Journal of Theoretical and Applied Information Technology
30th June 2016. Vol.88. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A MACHINE LEARNING APPROACH FOR IDENTIFYING DISEASE TREATMENT RELATIONS IN SHORT TEXTS

T.V.M. SAIRAM, DR. G. RAMA KRISHNA

M.Tech (Cloud Computing), K L University, India

Professor (Dept. Of CSE), K L University, India

E-mail: sairam.tadepalli1904@gmail.com, ramakrishnag_cse@kluniversity.in

ABSTRACT

The Machine learning (ML) region has proved its power in almost every industry and is currently a reliable technology in health care industry. Computerized study of the clinical industry includes suitable care choice guide, healthcare photo and DNA connections. ML is recognized as a tool employing computer systems integrating health care mechanisms resulting in more appropriate care and attention patients and further study or research on a disease. This paper provides powerful algorithms and techniques used in diagnosing illness using remedy associated phrases from brief published written text launched in health-care documents. The objective of this work is to show how Natural Language Processing (NLP) and Machine Learning strategies can be used for reflection of information and what class strategies are appropriate for determining & figuring out suitable care information in brief published written textual content. This paper additionally focuses on suitable care analysis therapy & prevention of contamination, infection harm in human. The system found out some assignment of clinical suitable care statistics, health-care control, and man or woman health data and so forth. The proposed method may be incorporated with any health-care management software to make better suitable care selection. The inpatient management application can instantly mine bio-medical data from virtual databases.

Keywords: *Health-Care, System Mastering, Natural Language Processing, Aid Vector Machine, Choice Aid System.*

1. INTRODUCTION

Machine Learning (ML) is well known for its performance in the pre-designed manner which will give the experimental results in apt manner which are expected. In the existing system we have a student data which will give the information regarding the understanding level of the student in his academics and also what are the problems they are being faced in their educational life. So researchers used Radian 6 tool which is a social networking monitoring tool which will monitor students twitter tweets and gather them as a dataset. Proposed system is not an extension for the existing system. The concept of providing most prioritized data to the researchers or to any individual is easy and important task for this tool. MEDLINE is like a datawarehouse which have the articles regarding the Biomedical data and the journal and research papers regarding the treatment and new inventions in the treatments of deadly diseases. But the case here to study the entire thousands of articles is difficult for a single person who want to perform a new research on specific field. So we provide the ranking system for the articles using machine learning technology and data mining. This tool will work on the process of data from the PubMed site and the analysis is performed based on the basis of Google analytics and Microsoft medical invention based.

Basically medical articles will have the images and their explanation based on the results researchers achieved. The below diagram will explain us the things we may have in medical articles mostly.

Journal of Theoretical and Applied Information Technology
30th June 2016. Vol.88. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

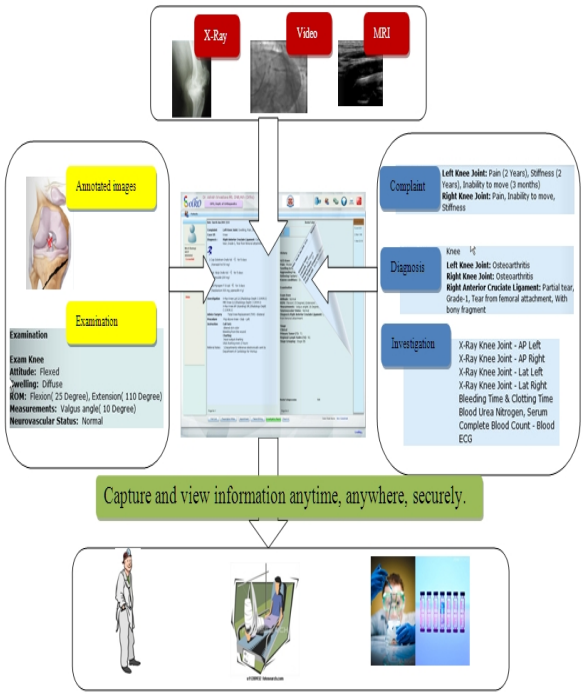

Figure 1: Automated Clinical Case Sheet Class For Fast Know-How Of Patient's Profile.

It's miles better to perceive and remove the phrase that doesn't include statistics appropriate to infection or remedies. With those kind of phrases we dont have any use, because of those the result may vary. The final terms can be labeled in step . It is going to be very complicated to understand the exact answer if the entirety is done in a single phase with the aid of classifying sentences relying on attention and also which include the sentences that do not offer suitable information. Relation Extraction is a conventional evaluation concern in natural Language. Health-care details are stored in textual shape some of the scientific information stored in "**MedLine**". Manually getting useful information from large volume of facts supply is an uneventful work. Moreover HTML web page displaying medical details carries scientific statistics and usually inappropriate additives which includes navigation pictures, guidelines, customer feedback, advertising, comments and so on. The suggested approach supports in clinical making selections via offering health-care practitioner(MBBS) with great to be had proof of scientific info. In this paper we choose textual content exploration centered with clinical papers relevant to scientific remedy. "**MedLine**" is chosen in this challenge to get bio-medical information as it provides answers related to man or woman therapy and it's the facts supply that's most widely utilized by the physicians and analysis college students in scientific vicinity. Even more essential it is frequently up to date and the materials are grew to become out to be specific in contrast to different health-care websites imparting details relevant to human illness, fitness, medicinal drugs, therapy and many others. With the growing variety of health-care dissertation, evaluation documents, research articles, scientists ought to face the issues of analyzing quite a few analysis files to obtain understanding in their discipline of interest. Google like Pub Med decreases this limit via gaining access to the correct papers suitable to the client question(Search pattern in online access). In this task all the inappropriate data like advertising etc defined inside the above passage are removed and text exploration is performed at the extracted report from which information or terms suitable to consumer detailed contamination is produced. From the produced report symptoms, reasons, treatment of the particular contamination is strained and shown to the consumer. Hence the consumer receives the required information by this tool which lets you keep his/her time and enhances the importance of the result. Health-care subheadings and situation going may be used to infer courting among scientific ideas.

2. BACKGROUND WORK

Info locations have been used for acquiring our application efficiently. The data set includes terms from "**Medline**" abstracts annotated with infection, cure, prevention etc and with 8 semantic interplay between illnesses and treatment options. The primary goal in their perform is on agency identity the illnesses and related treatments. In bio-medical literary works, rule-based techniques had been normally used for fixing regards elimination duties. The number one assets utilized by this approach are both syntactic: part-of-speech (POS) and syntactic structures; or semantic information with the aid of set styles that incorporate phrases that result in a certain regards. One of the risks of the usage of techniques relying on tips is they normally require more human-expert strive than statistics-driven techniques . MEDLINE is the source for the retrieving of the datasets based on the search key which is based on the semantic web concept. The semantic web is the concept of searching the warehouse based our search request which will give the apt result by excluding the unwanted data etc. This semantic concept is applied mostly on the social network data or on other kind of media data. Biomedical media is

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

used for our approach which reduces the burden to collect the data and form datasets from it. Sentence analysis is applied in this tool and the related data is collected in an order in the CSV file. So that it will have some ID which is unique and based on that ID we can recognise the paper or article which is having more rank and highest priority. The primary method (mission 1 or phrase choice) acknowledges terms from Medline launched abstracts that talk ailments and healing procedures. The technique is much like a test out of terms inside the subjective of an content in an effort to existing to the user-handiest terms which can be diagnosed as containing suitable info (disease treatment statistics). The second manner has a similarly semantic sizing and it is centered on determining disorder-treatment interplay within the terms already selected as being beneficial. Even though the syntactic information are resulting from sources that aren't a hundred percent unique, testimonials with those varieties of techniques were experienced in the bio-medical region. To try this work correctly historically performs two tasks for downloaded statistics units.

3. PROPOSED APPROACH

The two tasks used in this paper are the basis for the improvement of generation structure. This tool permits to recognize the health-care relevant info from abstracts. The first manner gives with removal all details concerning illnesses and treatments while the task offers with elimination of relevant information present among contamination and remedies. The shape designed with those initiatives is used by health-care care vendors, individuals who wish to manage their health-care relevant issues and corporations that build systematic views. The destiny item may be presented with browser plug-in and computer software so that it lets in the purchaser to get all info relevant to illnesses and treatment options and also the regards between those entities. It is also be beneficial to understand more about latest findings relevant to medication. The object can be designed and sold by using organizations that do analysis in well being right care domain, Terminology Managing like Natural Language Processing (NLP), and tool reading using Machine Learning (ML), and companies that create assets like Microsoft health container and Google health. This object is valuable in e-trade areas by way of displaying the information that the details presented here are particular and also offer all the modern findings applicable to health proper care. To make a item greater famous it need to be believe in deserving in order that individuals should purchase it. It's far the key component for any company to make object a hit. Wish-list coming to well-being maintenance structures it must be greater consider in deserving due to the fact that it's far dealing with health-care applicable problems. Businesses that wish to sell tool with right health-care structure need to create resources that mechanically draw out the wealth of analysis.

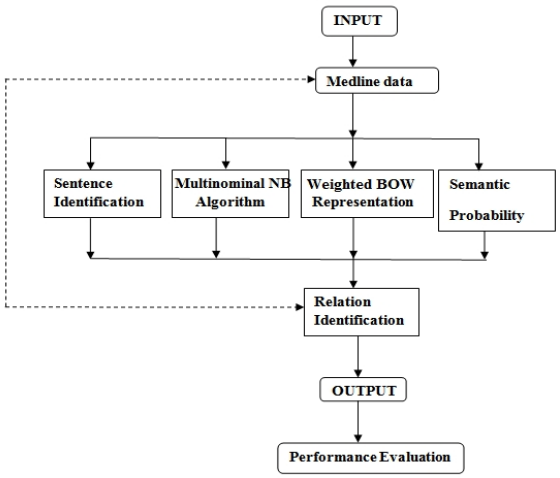

Figure 2: Proposed Architecture For Solving Data Sets Based On Their Relevancy Of Key Word Search.

For example the information presented for illnesses or treatments needs to be based on present day findings on well being right care vicinity in order that individuals can agree with it. The excessive exceptional also must be taken right care so that it affords dynamic content for customers. The primary method offers with the identification of terms from the Medline abstracts that offer the information approximately the illnesses and therapies. In other words it additionally seems like scanning the phrases from Medline abstracts that contain relevant information which the patron desires. Natural Language Processing (NLP), and Machine Learning (ML) are used to attract out particular info or it is able to also say that it flawlessly eliminates the undesirable details which are not applicable to infection or remedy. Natural Language Processing coping with (NLP) and Machine Learning (ML) itself contain in getting useful phrases. It's far trial to apprehend the informative phrases in regions including summarization and details elimination. In this propose approach we use hybrid classification approach with two classification algorithms. They are SVM (Support Vector Machine) and CNB (Component Naive Bayes

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Classification). In these both classifications we use same datasets which are downloaded from the MEDLINE repository. So that we need to assign one classification for one time and then for the next time we need to give another one. Consider if we use the SVM algorithm we need to select the type f ranking we want. That is based on three categories. For the CNB also we have the same three categories. Those three categories are 1) Phrases 2) Word Frequencies 3) Medical Concept Optimization. Phrases is considered as the data with some specific phrases which are related to the specific diseases. In some case we don't find a phrase related to the paper in the title. They will be seen in sentence based text in the article. So we need to get the phrases from that article using sentence analysis. Word Frequency is based on the specific words appearing the whole document. Which shows that the interest of author in the specifying the treatment of the specific disease and how we can find more information regarding that disease in that article. Medical Concept optimizing is with three argument. Those are cure, prevent , side effects.

4. PERFORMANCE EVALUATION

Records Extraction

Records extraction in this approach is getting the updated records over the internet from the PubMed site. PubMed repository will have all the published journals or articles in their data warehouse. If the researcher needs the updated journals information and if it is not available in the datasets location in our current system then we need to perform records extraction.For this we need to give a keyword such as "Cancer" in the search space then it will retrieve all the related records of Cancer and create a data set in the data sets location with current data and the name of the search key. This will give the complete info regarding the updated articles info in the form of data sets.

Bag of words

The Bag of Words(BOW) is a collection of the phrases related to the data sets and the context mention in he search key. For a search key the sentence analysis is based on the training set that is pre defined which can be downloaded from the PubMed site. For SVM and CNB classification we need to give two types of CSV files. One is Training set which is common and the other is Test set which is needed to upload for the testing and generating the ranks for the documents and articles.

Genie Tagger (Phrases)

Type of reflection is based on syntactic records: noun-phrases, verb-terms, and biomedical ideas diagnosed within the phrases. The tagger examines British phrases and consequences the speech labels, quantity labels, and business enterprise labels. The tagger is mainly up to date for biomedical written text together with Medline abstracts. The noun and verb-phrases diagnosed by way of the tagger are capabilities used for the second one reflection method. Here we have an example output which is expected to get after gunning the tool.

```
pancreatic cancer >> NP
pancreatic cancer >> NP
have >> VP
a unique relationship >> NP
a unique relationship >> NP
Genetic mutations >> NP
Genetic mutations >> NP
activation >> NP
activation >> NP
the KRAS2 oncogene >> NP
the KRAS2 oncogene >> NP
inactivation >> NP
inactivation >> NP
the tumor-suppressor gene CDKN2A >> NP
the tumor-suppressor gene CDKN2A >> NP
inactivation >> NP
inactivation >> NP
```

Figure 3: Practical Application Development In Pub Med Data Extraction.

For this present method, determining which words from the abstracts of Medline content material that comprise useful words of illnesses and treatments, the fine outcomes acquired. The reflection method that makes use of BOW features, noun and verb-terms, and biomedical data with the CNB classifier acquire a 90%.

5. EXPECTED EXPERIMENTAL RESULTS

This section talks about the results we will acquire for the two responsibilities on this study. The two classifiers (SVM and CNB) have continually been confirmed to execute properly on textual content category tasks. We calculate individual score of each relation primarily based on expressed biomedical family members of every preferred

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

system on records sets with semantic family members.

Table 1: Expected Data Score Of Related Data Items.

Abstract Number	SVM Classifier	CNB Classifier
1	92	80
2	90	70
3	90	80
4	90	45
5	96	79

One of the vast efforts of this work is the fact that the cutting-edge exams show that more information inside the reflection configurations supplies traits for the mission of figuring out useful phrases.

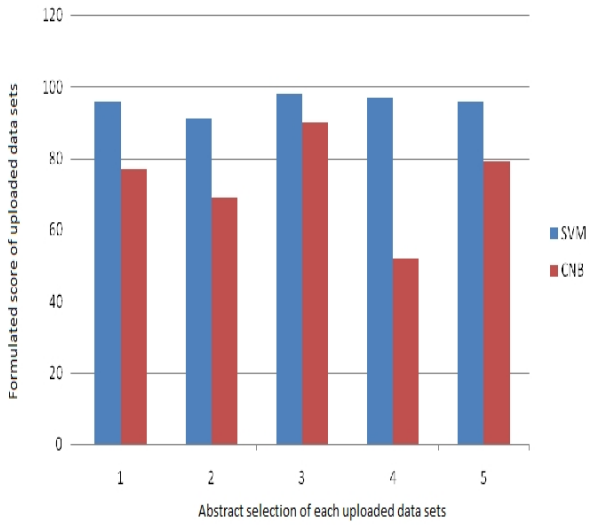

Figure 4: Expected Formulated Score Of The Uploaded Data Sets.

```
==================================================
******Classifying Abstract : 3********************
Abstract Clasification Score of 27026362 is : 98
Terms Belonging to "Cure" class        : 0
Terms Belonging to "Prevent" class     : 0
Terms Belonging to "SideEffects" class : 1
==================================================
******Classifying Abstract : 4********************
Abstract Clasification Score of 27025371 is : 91
Terms Belonging to "Cure" class        : 0
Terms Belonging to "Prevent" class     : 0
Terms Belonging to "SideEffects" class : 0
==================================================
******Classifying Abstract : 5********************
Abstract Clasification Score of 27020417 is : 91
Terms Belonging to "Cure" class        : 0
Terms Belonging to "Prevent" class     : 0
Terms Belonging to "SideEffects" class : 3
==================================================
```

Figure: 5: Expected Scores In SVM For Medical Concept Optimization

The expected experiment results which give the basic idea on the process of ranking the articles based on the BOW(Bag Of Words), SVM and CNB classifiers. The Classifiers will work based on the training and test data sets provided to the machine which will calculate the data sets and then they will be ranked according to the phrases, frequency f the words used in the context and mainly based on the abstracts. Graph is not shown in the application because, this tool is primarily being used and testing on the local system that is stand alone system. To generate the graph we need to use the Canvas technology and need to assign the results opted to that canvass to generate the graph. In the tabular form we can see the expected ranks which we can get on using the SVM and CNB. There will be a slite difference in the ranking in the CNB ans SVM. But the process will be same.

6. CONCLUSION

This tool is with respective of the datasets which we obtain as the result of the articles will help us fort assigning the ranks for the articles in the MEDLINE. With this pubmed data we can find out the clinical structure of the pubmed and the idea the researchers have in treating different diseases in their own approach. We need to concentrate on the better result which will be helpful for the researchers to know in which

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

research article the work done on a specific disease is apt and the results which they mentioned are correct and they can learn from those kind of articles. This written text excavated papers may be utilized in clinical hospital treatment area wherein a medical doctor can analyze numerous styles of treatment that can be given to person with precise health-care trouble. the experimental methodology whilst the primary setting up is used for the second method, to use greater resources as representation techniques, and to concentrate extra on strategies to combine the analysis findings in a structure to be deployed to customers.

7. FUTURE WORK

The current system is based on the local stand alone system. But we cannot restrict this tool for global access over the globe. We can convert the same application as a global tool for the sharing the access to all the researchers over the globe. So that we can get the more research opportunities and we can get more research documents and articles from the PubMed and also from MEDLINE.

REFERENCES

[1] Jerr D. M. Rennie, Lawrence Shih, Jaime Teevan, Bob R. Karger, "Tackling The POOR Supposition Of Naïve Bayes Text Classifier", Procedures Of The 20th Worldwide Conference On Device Studying (ICML-2003), California DC, 2003.

[2] T.Mouratis, S.Kotsiantis, "Increasing The Precision Of Discriminative Of Multi-nominal Bayesian Classifier In Text Classification", ICCIT'09 Procedures Of This year's Fourth International Conference On Pc Technological innovation And Convergence Information Technological innovation.

[3] B.Rosario And M.A.Hearst, "Semantic Regards In Bioscience Text", Proc. 42nd Ann. Conference On Assoc For Computational Linguistics, Vol.430,2004.

[4] M.Craven, "Learning To Draw out Interaction From Medline", Proc. Assoc. For The Progression Of Synthetic Intellect.

[5] Oana Frunza.et.al, "A Device Studying Strategy For Identifying Disease-Treatment Interaction In Brief Texts", May 2011

[6] L. Seeker And K.B. Cohen, "Biomedical Language Processing: What's Beyond Pubmed?" Molecular Cellular, Vol. 21-5, Pp. 589-594,2006.

[7] Mark Pasternack, Don Roth "Extracting Content Written text From Webb With Highest possible Subsequence Segmentation", WWW 2009 MADRID.

[8] Abdur Rehman, Haroon.A.Babri, Mehreen saeed," Feature Extraction Criteria For Category Of Written text Document", ICCIT 2012.

[9] Adrian Canedo-Rodriguez, Jung Hyoun Kim,etl.,"Efficient Text Extraction Algorithm Using Shade Clustering For Language Translation In Cellular Phone" , May 2012.

[10] Oana Frunza, Diana Inkpen, and Johnson Tran, Participant, IEEE "A Machine Studying Strategy for Determining Disease-Treatment Interaction in Brief Texts" IEEE dealings on knowledge and details technological innovation, vol. 23, no. 6, july 2011.