alzehimer disease	59.27	61.63	64.32
3	Virus: Fatigue Sexually transmitted diseases	70.58	75.52	75.52
4	Nervous system: Restless legs Anxiety Sleep wake disorder	72.33	77.62	77.62
5	Bone diseases Osteoporosis Arthritis	70.32	74.56	78.56
6	Female Urogenital Diseases Pregnancy Complications	74.11	75.6	77.6
7	Male Urogenital Diseases	65.3	68.12	71.23
8	Cancer	61.89	63.67	63.67
9	Heart diseases	65.63	70.32	70.32
10	Occupational diseases: Asthma Pneumoconiosis	71.23	82.45	82.45

V. CONCLUSION AND FUTURE WORKS

The main contribution of this study is to propose a feasible text mining pipeline of web based disease identification module which aims to extend user interactivity. Besides, we investigated the significance of including different kinds of entities (eg. demographic information) along with generic symptoms. This work is done in a limited computing environment on a manually labeled dataset. So our future work includes performing this experiment in an ideal computing setting. For this purpose, we are preparing a publicly available dataset on which, the significance of this experiment will be examined further.

REFERENCES

[1] M. Ko and S.-H. Myaeng, "Identifying disease definitions with a correlation kernel for symptom extractions from text," in *Healthcare Informatics (ICHI), 2014 IEEE International Conference on*. IEEE, 2014, pp. 320–327.

[2] S. Mukherjee, G. Weikum, and C. Danescu-Niculescu-Mizil, "People on drugs: credibility of user statements in health communities," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 65–74.

[3] www.healthdirect.gov.au/symptom-checker.

[4] www.patient.info/symptom-checker.

[5] www.isabelhealthcare.com.

[6] www.everydayhealth.com/symptom-checker.

[7] www.mayoclinic.org/symptom-checker.

[8] www.nhs.uk/symptom-checker.

[9] www.symptomchecker.webmd.com.

[10] P. Velardi, G. Stilo, A. E. Tozzi, and F. Gesualdo, "Twitter mining for fine-grained syndromic surveillance," *Artificial intelligence in medicine*, vol. 61, no. 3, pp. 153–163, 2014.

[11] E. Ammenwerth, P. Nykänen, M. Rigby, and N. de Keizer, "Clinical decision support systems: Need for evidence, need for evaluation," *Artificial intelligence in medicine*, vol. 59, no. 1, pp. 1–3, 2013.

[12] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 3111–3119.

[13] M. Bundschus, M. Dejori, M. Stetter, V. Tresp, and H.-P. Kriegel, "Extraction of semantic biomedical relations from text using conditional random fields," *BMC Bioinformatics*, vol. 9, no. 1, p. 207, 2008. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-9-207

[14] M. Thangamani, P. Thangaraj *et al.*, "Automatic medical disease treatment system using datamining," in *Information Communication and Embedded Systems (ICICES), 2013 International Conference on*. IEEE, 2013, pp. 120–125.

[15] Y. Ling, Y. An, and X. Hu, "A matching framework for modeling symptom and medication relationships from clinical notes," in *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*. IEEE, 2014, pp. 515–520.

[16] Q. Zhang, G. Zhang, J. Lu, and D. Wu, "A framework of hybrid recommender system for personalized clinical prescription," in *Intelligent Systems and Knowledge Engineering (ISKE), 2015 10th International Conference on*. IEEE, 2015, pp. 189–195.

[17] Y. Wang, K. Ng, R. J. Byrd, J. Hu, S. Ebadollahi, Z. Daar, S. R. Steinhubl, W. F. Stewart *et al.*, "Early detection of heart failure with varying prediction windows by structured and unstructured data in electronic health records," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2015, pp. 2530–2533.

[18] J. Bullard, R. Murde, Q. Yu, C. O. Alm, and R. Proaño, "Inference from structured and unstructured electronic medical data for dementia detection," *Proceedings of Operations Research and Computing: Algorithms and Software for Analytics*, pp. 236–244, 2015.

[19] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, no. 7232, pp. 1012–1014, 2009.

[20] M. Tahmid, M. Tahmid, A. Raihan, and N. Rashid, "Automated disease prediction system (adps): A user input-based reliable architecture for disease prediction," *International Journal of Computer Applications*, vol. 133, no. 15, pp. 24–29, 2016.

[21] www.dailystrength.org.

[22] P. Shrestha, N. Rey-Villamizar, F. Sadeque, T. Pedersen, S. Bethard, and T. Solorio, "Age and gender prediction on health forum data," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, N. C. C. Chair), K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. Paris, France: European Language Resources Association (ELRA), may 2016.

[23] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," in *Soviet physics doklady*, vol. 10, 1966, p. 707.

[24] Wordnet.princeton.edu.

[25] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the eighteenth international conference on machine learning, ICML*, vol. 1, 2001, pp. 282–289.

[26] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[27] P. Ernst, C. Meng, A. Siu, and G. Weikum, "Knowlife: a knowledge graph for health and life sciences," in *2014 IEEE 30th International Conference on Data Engineering*. IEEE, 2014, pp. 1254–1257.

[28] J. Leskovec, A. Rajaraman, and J. D. Ullman, *Mining of massive datasets*. Cambridge University Press, 2014.

[29] Www.healthline.com/symptom/eye-redness.

[30] Www.healthline.com/symptom/conjunctivitis.

[31] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[32] Www.ncbi.nlm.nih.gov/mesh/1000067.

[33] Www.rightdiagnosis.com/a/all/subtypes.html.