THINK INDIA JOURNAL

ISSN:0971-1260

Vol-22-Issue-10-November-2019

Implementation Of Stemming Algorithms In Healthcare With Special Reference To Viral Infective Diseases

Dr. Ashwini Manish Brahme [1]
Sinhgad Institute of Managemnet and Computer Application (SIMCA), Pune
Savitribai Phule Pune University Pune, India
ashwiniak47@gmail.com , ashwini_kulkarni21@gmail.com

Dr. Shivaji D. Mundhe[2]
Director, Yashaswi Education Society's International Institute of Management Science, Pune
Savitribai Phule Pune University, Pune, India
drshivaji.mundhe@gmail.com

THINK INDIA JOURNAL

ISSN:0971-1260

Vol-22-Issue-10-November-2019

Abstract— The data available on the internet and World Wide Web is very huge and vast only 10% of data is in structured form and approximate 90% is either unstructured or semi-structured. Data mining is only feasible for structured data and not for unstructured and semi-structured. The paper entitles towards the text mining phases such as text transformation, text pre-processing, filtration and stemming. The paper also aimed towards the high frequency viral infective diseases textual online news from various newspapers and processing it for better information retrieval. The research is aligned on various stemming techniques and theircomparison.

Keywords—Text mining, Stemming, Viral Diseases, Data Mining, Text Pre-Processing.

I. INTRODUCTION

The ample amount of available information and data is stored on the internet is in the text, images, video, audio, graph, email, blogs, comments, etc. formats. To work with a such as data text mining is useful as there are diverse applications of text mining in the field of healthcare, publishing and media, telecommunications, energy, Information technology sector, Internet, banks, insurance and financial markets, public administration and legal documents, political institutions, political analysts, pharmaceutical research companies and manymore.

Text mining is one of the biggest challenges in healthcare sector too. There are various challenges of healthcare text mining specifically an Intermediate form of medical data, multilingual text refining, domain knowledge integration, large dimension of healthcare data, the complexity of natural language processing, ambiguity and context sensitivity in data, discover the relationship between the medical terms and concepts. The ample amount of scientific and medical information is in the unstructured and semi-structured form in some cases longitudinal data is available but there is need to generate meaningful results, sizable portion of mined data and proper knowledge in all these cases text mining plays a vital role. For healthcare system, there is need to learn what type of data to be collected, how to collect it and how to analyze it effectively. Hence, there is a necessity to select text data mining and analytics for better prediction and knowledge discovery.

II. TEXTMINING

To perform the text mining of diseases news and information the following steps are carried out.

- TextTransformation
- TextPre-Processing
- Apply data mining processedtext
- Visualization and KnowledgeDiscovery

A. Text Transformation: This data is in textual form that is in an unstructured form; to work with the unstructured data there is need to convert an unstructured information into a structured format. The text transformation phase converts an unstructured data into a structured form which is beneficial to achieve better performance for information retrieval and further processing

B. Text pre-processing: The text pre-processing is the first steps of text mining and it plays a significant role intertwining. The document is converted into XML form so that the pre-processing becomes easy on the document. The text pre-processing contains filtration, tokenization, stop word removal and stemming. The complete textual data is converted into small characters and all the special characters from the text as @! #$%^7*()/<? etc. are removed.

Tokenization: Tokenization is the process of splitting the text into words or terms. The main objective is to convert the unstructured document into a structured format. Normally, the web page contains information in an unstructured/semi-

ISSN:0971-1260

Vol-22-Issue-10-November-2019

structured form which creates a problem for accurate and relevant text mining; therefore, there is need to convert the unstructured diseases data, online news and information into the structured form to perform the better information retrieval. This phase is very important for association rule generation also. The diseases news and data are fragmented into the words, characters, phrases, symbols and so forth which are used for further processing. The some of the sample tokens are signified in the following figureno.1.

Figure 1: Tokenization of Diseases News

id	type	filename	transaction	token	dt
1	SwineFlu	7-yr-old-dies-swine-flu-taking-toll-pune-81-3821	1	pune	2017-11-12 15:4
2	SwineFlu	7-yr-old-dies-swine-flu-taking-toll-pune-81-3821	1	a	2017-11-12 15:4
.....	.....	.....	.....	...	.....
15418	Dengue	61098492.cms	44	vip	2017-11-12 15:5
15419	Dengue	61098492.cms	44	homes	2017-11-12 15:5
15420	Dengue	61098492.cms	44	additional	2017-11-12 15:5
.....	.....	.....	.....	.....	.....
60777	HIV	61478954.cms	167	in	2017-11-12 16:1
60778	HIV	61478954.cms	167	the	2017-11-12 16:1
60779	HIV	61478954.cms	167	medical	2017-11-12 16:1
60780	HIV	61478954.cms	167	college	2017-11-12 16:1

Filtration and Stop Word removal: Filtering the web documents is an essential segment to remove unimportant (irrelevant) keywords from the document such as stop words. Stop words do not carry any meaningful information. Therefore, the stop words are discarded and more important and highly relevant words/tokens are used in the present study for stemming and association rule mining. Examples of stop words: a, an, the, where, why, so, after, before, of, no, yes, true, false, therefore, hence, he, she, it, they, whom, how, to, up, down,etc.

The stop words are removed from the disease's news and information. The more significant and highly relevant words/tokens are filtered out from the document for further processing. Therefore, it is easy to work with significant words for better informational retrieval, stemming, association rule mining and knowledgediscovery.

Table 1: Tokenization and Stop word removal

Sr. No.	Particulars	Count
1	Number of tokens Generated	60,800
2	No. of Stop Words Removed	28,595
3	No. of Single Character Words	1685
4	Total Tokens remained for further Processing	30,520

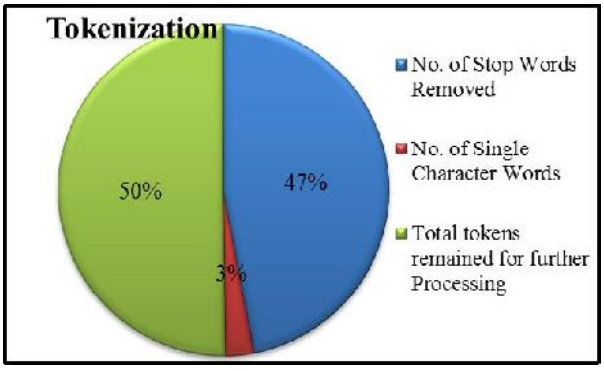
Source: Compiled by researcher

Ithasresultedthat,47.03%tokensarestopwordsor least important words which do not play any role in information retrieval and further processing, 2.77 % tokens are single letter word. Therefore, the total 49.80% tokens are least significant and not considered for further processing. It has also depicted from the subsequent graph no.1 that approximate 50 % tokens are least significant and 50% tokens are significant which plays a key role and hence these 50% (30,520) tokens are used for stemming, association rule generation and knowledge discovery in the presentstudy.

Hence, it has concluded that there is need to filter out the textual data by removing the special characters, stop words, least significant words to reduce the space, the processing time to achieve the better information retrieval, stemming, association rule mining and knowledge discovery.

ISSN:0971-1260
Vol-22-Issue-10-November-2019

Graph 1: Filtered tokens and stop word removal



III. STEMMING

Stemming is very important in indexing and searching which is a key point of text mining, natural language processing, information retrieval, text categorization, summarization and clustering as a part of pre-processing before applying to any algorithm. The morphological deviations of words have a similar semantic interpretation; since the meaning is same but the word form is different it is necessary to identify each word form with its base form. To get the proper stems various stemming algorithms are evolved to convert the morphological variants of a word like an introduction, introducing, introduces, etc. to get mapped to the word 'introduce'. Some algorithms may map them to just 'introduc', 'introd', 'introdu', etc. but that is allowed if all of them map to the same word form or more popularly known as the stem form. The main objective is to reduce the various terms used to reduce the processing time.(1)

The stemming algorithms are broadly classified into three types namely truncating, statistical and mixed. The truncating technique contains Lovins, Porters, Paice/ Husk and Dawson algorithms which works on suffix and affix removal. The statistical stemming contains N-Gram, HMM and YASS; and the third type mixed techniques includes Korvetz, Xerox, Corpus-Based and Context Sensitive stemming algorithm.

IMPLEMENATION OF STEMMING TECHNIQUES ONVIRAL DISEASES ONLINENEWS

The present study aimed at the selected truncating and mixed stemming technique which aims to remove prefixes and suffixes are also known as affix removal method. The present study is focused towards the implementation of Porters,Lovins, Paice/Husk and Krovetzstemming for selected viral infective diseases online news and information explicitly Dengue, Chikungunya, HIV, Influenza Flu, Swine Flu, Diarrhea andHIV.

A. Porters Stemming: The Martin Porters had devised this algorithm in 1980; it encompasses 5 steps and 60 rules to remove the suffixes from the word and convert the word to its original stem/root. This is most widely used stemming technique used by many researchers in various languages for stemming and information retrieval.

B. Lovins Stemming: The Lovins had designed suffix stripping stemming algorithm in 1968. The main objective of it is to remove the longest suffix from the word. It has 294 endings, 29 conditions and 35 transformationrules.

C. Paice/ Husk Stemming: This stemming is developed by Charis Paice and Gareth Husk in 1980; It contains 120 rules and it is a simple and iterative algorithm. It is a conflation based iterative stemmer, removes and replaces the endings. This algorithm is very complex, and the main pitfall is over stemming is occurs due tothis.

D. KrovetzStemming Algorithm: The Krovetzalgorithm was developed in 1993 by Robert Krovetzfor removing the inflectional suffixes. It has three main steps and an unknown number of rules. It is complicated and not implemented other than the

ISSN:0971-1260
Vol-22-Issue-10-November-2019

Englishlanguage.

IV. COMARISON OF CORRECT AND INCORRECTSTEMMING There are 30520 significant tokens available after filtration and stop word removal. The selected stemming techniques specifically Porters, Lovins, Paice/Husk and Krovetzare applied on these 30520 tokens. Each technique givesvarious results while generating properstem.

Table 2: Correct and incorrect stemming

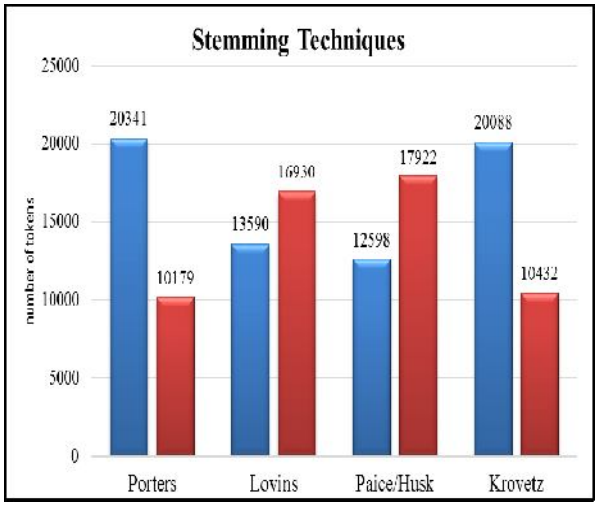
Particulars	Stemming Technique			
	Porters	Lovins	Paice/Husk	Korvetz
Number of Correct Stems found (out of 30520)	20341	13590	12598	20088
Number of Incorrect stems found (out of 30520)	10179	16930	17922	10432
Percentage of Correct Stemming	66.65	44.53	41.28	65.82
Percentage of Incorrect Stemming	33.35	55.47	58.72	34.18

Source: Compiled by researcher

The above table no. 2. and graph no 2 it has revealed that the Porters stemming results 66.65 % correct stemsand 33.35 % incorrect stemming, Lovins outcomes 44.53 % correct and 55.47 % incorrect stems, Paice/Husk gives41.28 % correct and 58.72 % incorrect stems while as Korvetz outcomes 65.82% correct stemming and 34.18 % incorrect stemming out of 30520 tokens. It has observed that Porters stemming gives more accurate and correct stems as compare to Korvetz, Lovins, and Paice/Husk sequentially.

Therefore, it has concluded that the Porters stemming results better stemming than Korvetz, Lovins and Paice/Husk stemming algorithms consequently.

Graph no. 2: Correct and Incorrect Stemming of Porters, Lovins, Paice/Husk and Korvetz Techniques



V. LIMITATIONS OF PORTERS STEMMING ALGORITHM The above result and comparison represent that out of all the selected stemming algorithms/techniques the Porters stemming performs better results for identifying the proper stems. Most of the researcher also done a lot of research on various stemming algorithms and tried to improve the techniques. But still there are some problems with the porters stemming briefed asfollows:

- It has five steps and 60 rules for suffixstripping
- There are main two problems in porters stemming one is over stemming and another is understemming.
- Stemming generates the stem by applying a set of rules without bothering about the Part of Speech(POS).
- Most of the time it does not gives proper and correct stems or rootwords.
- Because of improper stems, it results ininsig
- Similarly, improper stems differ in the meaning of words also.

VI. FUTUERSCOPE

ISSN:0971-1260

Vol-22-Issue-10-November-2019

To overcome above mentioned demerits of porters stemming algorithm there is necessity and scope to improve the porters stemming and come with the innovative and effective approach so that better stemming will be performed. Getting correct stem through the textual information is very important as well as the good stemming results in better information retrieval, association rule mining, knowledge discovery and further processing. Therefore, the researcher had aimed towards proposing/ devising a new algorithm of stemming to overcome the limitations of existing porters stemming algorithm.

CONCLUSION

There is an enormous amount of viral infective diseases data and information available on the internet; but it is in an unstructured and semi-structured format such as text, emails, images, graphs, audio, videos, and blogs and so on. The study was carried out on text mining of online news and information of selected viral infective diseases. The mined text is filtered by removing stop words and the significant tokens are considered for stemming. The Porters, Paice/Husk, Lovins and Korvetz suffix stripping algorithms were implemented on the filtered tokens and correct and incorrect stems are generated. Out of these Porters stemming results better stemming than the others. But still to achieve more accuracy and correct stemming there is need to improve the exiting stemming techniques and algorithms which rises the new avenue for the researchers and for better information retrieval process.

ACKNOWLEDGMENT

The authors would like to express her sincere thanks to the Dr. Shivaji D. Mundhefor their support, excellent knowledge and guidance for the research work presented in thispaper.

REFERENCES

1. Jivani A. (2011). A Comparative Study of Stemming Algorithms. Int. J. Comp. Tech. Appl. (IJCTA), 2 (6), 1930-1938,www.ijcta.com

2. S. Giridhar, V. Prema, Reddy S. (2011). Giridhar N S., Prema K.V., N. V SubbaReddy (2011). A Prospective Study of Stemming Algorithms for Web Text Mining. Ganpat University Journal of Engineering & Technology,1(1),28-34

3. Sharma D. (2012). Stemming Algorithms: A Comparative Study and their Analysis. International Journal of Applied Information Systems (IJAIS), Foundation of Computer Science FCS, New York, USA, 4(3), 7-12,www.ijais.org

4. N. Sandhya, Y. Sri Lalitha, Sowmya, K. Anuradha, A. Govardhan (2011). Analysis of Stemming Algorithm for Text Clustering. IJCSI International Journal of Computer Science, 8(5), 352-359,www.IJCSI.org

5. Ramasubramanian C., Ramya R. (2013). Effective Pre-Processing Activities in Text Mining using Improved Porter's Stemming Algorithm. International Journal of Advanced Research in Computer and Communication Engineering, 2(12), 4536-4538, www.ijarcce.com

6. S. Vijayarani., J. Ilamathi, Nithya. Preprocessing Techniques for Text Mining - An Overview, International Journal of Computer Science & Communication Networks, 5(1),7-16

7. Kulkarni M., Kulkarni S. (2016). Knowledge Discovery in Text Mining using Association Rule Extraction, International Journal of Computer Applications (0975 – 8887), 143(12),30-35

8. M.F. Porter, 1980, An algorithm for suffix

ISSN:0971-1260

Vol-22-Issue-10-November-2019

stripping, Program, 14(3) pp130−137

9. Kulkarni A., Mundhe S. (2018). Implementation of Effective Stemming Algorithm of Text Mining for Knowledge Discovery in Healthcare. Savitribai Phule PuneUniversity

10. http://snowball.tartarus.org/algorithms/porter/stemmer.html

11. http://www.punediary.com/html/press.html

12. https://www.sccollege.edu/Library/Pages/primarysources.aspx

13. https://www.practo.com/pune, (Accessed this website on 12/11/2016)

14. http://timesofindia.indiatimes.com/archive.cms

15. http://www.news-medical.net/

16. http://www.punecorporation.org/en/health-department-3

17. http://imapune.org/, Indian Medical Association PuneBranch