Design of a Support Vector Machine Aided Real Time Health Information Text Extraction System

Enikuomehin O.A.
Department of Computer Science
Lagos State University, OJO
Lagos, Nigeria
Email: toyinenikuomehin@gmail.com

Abstract

The exponential growth of available information sources has greatly affected the access to useable health information. As a consequence, medically biased information has become difficult to use for decision making. In this paper, we consider these consequences and present an enhanced method for accessing health information in real time. The approach involves the use of the vapnik Support Vector Machine process for text classification. The proposed method was frameworked on php/mysql for web user. Experimental setup shows that the method outperforms the baseline in the Precision, Recall and F1 measures. An extension using the Gaussian kernel is recommended in the paper.

Key words: Health Information, Support Vector Machine, Information Extraction

Received:05.07.2016. Accepted:02.05.2017

1. INTRODUCTION

Health services delivery using Information and Communication Technology (ICT) facilities has become one of the fastest growing sections of research in recent times with very large variety of application areas ranging from health tips awareness, web biased surgical processes to brain mapping, amongst others. The health care industry has always aimed at improving the quality of health care delivery at reduced cost. Adequate medically biased information, Health Information (HI), which deals essentially, with the resources, devices and methods required to optimize the acquisition, storage, retrieval and the use of information in health institutions, plays a major role in achieving the stated goal.

HI help doctors to make more informed decisions without wasting much time such that their actions cumulatively add to the process of saving a patient's life. However, due to the heterogeneous nature at which most of the information is received; there has been growing concerns on the effective usage of this information because they are snot well structured. Doctors providing care for their patients need to know what symptoms the patient has, allergies, drugs previously taken; test carried out, amongst others and in most cases, require the consultation of more recent medical journals so as to be in tune with latest medical practice as applicable. For a patient with say X and Y diseases, medical practices try to get relevant information about the diseases, such information include the symptoms, the contagious form, type(virus, bacteria etc), but this is only achievable when related records are appropriately kept in searchable repositories. Thus, problem occurs when the information is not readily available. If the information is digitized and formatted appropriately, it can be easily be queried. The phrase "text extraction" comes to play a vital role in HI by addressing the deficiencies of unstructured information. Information Extraction is used to denote any system that analyzes large quantities of natural language text and detects lexical or linguistic usage patterns in an attempt to extract probable useful (although not probably correct) information from multiple free texts. In this paper, the effect of the continuous growth of information system is considered and appropriate method for real-time health information extraction is developed.

An enormous amount of information exist only in natural language form and to allow this information to be automatically manipulated and analyzed, it must first be distilled into a more structured form in which the individual facts are accessible. For example, the curators of genome databases would like to have tools that could accurately extract information from scientific literature about entities such as genes, proteins, cells, diseases, etc. Health IT can help in monitoring patients' health status and make specific and targeted recommendations to improve patients' health. Access to real-time data through electronic health records and health IT helps, by using clinical decision support to highlight care options tailored to patients, improving safety by highlighting drug interactions or allergies when ordering medications, connecting patients with community and educational resources to better manage their health [3].

2. STATEMENT OF PROBLEM

The growing use of modern information technology increases the amount of documents, information and data accessible to health professionals. To overcome this overload, intelligent technology are demanded to support the information seeking tasks of health services provider[1] but the problem of identifying useful knowledge from unstructured text is becoming an important aspect to consider [2]. Due to the complexity of dynamically changing biomedical terminologies, term identification has been recognized as the current bottleneck in text extraction, and consequentially, topics in both natural language processing and biomedical communities are now being deeply studied [4]. The aim is to design a system that extract health information from unstructured text document in real time such that health professionals can automate parts of the diagnostic and treatment procedure. The process will adopt a supervised machine learning approach to extract health information, in which models of the text are traced from human annotated example and used to examine the use of pre-existing medical terminologies from their knowledge resources.

The structured and narrative components of health records collectively provides a comprehensive account for the medication of patients in which continuity of care is supported by preventing both redundant and repeated action to save time.

2.1 Information Retrieval And Information Extraction Task

The task of IR is to select from a collection of textual documents, subsets which are relevant to a particular query, based on key-word search and possibly augmented by the use of a thesaurus. The IR process usually returns a ranked list of documents, where the rank corresponds to the relevance score that the system assigned to the document in response to the query.

In some cases, ranking of documents do not necessarily correlates to relevance to users. This is compounded with the user's inability to precisely define what he or she needs. Thus, the major task in extraction is to aid the processes of finding the most appropriate information.

The task of Information Extraction (IE) is to identify a predefined set of concepts in a specific domain, ignoring other irrelevant information, where a domain consists of a corpus of texts together with a clearly specified information need. In other words, IE is about deriving structured factual information from unstructured text [5]. There are many ways of expressing the same fact, which can be distributed across multiple sentences [6], documents, or even knowledge repositories. Furthermore, a significant amount of relevant information might be implicit, which may be difficult to discern and an enormous amount of background knowledge is needed to infer the meaning of unrestricted natural language.

Information Extraction has not received as much attention as Information Retrieval (IR) and is often confused with the latter. IE systems are in principle, more difficult and knowledge-intensive to build than IR systems. However, IE and IR techniques can be seen as complementary and potentially combined in various ways. IR is often used in IE for pre-filtering very large document collection to a manageable subset, to which IE techniques could be applied. Alternatively, IE could be used as a subcomponent of an IR system to identify structures for intelligent document indexing.

Research in the late 1990s and the beginning of the twenty-first century resulted in significant advances in IE (and in NLP in general) in terms of emergence of Knowledge Engineering-based, modular IE systems that was able to process vast amounts of textual data robustly and efficiently, including in languages other than English. However, the process of handcrafting language and domain-specific resources and components remained a very time consuming and difficult task. This stimulated research on trainable IE systems that deploy machine-learning (ML) techniques to offload some of the burden of customizing a general-purpose IE engine to a new domain or task.

The first "wave" in the shift away from KE-based approaches and toward trainable systems focused on supervised machine-learning approaches. The principal motivation behind this trend is to shift the human effort in customization of the IE knowledge bases away from knowledge engineering and toward annotating training data, which serves as input to machine-learning algorithms [15]. This promotes further modularity in development, since knowledge engineering requires effort from both the system developer and the domain expert, whereas data annotation, in principle, requires mostly an effort on the part of the latter. The broad array of tools and algorithms from supervised learning that were already in wide use in other areas of NLP came to be applied to IE tasks as well. This trend affected all IE-related tasks [15]. These features will typically include words and terms (possibly consisting of multiple words) present (or absent) in the segment, entities found in the segment, other relations, and other syntactic and semantic information extracted from the segment. Some of the earlier literature on application of supervised learning to IE tasks includes, e.g., [7], [8]. Conditional Random Fields (CRFs), another supervised learning method that has become popular in the recent years, has also been applied to IE tasks. CRFs are similar to HMMs in that they are good at modeling local dependencies, and they perform especially well on the "lower-level" tasks—dereferences are available for application of supervised learning algorithms [9],and in particular to NLP task [10], named entity classification [10], as they are well suited for sequence-labeling tasks in general. CRFs can be used to model certain variants of the relation detection task as well.

2.2 Health Information Text Extraction

Techniques for automatically encoding textual document from medical records have been evaluated by several groups. Examples are the Linguistic String Project and MedLEE (Medical Language Extraction and Encoding system) [11]. MedLEE has been recently adapted to extract UMLS concepts from medical text documents, achieving 83% recall and 89% precision. Other systems automatically mapping clinical text concepts to a standardized vocabulary have been reported, like Meta Map , Index Finder, and Knowledge Map. Meta Map and its Java version called MMTx (MetaMap Transfer) were developed by the US National Library of Medicine (NLM). They are used to index text or to map concepts in the analyzed text with UMLS concepts. Meta Map has been shown to identify most concepts present in MEDLINE titles. Meta Map has been used for Information Retrieval and for Information Extraction in biomedical text, and to extract different types of information like anatomical concepts or molecular binding concepts. Meta Map has also been used with patient's electronic messages to automatically provide relevant health information to the patients [3]Open-source clinical NLP systems such as HITEx and cTAKES were reported in 2006 and 2010, respectively. Finally, in a study by Shadow and McDonald, the system extracted the most critical pathology findings in documents.

Most recently, multiple research efforts focused on medication information extraction, such as MERKI, MedEx and others. In 2009, the Third i2b2 workshop on NLP Challenges for Clinical Records, also referred as the medication challenge, focused on the extraction of medication information from discharge summaries. Many NLP applications in biomedical informatics automatically encode clinical text to concepts within a standard terminology. UMLS is widely used since it includes various controlled vocabularies and provides mappings among them [12]. Some studies map terms in clinical text to a specific terminology, such as MESH terms, SNOMED_CT, ICD-9-CM, or RxNorm. These studies applied approaches based on string matching, statistical processing, and NLP techniques (e.g., term composition, noun phrase identification, syntactic parsing, etc). In addition, many previous NLP studies reported evaluation of their systems' performance using inpatient reports, such as radiology reports and discharge summaries. Outpatient clinical notes often contain unique formats and characteristics. The proposed Data Extraction Framework is of the form:

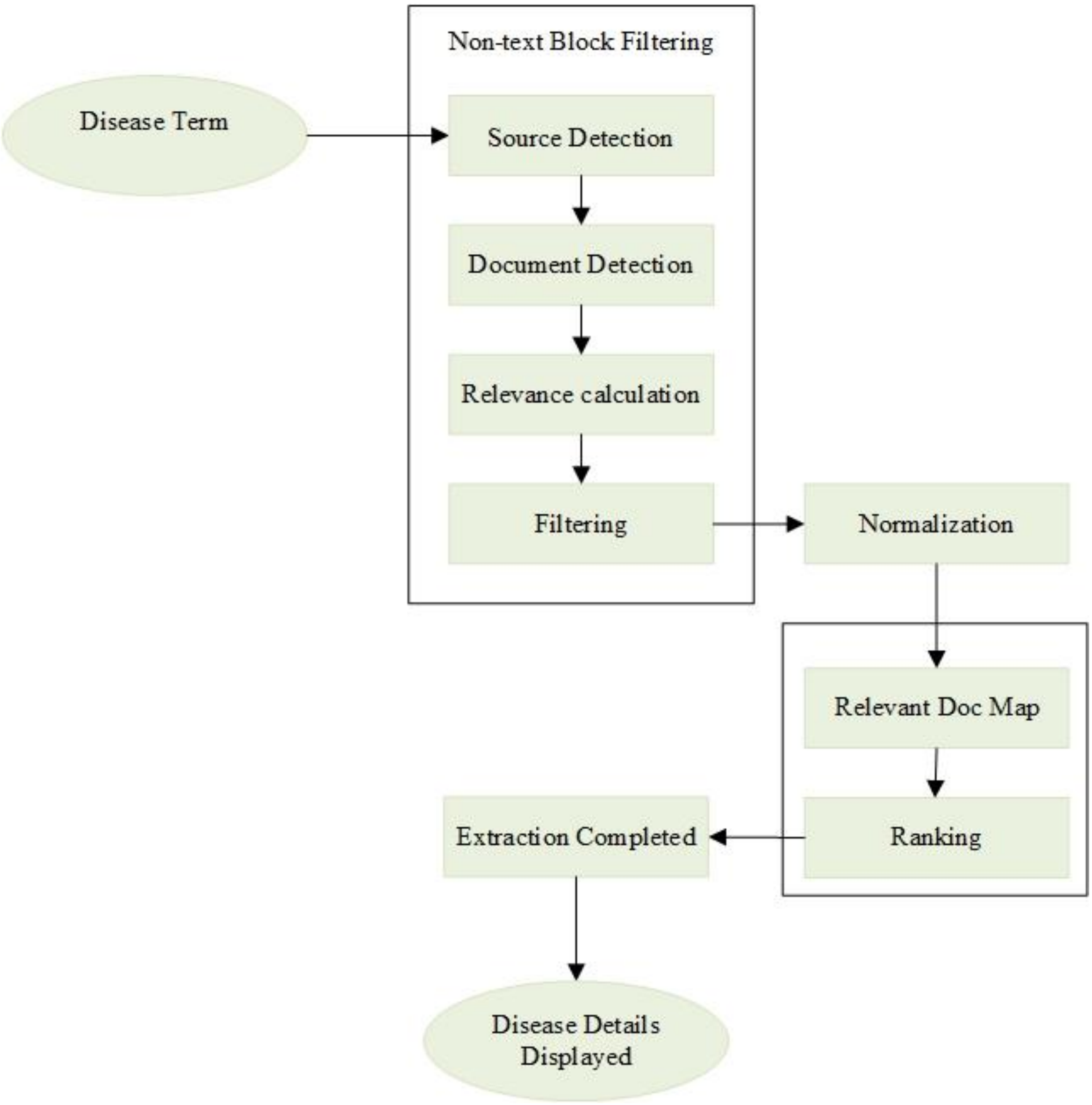

Figure 1. Proposed Data Extraction Framework. Literally, a data flow diagram for the above can be simplied as shown in figure 2 below:

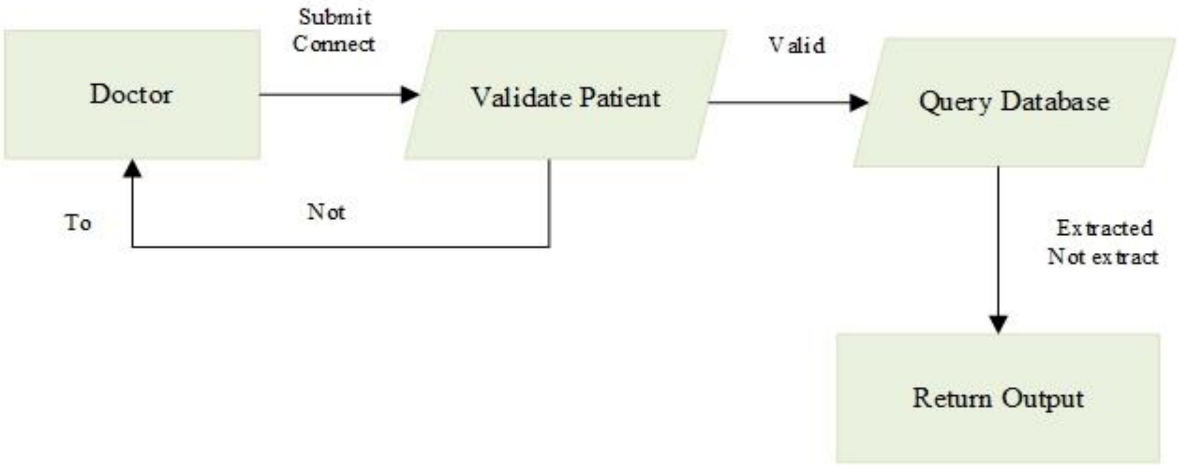

Figure 2. Dataflow Diagram for the new system. Analogically, the developed program allows the Doctor to submit the patient's symptoms and connects to the network, Validates patient's symptoms, Query database then the system sends output back to the doctor. The application was developed in Php/mysql and some snapshot of the system is provided in further session.

3. SCHEMATIC OVERVIEW OF THE PROPOSED SYSTEM DESIGN

Top-down development was used so as to achieve a structured programming which is directed at developing the program in an orderly way, decomposing the requirement into well specified high level model. An architectural design is first made to break the system into modules and then a detailed design of all the modules is carried out. The system design is logically divided into two (2). They are a)
The User Interface b) The Database

.

a. The User Interface

The simple and interactive user interface has a page which has a form where the user submits the patient's data (symptoms).after been submitted, the output which includes the extracted item is displayed.

b. The Database

This database is in a defined format in a location. The patient's symptom supplied serves as the primary key that uniquely identifies a record in the database. The database used to represent the different diseases is MySql and is structured as shown in fig 3 below:

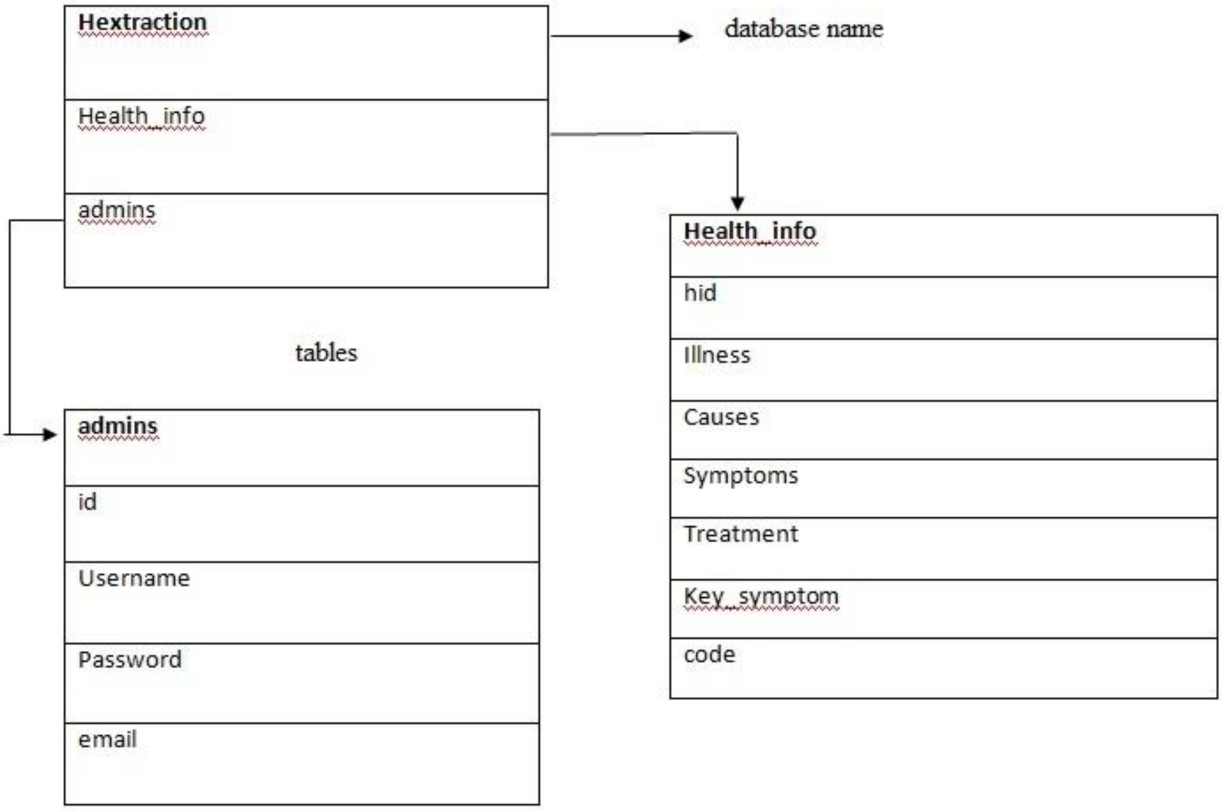

Figure 3. Database diagram for the new system

3.1 Input-Output Design

The way this program is designed, input is only added at the point of usage i.e. during the implementation of the work. However, the data that should be inputted into the system is only the patient's data which in this case is the symptoms of the patient. The way the program is designed, output is only generated after querying the databases and the likely output are; patient's diseases, causes, treatment or the remedy. This is detailed in the flowchart below:

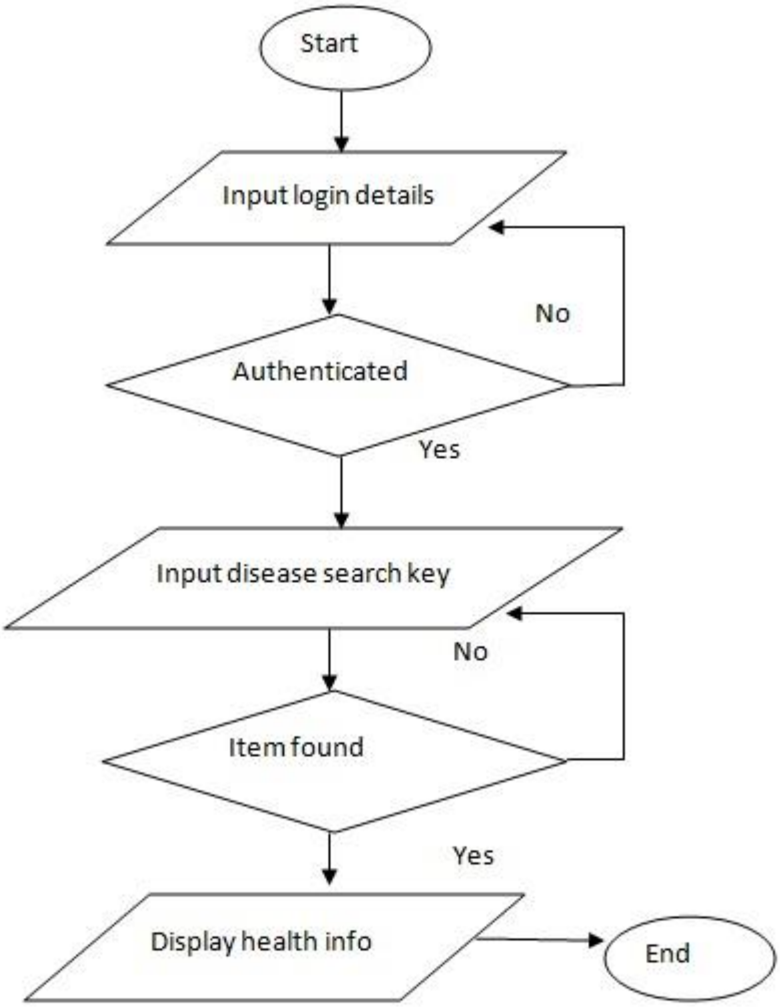

Figure 4. flowchart diagram for Health Information Text Extraction System.

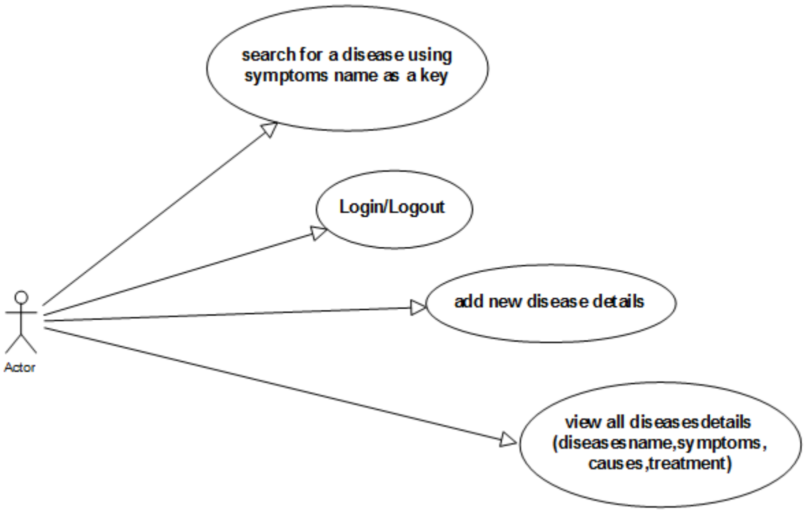
search for a disease using symptoms name as a key
Login/Logout
add new disease details
view all diseases details (diseases name,symptoms, causes,treatment)
Actor

Figure 5. use case diagram for Health Information Text Extraction System.

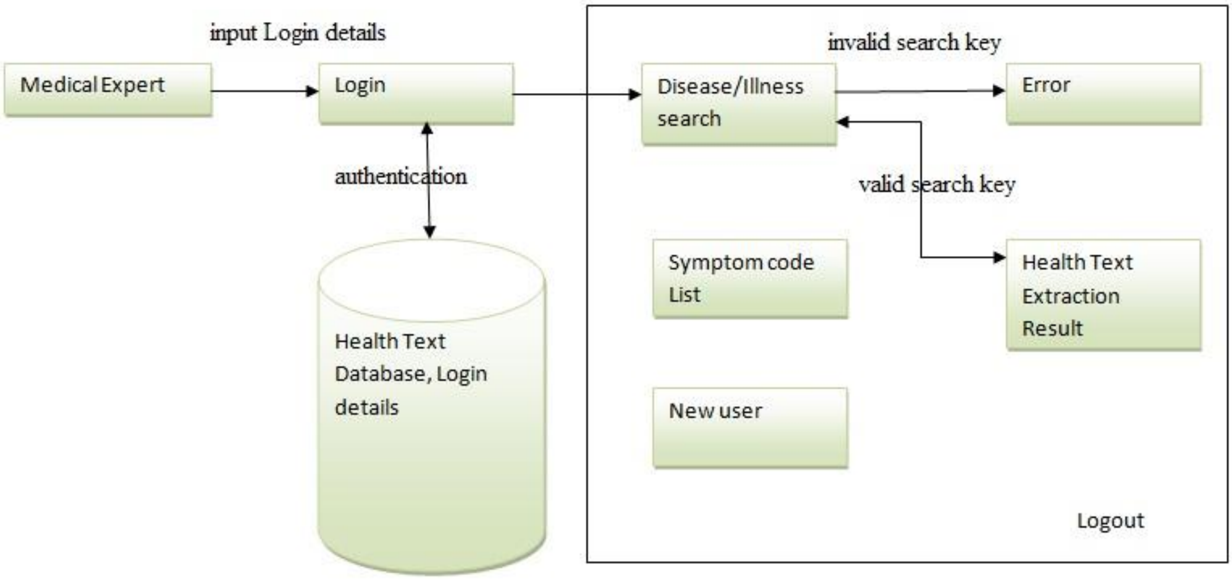

Figure 6. Data flow diagram of health information Text Extraction System.

4. THEORETICAL MODELLING

If we consider a two set classification problem of the form x and y and we assume thatt {(x_1, y_1), … , (x_N, y_N)} is set up as a training data, where x_i denotes a concept (a feature vector) and $y_i \in \{-1,+1\}$ denotes an index class. The general approach is to develop a separating paradigm for two set where one corresponds to the linear SVM class. Studies has shown that error bounds for such sets are always minimal then the Linear SVM can be further extended into non-linear Support Vector Machines by using as Gaussian kernels. Using Vapnik's Support Vector Machine [13, 14], we can classify the search terms such as disease name, symptoms, previous treatment etc whose framework can be used to perform some data analysis on this method. In the experiment setup, set of disease, their symptoms and treatment were individually run on search engines. The performance as search result were documented and analyzed as follows.

In the experiments on extraction, we conducted evaluations in terms of Precision, Recall, F1-measure, and Accuracy. The evaluation measures are defined as follows:

$$\text{Precision: } P = \frac{X}{X+Y}$$

$$\text{Recall: } R = \frac{X}{X+K}$$

$$\text{F1} - \text{Measure: } F1 = \frac{2PR}{P+R}$$

$$\text{Accuracy: Accuracy} = \frac{X+Q}{X+Y+Q}$$

where X,Y,K and Q denote number of
be the disease symptoms or the search

Table 1: Instances table on results of extraction

	Is Required	Is Not Required
Available	A	B
Not available	C	D

In the evaluations, the item and information source needed is tagged as a ‘Required’. If the method can find the required, then we conclude that the approach is effective and otherwise we consider the approach not of major significance. Precision, Recall, and F1-measure are calculated on the basis of the result. For symptom, allergies, and treatment pattern, we conduct evaluation at the document level and for the other tasks like disease name, we perform evaluation at term level

4.1 Experiment

We evaluated the performances of retrieval process for the Class A search terms and Class B terms were first normalized on the first data sets used. We conducted the experiment in the following way. First, we conducted search using the disease name and consider the returns, then we performed the symptoms search with expectation that the symptoms search will have a larger return than the disease search which can allow for a more proper diagnostics. Then, we conducted the matching. Finally, we conducted sentence normalization and term normalization with precision, recall and F1 analysis being carried out at each level. We also made comparisons with the baseline methods as described above.

Table 3 shows the result of the extraction task on a set of search terms for the disease name and symptoms.

Table 3. Performances of non-text filtering and text normalization (%)

Extraction Task		Precision	Recall	F1-Measure
Disease Name	Our Method	96.95	97.42	97.19
	Baseline	99.81	64.42	78.30
Symptoms	Our Method	91.33	88.38	89.83
	Baseline	88.54	23.68	37.36
Report		98.18	92.01	95.00
Program Code		92.97	72.17	81.26
Drugs/Treatment	Our Method	85.53	97.65	91.19
	Baseline	63.55	98.13	77.15
Term Identification		94.93	93.91	94.42

The experiment shows that our approach achieves a higher performance in the set experiment. To statistically verify our result, we carried out a test on the outcome and found that the *p* values are much smaller than 0.01, which means the proposed approach is statistically significant. The recall is recommended to be improved upon

5. CONCLUSION AND RECOMMEDATION

Health Information Text Extraction, which is the area of Computer Science studied in the course of this work, is still relatively a green field that actually has been merely tilled. It has come to stay and more research work is still being carried out to fully explore the endless possibilities that it can bring to the data extracted as well. The program developed in this work can easily be extended and modelled for different other information extraction related purposes.

In conclusion, Health Information Text Extraction System provides a solution for quicker and easier health text extraction. A system that can scale to meet organizational demands for timely, relevant, trustworthy data has been developed and experimental framework has shown the possibility of using the enhanced SVM to achieve a better retrieval result for medical searches.

REFERENCES

Kannampallil, T. G., Franklin, A., Mishra, R., Almoosa, K. F., Cohen, T., and Patel, V. L. (2013). Understanding the nature of information seeking behavior in critical care: implications for the design of health information technology. Artificial intelligence in medicine 57(1), 21-29.

Savolainen, R. (2015). Cognitive barriers to information seeking: A conceptual analysis. Journal of Information Science, 0165551515587850.

Krist, A. H., Beasley, J. W., Crosson, J. C., Kibbe, D. C., Klinkman, M. S., Lehmann, C. U., ... and Peterson, K. A. (2014). Electronic health record functionality needed to better support primary care. Journal of the American Medical Informatics Association 21(5), 764-771.

Prince, V. (Ed.). (2009). Information Retrieval in Biomedicine: Natural Language Processing for Knowledge Integration: Natural Language Processing for Knowledge Integration. IGI Global.

Piskorski, J., and Yangarber, R. (2013). Information extraction: past, present and future. In Multi-source, multilingual information extraction and summarization (pp. 23-49). Springer Berlin Heidelberg.

Aone .C., Halverson .L., Hampton .T., Ramos-Santacruz .M., and Hampton .T. SRA: (1999), “description of the IE2 system used for MUC-7 In: Proceedings of MUC-7”.

McCallum, A., and Li, W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 4, 188-191). Association for Computational Linguistics.

Aone .C., Ramos-Santacruz .M. (2000) “REES: a large-scale relation and event extraction system. In: Proceedings of the 6th Conference on Applied Natural Language Processing, Seattle 76–83.

Xu, Y., Zeng, X., and Zhong, S. (2013). A new supervised learning algorithm for spiking neurons. Neural computation 25(6), 1472-1511.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature 521(7553), 436-444.

Doan, S., Conway, M., Phuong, T. M., and Ohno-Machado, L. (2014). Natural language processing in biomedicine: a unified system architecture overview. Clinical Bioinformatics 275-294.

McInnes, B., Liu, Y., Pedersen, T., Melton, G., and Pakhomov, S. (2013). Umls:: similarity: Measuring the relatedness and similarity of biomedical concepts. Association for Computational Linguistics.

Schölkopf, B., Smola, A. J., Williamson, R. C., and Bartlett, P. L. (2000). New support vector algorithms. Neural computation 12(5), 1207-1245.

Keerthi, S. S., and Lin, C. J. (2003). Asymptotic behaviors of support vector machines with Gaussian kernel. Neural computation 15(7), 1667-1689.

Poibeau, T., Saggion, H., Piskorski, J., and Yangarber, R. (Eds.). (2012). Multi-source, multilingual information extraction and summarization. Springer Science and Business Media.