IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10427

A Domain-Based Model for Knowledge Acquisition and Text Mining in A Database

Bennett, E.O.[1], Chidiebere, F.E.[2], SAKO D.J.S[3]

Department of Computer Science, Rivers State University, Port Harcourt, Nigeria[1,2,3]

Abstract: Extracting specific information from large volumes of text document possess a tremendous challenge. Medical records tend to have information glut problem (Masses of continuously increasing information) resulting in delay of early detection of diseases. This research concentrates on modeling a domain-based technique for knowledge acquisition and text mining in a database. The Term Frequency-Inverse Document Frequency (TF-IDF) was used to extract the entity's text features and the frequency of the terms is rescaled in TF-IDF by considering how often the words appear in all the documents. Random Forest Classifier (RFC) has been used to model the class of objects associated with symptoms' illnesses to facilitate accurate prediction of likely diseases the knowledge -base and text mining system calculates the possibilities of diseases according to given symptoms and displays the probabilities of disease accuracy.

Keywords: Domain Based, Knowledge Acquisition, Text Mining, TF-IDF,RCF.

1. INTRODUCTION

Knowledge Acquisition is used to describe the process of extracting, structuring, and organizing domain knowledge from domain experts into a system [1]. Due to the fast growth of data in real-life applications, there are huge amounts of data and information in different formats, and in various quality which leads to slow processing and extraction of useful information from large databases for decision making. Knowledge Acquisition techniques such as rough set algorithm, decision tree and fuzzy set algorithm where approaches used in faster discovery and extraction of useful knowledge from large databases. [2]

The Knowledge Acquisition techniques became very important tools for processing due to their ability to deal with huge amount of data with different formats and characteristics to prepare the data for fast processing [3].

Text mining refers to the discovery of non-trivial, previously unknown, and potentially useful knowledge from a collection of texts. Since its origin, text mining has been considered an analog of data mining (interpreted as Knowledge Discovery in Databases) applied to text repositories. Text mining is very important since nowadays because most of the information stored in computers (not considering audio, video, and images) consists of text. One common technique associated with text mining is an approach in which each document has a set of label to perform knowledge-discovery operations. This approach has been to assume that labels correspond to keywords, each of which represents that a given document is about the topic associated with that keyword. However, to be effective, this requires either: manual labeling of documents, which is infeasible for large collections; hand coded rules for recognizing when a label applies to a document, which is difficult for a human to specify accurately and must be repeated anew for every new keyword; or automated approaches that learn from labeled documents rules for labeling future documents, for which the state of the art can guarantee only limited accuracy and which also must be repeated anew for every new keyword [4].

The aim of this paper is to develop a domain-based model for acquiring knowledge alongside easy and fast extraction of vital information from a large medical dataset using Term Frequency-Inverse Document Frequency (TF-IDF) and Random forest classification.

This research is focused on exposing the following problems:

i. Determination of definite patterns and trends to examine a textual data is a challenge in text mining.

ii. Increase in volume, variety, velocity of data leads to delay in extraction of useful information from unstructured data.

iii. Organizational records tend to have information glut problem (Masses of continuously increasing information) which makes early extraction of text features in these records difficult.

iv. High tendency of inconsistency and inaccuracy as a result of large proportion of data.

Therefore, the importance of knowledge acquisition and text mining in this research are to enable much more efficient analysis of extant knowledge and also to unlock hidden information that leads to the development of new knowledge/idea.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10427

2. LITERATURE REVIEW

Genetic Algorithms (GAs) based approach for mining breast cancer pattern was proposed. [5] The approach extracted breast cancers' pattern, decision rules, threshold values and finally decision-making model with high degree of prediction accuracy. Their experimental results showed that the accuracy of prediction was improved.

A framework [6] was integrated into MEDLINE biomedical database and the unnecessary details are eliminated and valuable information is extracted using this new framework. [7] used text mining patterns for analyzing texts synonyms and polysemy are not analyzed properly by term based approaches. A prototype model was designed by them. This prototype model was used for arranging the patterns in various terms and conveying weights according to their distribution. Due to this approach the productivity of text mining was enhanced.
An approach based on text mining for knowledge acquisition in diagnostic systems. [8] applied naive Bayes technique with knowledge base of 38 disorders that can occur in a corn plantation. The limitation of the approach lies on mining 38 diseases in corn plantation and application of naive Bayes which has been used by several researchers.

3. SYSTEM DESIGN

The proposed system uses Term Frequency-Inverse Document Frequency (TF-IDF) to extract text features in the database and Random Forest (RF) to categorize text. It can learn continually which is very desirable in classifying text and estimate missing data. Figure 3.1 represents architectural design of text mining system and its components: text processing, word segmentation, text feature extraction using (TF-IDF), text classification using random forest classifier.

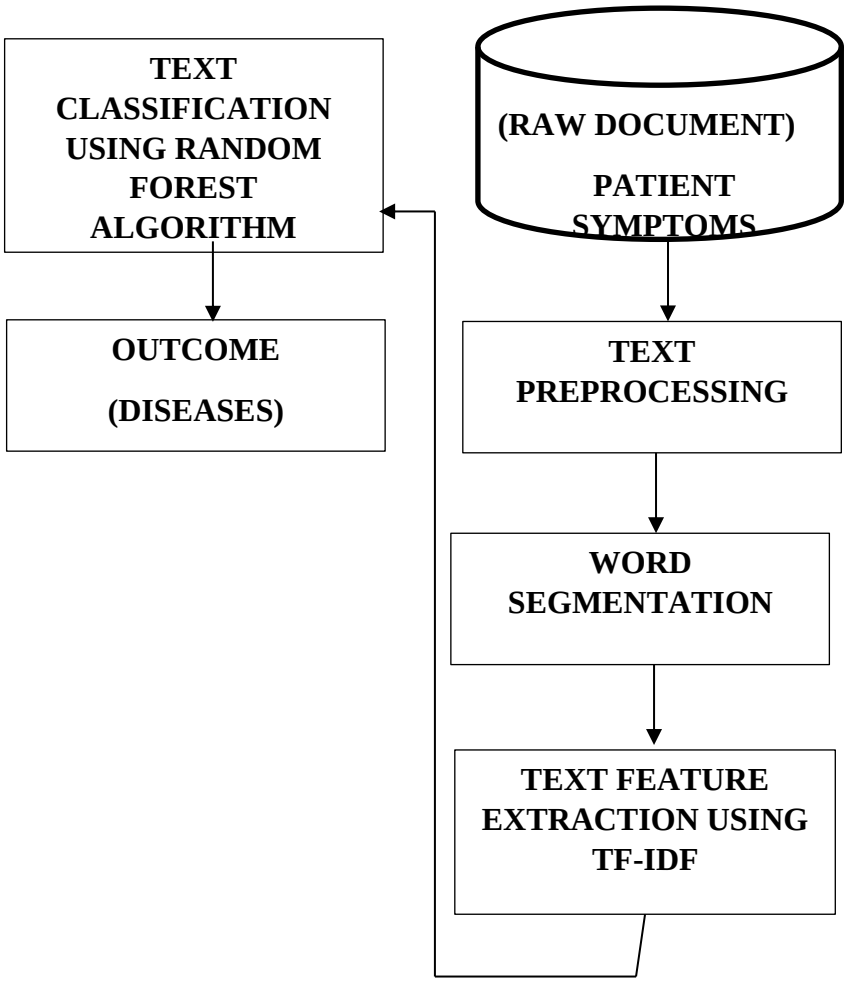


Figure 3.1. **System Architecture design**

The input to the system is derived from medical records of patient symptoms gotten from doctors and internet sources. Text preprocessing is a step of tokenization where longer strings of words are split into smaller pieces. Hence, Text Preprocessing reduces larger chunk of text in sentences then further tokenized into words.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10427

Word segmentation picks up the task of splitting text into contiguous coherent sections. The Text Feature Extraction is the process of taking out a list of words from the text data and then transforming them into a feature set which is usable by a classifier. This process is carried out by the Term Frequency-Inverse Document Frequency (TF-IDF).

Random text classifier takes the mined symptoms and carry out classification whereby Voting is done to classify symptoms which is performed by each tree i.e., the trees mark their votes for that class. The class having most number of votes decides the classification label which produces the outcome of possible diseases. Possible diseases with their percentages are expected to be the output from the system.

3.1 Input Specifications.

Inputs are resources needed for executing processes of a system. In this system, there are some basic input specifications which are as follows; The Login Validation: a sign in interface where the user login details are displayed followed by the User's Profile where a detailed profile of the said user is displayed next the Text Mining Module shows the search and message interface where a query is done by inputting a complete sentence or selection of symptoms then the Text Mining Process comes up which shows how text mining process is carried out through the search option. Alternatively, Disease Selection Process shows how text mining process is carried out through selection of symptoms. The system then displays the Disease probability process to show the possible results from the text mining process being carried out. Finally, the Symptoms Acquisitions message interface where all search history are being stored

3.2. Word segmentation

The segmentation task is to split a text into contiguous coherent sections. Firstly, build a representation of the text, by splitting it into *N* basic elements, V_i (i = 1, ... ,N), each a D-dimensional feature vector $v_{i\alpha}(\alpha = 1, \dots, D)$representing the element. Then we assign a score α (i; j) to each candidate segment, comprised of the i^{th}through α (j - 1)th elements, and finally determine how to split the text into the appropriate number of segments. Denote a segmentation of text into K segments as a list of K indices $S = (s_1, \dots, s_k)$ where the k-th segment includes the elements V_i with $s_{k-1} \leq i < s_k \; with \; s_0 \equiv 0$.

For example, the string "aaabbcccdd" considered at the character level would be properly split with S = (3, 5, 8, 10) into ('aaa', 'bb', 'ccc', 'dd').

To assess the utility of word vectors in segmentation, we consider more general scoring functions based on our word vector representation. As the representation of an element, we take

$$v_{i,k} = \sum_{w}^{k} f_{wk} \, v_{wk} \qquad (3.1)$$

$f_{i,w}$represent the frequency of word *w* in element ***i***.

$v_{w,k}$represent the k^{th} component of the word vector for word ***w***.

The length of word vectors varies strongly across the vocabulary and in general correlates with word frequency. In order to mitigate the effect of common words, the sum will be weight by the inverse document frequency of the word in the corpus:

$$v_{i,k} = \sum_{w}^{k} f_{iw} \, log \frac{|D|}{df_w} \, v_{wk} (3.2)$$

df_wrepresents the number of words ***w*** that appear in a document.

3.3. Text Feature Extraction

Text feature extraction is the process of taking out a list of words from the text data and then transforming them into a feature set which is usable by a classifier. This work emphasizes on feature extraction method known as Term Frequency-Inverse Document Frequency (TF-IDF).

3.3.1. Term Frequency-Inverse Document Frequency (TF-IDF)

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10427

In TF-IDF, the frequency of the words is rescaled by considering how frequently the words occur in all the documents. Due to this, the scores for frequent words are also frequent among all the documents are reduced. This way of scoring is known as Term Frequency – Inverse Document Frequency.[9]

Term Frequency (TF) is the frequency of the word in the current document. Inverse Document Frequency (IDF) is the score of the words among all the documents.

These scores can highlight the words that are unique that is the words that represent needful information in a specified document. Therefore, the IDF of an infrequent term is high, and the IDF of a frequent term is low.

Suppose we have a document(or a collection of documents i.e, corpus), and we want to summarize it using a few keywords only. In the end, we want some method to compute the importance of each word.Term Frequency-Inverse Document Frequency (TF-IDF) is given as:

$$tf - idf(t, D) = tf(t, d) \,.\, idf(t, D) (3.3)$$

$$tf(t, d) = f_{t|d} = \frac{number\ of\ time\ t\ appears\ in\ a\ document}{total\ number\ of\ terms\ in\ the\ document}$$

$$idf(t, D) = \log\left(\frac{N}{number\ of\ documents\ with\ t\ in\ it}\right)$$

Where,

N is the total number of documents

tf is TF (Term Frequency)

idf is IDF (Inverse Document Frequency).

To further illustrate the performances of TF-IDF for text extraction is the Table 3.1.below

Considering two documents D_1 and D_2

D_1 contains “Fortune is tall”.

D_2 contains “Fortune is not tall”.

Here,

N = |D|= 2

Table 3.1: TF-IDF for Text Extraction

Text	**Tf**		**Idf**	**tf-idf**	
	D1	**D2**		**D1**	**D2**
Fortune	1	1	Log(2/2)	0	0
Is	1	1	Log(2/2)	0	0
Tall	1	1	Log(2/2)	0	0
Not	0	1	Log(2/1)	0	Log(2)

The only thing that differentiates D1 and D2 is the word “not”. The TF-IDF reflects that “not” is really important in terms of IDF given as log(2) for the word “not”

Algorithm: Text Extraction

1. //text collection needed
2. load dataset D
3. Segment X into a sentence s1, s2,...,sn

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10427

4. if word wi appears in document Di
5. For i in range(iteration)
6. set the number of iteration as 100
7. y = compute equation (3.3)
8. end for
9. compute RandomForest(y)
10. end

3.4. Text Classification- Random Forest Classifier

Random Forest (RF) is a term used for an ensemble of decision trees. The Random Forest classifier is an ensemble learning method which involves collection of decision trees. Voting is done to classify a new object which is performed by each tree i.e., the trees mark their votes for that class. The class having most number of votes decides the classification label.

The decision tree classifier consists of a rooted tree, which contains nodes $t_0, \dots, t_n, n \in N$ that each represent a subspace$X_{t_0} \subseteq X$. The root node t_0 corresponds to the input space X. Each node t is labeled with a split S_t. The splits divide the nodes' subspace X_t into two subspaces, which are represented by the nodes' children. Random Decision Forest (RDF) for text classification is demonstrated in Figure 3.2 with an example that classifies eye problem based on symptoms.

Where;

$t_0, \dots, t_n$	represent a nodes
X_{t_0}	represent a subspace
X	represent an input space

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 10, Issue 4, April 2021
DOI 10.17148/IJARCCE.2021.10427

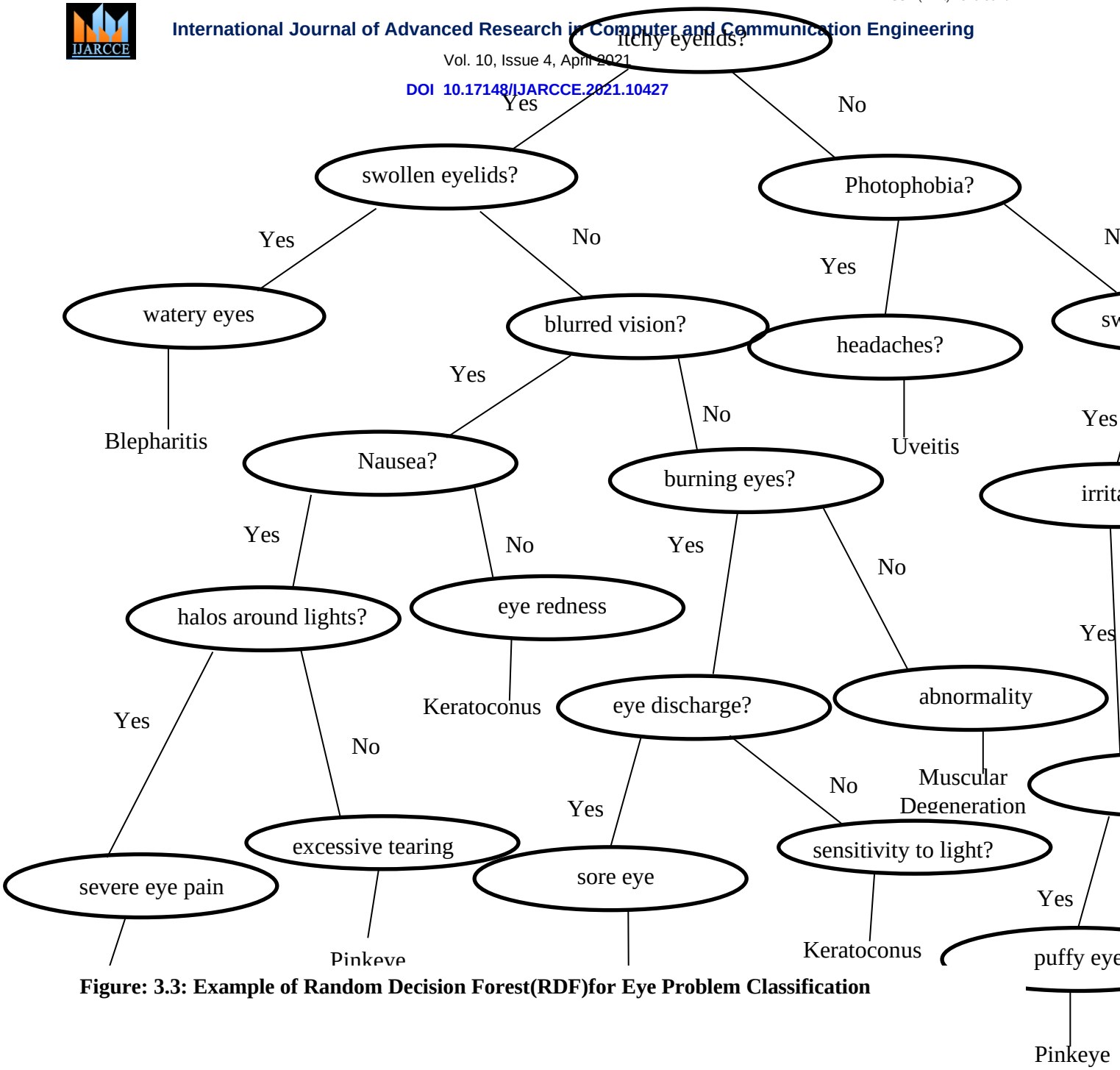


Figure: 3.3: Example of Random Decision Forest(RDF)for Eye Problem Classification

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10427

In the decision tree, each node makes a decision based on attributes of eye disease such as symptoms. Glaucoma and pinkeye might have one similar symptom like 'blurred vision' but will differ in other symptoms like 'eye discharge'. A node that makes a decision might send 'blurred vision' to the right and 'eye discharge' to the left. The output classes are eye diseases such as glaucoma, blepharitis, uveitis, keratoconus, cornel ulcer, pinkeye which are the terminals. However, the class “keratoconus” has the most number of votes; it appears three times in classification.

4. RESULTS AND DISCUSSION

Easy identification of disease in healthcare industry is achieved through knowledge acquisition and text mining. The knowledge base has been populated after a thorough search through the web and interviews with doctors for disease and symptoms. There exist 103 disease types and 228 symptoms in knowledge base. The system allows users to search for their symptoms by either selecting from 228 symptoms or typing their symptoms in a sentence form. According to the given symptoms, the knowledge base and text mining system calculates the possibilities for each disease, and displays the probabilities as depicted in Table 4.1.

Table 4.1: Mined Symptoms with Disease Possibilities

S|N	Text	Mined Diseases	Disease Possibilities
1	Bulging of the eye, Lump in eye, Pain in or around eye, Eye pain	Eye Cancer	30 %
		Retinoblastoma	20 %
		Nasal Cavity and Paranasal Sinus Cancer	20 %
		Retinoblastoma	20 %
		Histiocytosis, Langerhans Cell	10 %
2	Vomiting, Diarrhea, Abdominal pain	Appendix Cancer	6.82 %
		Wilms Tumor	4.55 %
		Islet Cell Tumors- Pancreatic Neuroendocrine Tumors	4.55 %
		Cholera	4.55 %
		Gallbladder Cancer	4.55 %
		Hepatocellular (Liver) Cancer	4.55 %
		Histiocytosis, Langerhans Cell	4.55 %
		Gastrointestinal Stromal Tumors (GIST)	4.55 %
		Bile Duct Cancer	4.55 %
		Pancreatic Cancer and Pancreatic Neuroendocrine Tumors (Islet Cell Tumors)	4.55 %
		Parathyroid Cancer	4.55 %
		Rectal Cancer	4.55 %
		Gliomas	2.27 %
		WaldenstromMacroglobulinemia	2.27 %
		Liver Cancer (Primary)	2.27 %

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10427

		Paraganglioma	2.27 %
		Neuroblastoma	2.27 %
		Anthrax	2.27 %
		Gestational Trophoblastic Disease	2.27 %
3	He has pain, joint stiffness and joint swelling	Bone Cancer	33.33 %
		Anthrax	11.11 %
		Appendicitis	11.11 %
		Chordoma	11.11 %
		Liver Cancer (Primary)	11.11 %
		Mouth Cancer	11.11 %
		Sarcoma	11.11 %

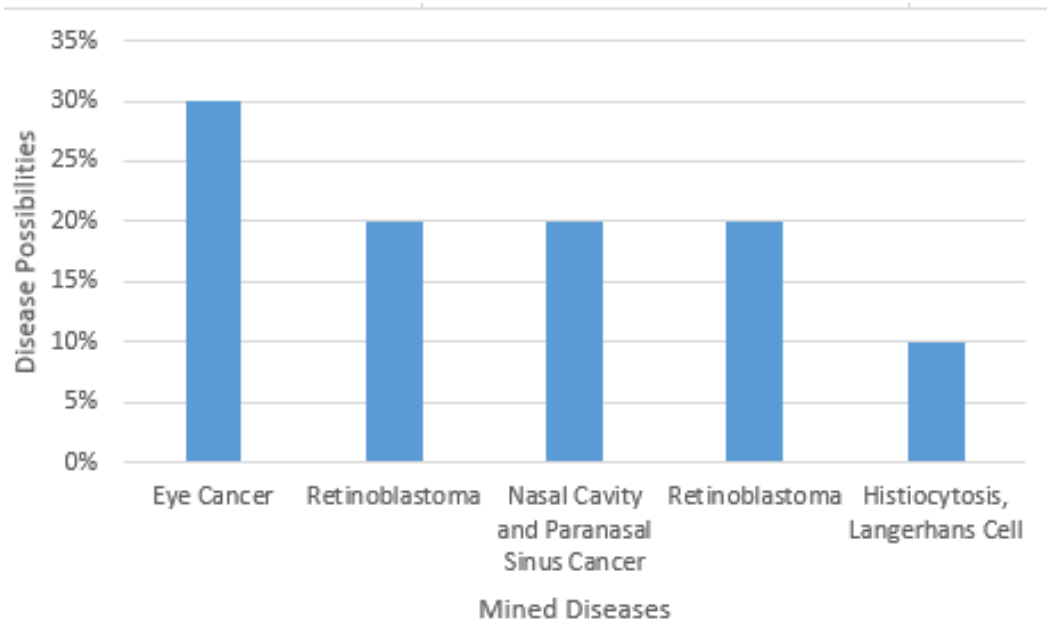


Figure 4.1: Symptoms (Bulging of the eye, Lump in eye, Pain in or around eye, Eye pain) with Disease Possibilities

IJARCCE

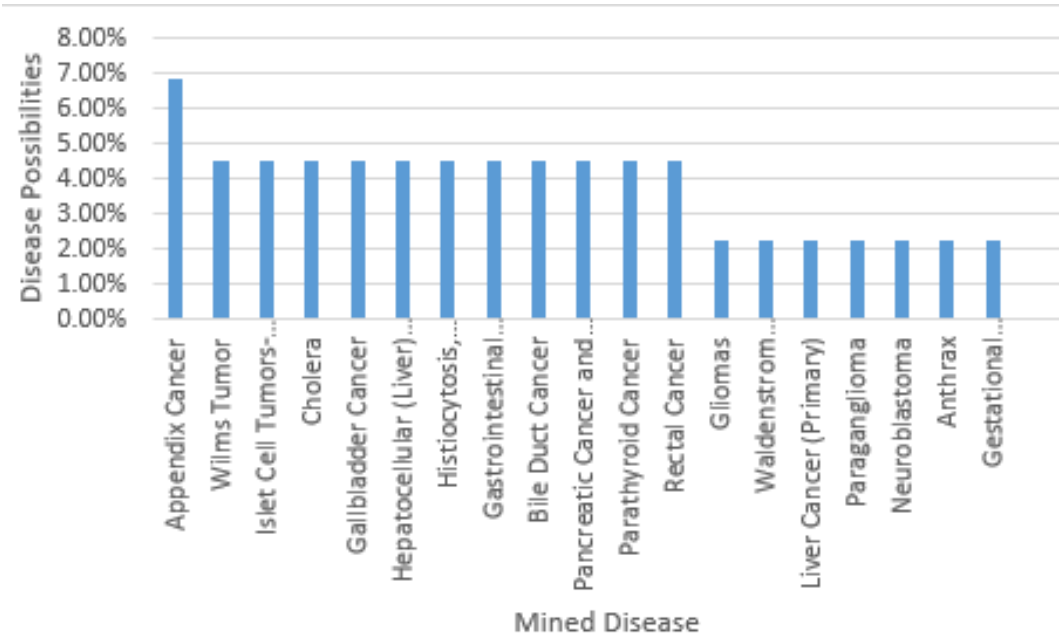
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10427

Figure 4.2:



Symptoms (Vomiting, Diarrhoea, Abdominal pain) with Disease Possibilities

This system is a database driven identification system regarding different disease types such as (Eye Cancer, Liver Disorder, Anthrax, Flu, Diphtheria, Cholera, Tetanus, Mumps, Rabies, Appendicitis, Acute Lymphoblastic Leukemia(ALL), Acute Myeloid Leukemia (AML), Adrenocortical Carcinoma, Adrenal Cortex Cancer, Kaposi Sarcoma, Lymphoma, Primary Cerebral (CNS) Lymphoma, Anal Cancer, Appendix Cancer, Astrocytomas, Childhood Atypical Teratoid, Basal Cell Carcinoma (Non-melanoma), Bile Duct Cancer, Bladder Cancer, Bone Cancer, Brain Tumors, Breast Cancer, Bronchial Tumors, Burkitt Lymphoma (Non-Hodgkin), Carcinoid Tumor (Gastrointestinal), Cardiac (Heart) Tumors, Cervical Cancer, Chordoma, Colorectal Cancer, Cutaneous T-Cell Lymphoma, Ductal Carcinoma In Situ (DCIS), Endometrial Cancer, Ependymoma(Childhood), Esophageal Cancer, Esthesioneuroblastoma, Ewing Sarcoma, Extragonadal Germ Cell Tumor, Intraocular Melanoma, Retinoblastoma, Fallopian Tube Cancer and many more. Most patients with disease symptoms don't investigate their symptoms due to their symptoms being the same as an everyday sickness.

There are 103 types of diseases in the knowledge base. Diseases can be serious because the first signs are not diagnosed. There are cancers that do not cause any symptoms in the case of cancer until the tumor grows large. Patients understand symptoms in some situations, but are not sufficiently resourceful to investigate the causes of their symptoms.

After a detailed search across the internet, interviews with doctors for sickness and symptoms, the information base has been populated. Several sources were utilized and the symptoms list in the knowledge base was created for each disease. By choosing from 228 symptoms or typing their symptoms in a sentence form, the device allows users to check for their symptoms. The knowledge base and text mining method measures the possibilities for each disease according to the given symptoms, and shows the probabilities. A hyperlink that describes each disease is generated by each possibility.

5. CONCLUSION

This paper was set out to solve "information-glut" issues when handling large volumes of text documents and inability to easily locate relevant details from big data. The application of text feature extraction (TF-IDF) and Random Forest Classifier provides a simple and effective solution to the above problem and ensure early mining of relevant data

A domain based technique for knowledge acquisition and text mining in a database was developed. In summary, the results and contributions obtained from this dissertation are:

i. The knowledge based system developed for disease acquisition.
ii. An informative text mining tool for people with disease symptoms.
iii. Identification of a class of disease associated with its symptoms.
iv. A novel strategy of applying text mining module to check and extract disease and calculates their possibilities.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10427

REFERENCES

[1]. Shang, Y., (2005). Expert Systems. In: The Electrical Engineering Handbook. Wai Kai Chen Editors. *Academic Press.* 1171 – 1208.

[2]. Mahmoud A, Abd El-Aziz A.A and Hesham A.H.(2018) A Survey on Big Data and Knowledge Acquisition Techniques.*Dept. of Information Systems & Technology, Institute of Statistical Studies and Research Cairo University*

[3]. Giadomi, A. and Haider, M., (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management. 35*(*2*): 137 – 144.

[4]. Feldman R., Aumann Y., Amir A., Klösgen W. and Zilberstien A. (1997). Maximal Association Rules: A New Tool for Mining for Keyword co-occurrences in Document Collections, In *Proceedings of the 3rd International Conference on Knowledge Discovery*, KDD.

[5]. Chen, T. C. and Hsu, T. C. (2006). A GAs based approach for mining breast cancer pattern. *Expert Systems with Applications*, 30, 674–68

[6]. Henriksson. A, Zhao.J, Dalianis.H, and Bostrom.H. 2016. "Ensembles of ¨randomized trees using diverse distributed representations of clinical events," *BMC MedicalInformatics and Decision Making*, 16(69*).*

[7]. Laxman. B and Sujatha. D,"improved method for pattern discovery in text mining"*international journal of research in engineering and technology,vol.2,no. 1,pp.2321-2328,2013*

[8]. Silvia. M, Rodrigo. M.(2018) *An Approach based on text mining for knowledge acquisition in diagnostic systems*

[9]. *Waykole. R and Thakare. A. (2018). A reviewof feature extraction method for text classification international journal of advance engineering and research development, 5(4), 351-354*