**International Journal of Management, IT & Engineering**
Vol.7 Issue 11, November2017,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

# AN APPLICATION OF PORTERS STEMMING ALGORITHM FOR TEXT MINING IN HEALTHCARE

**Ashwini Rajendra Kulkarni***

**Dr. Shivaji D. Mundhe****

**Abstract**

Text mining has diverse applications in variety of fields where manual analysis and generating effective knowledge discovery from information is not possible because of huge availability of information on website. There is 90% of web data is in unstructured and semi structured form and only 10 % is in structured format. The data analysis and decision making from unstructured and semi structured data arises biggest challenge and research opportunities.

The paper is focused towards text mining and text pre-processing aspects. The paper is briefed about the need of stemming for healthcare, stemming of various viral infective diseases information. Also the paper elaborated on comparison of various stemmers. This paper is focused on implementation of porters stemming algorithm for viral diseases information and textual data. The output and results are disused along with further opportunities in stemming technique improvements.



***Keywords:***

Text Mining;
Stemming;
Data Mining;
Algorithm;
Diseases.

***Author correspondence:***

Ashwini Rajendra Kulkarni ,
Research Scholar, Sinhgad Institute of Management And Computer Application, Pune
Savitribai Phule Pune University
Email: ashwiniak47@gmail.com

## 1. INTRODUCTION

Text mining is used to discover hidden, useful, and interesting patterns from unstructured text documents. Text mining plays vital role in numerous applications such as There are various applications of text mining like telecommunication, bank, IT, media, insurance, political analysis, pharmaceutical, health care, bioinformatics, business intelligence, national security, etc. To perform effective and efficient text mining there is need to go through the various steps like text pre-processing, text transformation, feature selection

*Research Scholar, Sinhgad Institute of Management And Computer Application(SIMCA), Pune

**Director-MCA, Sinhgad Institute of Management And Computer Application(SIMCA), Pune

and applying data mining technique on it. Text pre-processing includes tokenization, stop word removal and stemming. Stemming is important for indexing and searching and highly beneficial to reduce the database size and improve the information retrieval effectiveness; also it is beneficial for text categorization, text classification, text clustering, etc. There are mainly three types of stemming truncating stemming, Statistical stemming and Mixed Stemming which is broadly classified among various algorithms and techniques detailed in below

**2.LITERATURE REVIEW**

The researcher Y. Jahnavi and Y. Radhika (2012) discussed on preprocessing, term weighting algorithms, concept based term weighting algorithms, pattern discovery, categorization, domain ontology based frame work for text mining and summarization techniques in the paper entitled "A Cogitate Study on Text Mining". The researcher had studied on Porters stemming algorithm which is implemented by using suffix list and it is speedy small and easy; but it may not work for all the words. [1]

An author Jivani A. (2011) explained the stemming algorithms in her research paper entitled "A Comparative Study of Stemming Algorithms". The truncating stemming algorithms are Lovins, Porters, Paice/Husk, and Dawson; statistical contains N-Gram, HMM, YASS; and mixed stemming are Krovetz, Xerox, Corpus based, Content Sensitive. The researcher interpreted that out of all these stemming Porters Stemming is produces best output as compared to the other and which is having less error rate; but the main problem with the Porters stemming is the stems produced are not always real words/ root words. [4]

A research paper entitled "A Prospective Study of Stemming Algorithms for Web Text Mining", S. Giridhar, V., et.al. (2011) elaborated on various stemming techniques and algorithms which are broadly classified into brute force algorithm and affix (suffix/prefix) stripping algorithm. The affix algorithm contains a list of rules in algorithm to find out the stems or root form. The researcher had focused on affix stemming algorithms which are useful for web text mining such as Paice/Husk Stemming , Porter Stemming, Korvetz stemming and Dawson Stemming. This paper also guides about how the stemming will be beneficial in mining to reduce the database size, the stemming algorithms differs in performance and accuracy. [2]

Sharma D. (2012) compared and analyzed the stemming algorithms and recommended the Porter's stemming is best than the other stemmers in a research paper entitled "Stemming Algorithms: A Comparative Study and their Analysis". This paper also discussed about the rule based and statistical approach of stemming which is useful for language specific and statistical information consecutively. Furthermore an author had focused on advantages and disadvantages of stemming and portray on the problem of size of stemming; for smaller sample size stemming is performed faster and for larger samples stemming takes long time. [3]

**3. VIRAL INFECTIVE DISEASES**

There are several infectious diseases have increased in incidence and expanded into new geographic areas. There are multiple factors that contribute to the spread of disease, including increasing urban population density, more international travel, and widespread international import/export of goods. There are several viral infective diseases such as Dengue, Chikungunya, Influenza Fli, Malaria, Rabies, Swine Flu, etc. Out of all these diseases, in the past decade, India has shown an increasing trend in the number of reported dengue cases, which have steadily increased, in 2007 it was 5534 and in 2013 it was 75,454; according to the National Vector Borne Disease Control Programme (NVBDCP). There are 35 states of the country out of which 18 are

now being considered endemic for dengue and the spread of the disease from urban to suburban and rural areas, the actual number of cases may count in millions. World Health Organization (WHO) estimates that 50-100 million dengue infections occur worldwide annually. Explosive dengue epidemics are being reported every year from more than 100 endemic countries spanning South-East Asia, Western Pacific, Africa, the Americas and the East Mediterranean.

## 4. TEXT MINING OF DENGUE DISEASES

The news of dengue diseases are taken from online English newspaper. To work on the textual data there is need to perform text mining and for the same the text mining steps are carried out. It contains text preprocessing where three main tasks are carried out:

- Tokenization: the unstructured data is converted into structured form. Simply the text is split into the words.
- Stop word removal: The least important words are removed from the text.
- Stemming: after stop word removal the list of words are available, for this stemming is performed to transform these words into its original or root word form.

## 5. PORTERS STEMMING ALGORITHM

In 1980 , Martin Porter designed a stemming algorithm for English Language. This stemming algorithms plays vital role for information retrieval and finding out the stems or root words from textual data. This stemmer has 5 steps and a set of 60 rules, which are applied to remove the suffixes from word until none of the rule will apply. The Porter stemmer applies a set of rules. It is widely used stemmer for truncation of words. The algorithm is as follows:

1. For each document Di in the document set
   For each word Wj in the document
   If (not stop-word)
   Stem Wj using Porter Stemmer
   Increment TF[i, j]

2. Build the Document Frequency vector using the Term Frequency matrix
   DF[j] = Total values of i for which TF[i, j] is non zero.

3. Build Inverse Document Frequency vector
   IDF[j] = log(D/DF[j]) , D is total number of documents

4. Normalize the Term Frequency matrix
   $TF[i, j] = TF[i, j] / \max(TF[i, j])_{j=1..N}$ , N is the total terms

5. Multiply Term Frequency & Document Frequency for each term
   TF-IDF[i, j] = TF[i, j] x IDF[j]

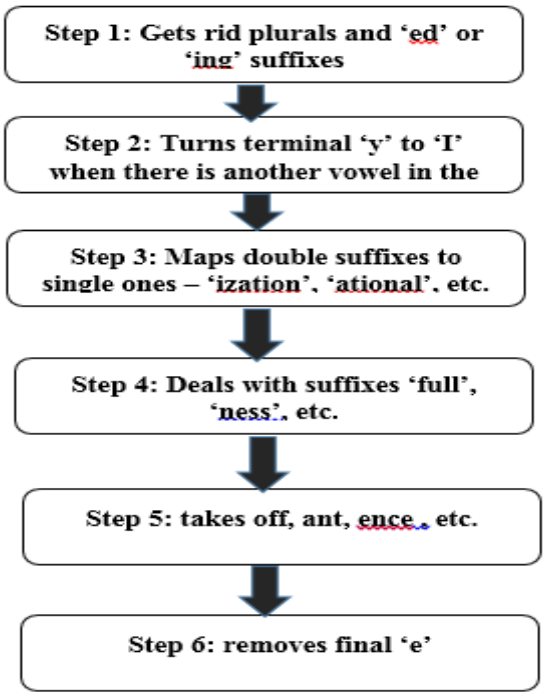


**Diagram: Porters Stemming Algorithm Flowchart**

## 6. RESULT AND DISCUSSION

The porters stemming algorithm is applied on set of 500 words/ indexes and the following is the list of root words or stems which are identified through this algorithm.

| Sr. no | Root Word | Porters Stemming | Root Word (Validity) |
|---|---|---|---|
| 1 | pune | pune | TRUE |
| 2 | pimpri | pimpri | TRUE |
| ……. | | | |
| 9 | **boy** | **boi** | **FALSE** |
| 10 | ahmednagar | ahmednagar | TRUE |
| 11 | lost | lost | TRUE |
| 12 | **battle** | **battl** | **FALSE** |
| 13 | swine | swine | TRUE |
| 14 | flu | flu | TRUE |
| 15 | hour | hour | TRUE |
| 16 | young | young | TRUE |
| 17 | girl | girl | TRUE |
| 18 | **declare** | **declar** | **FALSE** |
| 19 | dead | dead | TRUE |
| 20 | **friday** | **fridai** | **FALSE** |

| Sr. no | Root Word | Porters Stemming | Root Word (Validity) |
|---|---|---|---|
| …….. | | | |
| 486 | **office** | **offici** | **FALSE** |
| 487 | share | share | TRUE |
| 488 | detail | detail | TRUE |
| 489 | swine | swine | TRUE |
| 490 | flu | flu | TRUE |
| 491 | ailment | ailment | TRUE |
| ……. | | | |
| 499 | **city** | **citi** | **FALSE** |
| 500 | **thursday** | **thursdai** | **FALSE** |

**Table 1: Implementation of porters stemming**

The root words generated through the porters stemming are compared with the original root words and in the set of 500 indexes there are 72% proper root words and approximate 28% stems are not original root words. For the text preprocessing, indexing and further processing on stems there is need to come up with the effective stemming technique.

| Check Validity of root words | Stemming performed through Porters Stemming |
|---|---|
| TRUE | 360 |
| FALSE | 140 |
| Total | 500 |
| Per (%) | 71.85629 |

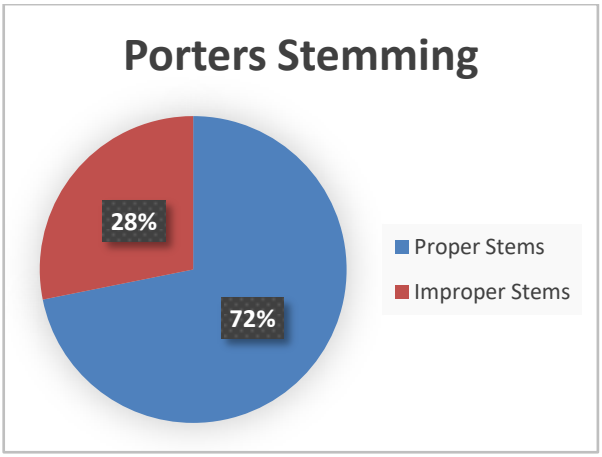


**Diagram: Porters stemming performed on words**

## 7. CONCLUSION

To achieving the correct text mining results there is need to use proper stemming technique.There are various stemming techniques available, through the literature review and previous research it was found that among all the stemming techniques porters stemming gives better results. But still porter is also having some pitfalls while identifying the root word in terms of under stemming and over stemming, also most of the time it does not give the original words or stems. Most of the researcher has worked on porters stemming and came up with the advanced version of porter but still it has exists. Therefore, there are research opportunity for modifying the porter's algorithm to get the maximum number of root words which is beneficial for decision making, association rule generation and further processing.

## 8. REFERENCES

[1]. Y. Jahnavi, Y. Radhika. (2012). A Cogitate Study on Text Mining. *International Journal of Engineering and Advanced Technology (IJEAT)* 1(6), 189 – 196

[2]. S .Giridhar, V. Prema, Reddy Subba. (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]. Jivani A. (2011). A Comparative Study of Stemming Algorithms. *Int. J. Comp. Tech. Appl. (IJCTA),* 2 (6), 1930-1938, www.ijcta.com

[5]. Kodimala S., (2010). Study of stemming algorithms. *UNLV Theses, Dissertations, Professional Papers, and Capstones.* 754. http://digitalscholarship.unlv.edu/thesesdissertations/

[6]. Kulkarni A., Mundhe S. (2016).A Theoretical Review on Text Mining: Tools, Techniques, Applications and Future Challenges. *International Journal of Innovative Research in Computer and Communication Engineering*, 4(11)

[7]. Brahme A., Mundhe S. (2015).A Conceptual Study of Knowledge Discovery Using Text Mining and Its Applications. *International Journal of Management, IT & Engineering*, 5(9) , page no. 174-181

[8]. Raja U., Mitchell T. Day T., Hardin J. TEXT MINING IN HEALTHCARE: APPLICATIONS AND OPPORTUNITIES. Retrieved from Research Gate, https://www.researchgate.net/publication/24182770_Text_mining_in_healthcare_Applications_and_opportunities

[9]. Moral. C., de Antonio, A., Imbert, R. & Ramirez, J. (2014). A survey of stemming algorithms in information retrieval Information Research, 19(1) paper 605

[10]. Feldman R., Sanger J. (2007).The Text Mining Handbook Advanced Approaches in

[11]. Analyzing Unstructured Data, *Cambridge University Press* www.cambridge.org/9780521836579

[12]. Chakraborty G., Pagolu M. , Garla S.(2013) Text mining and analysis Practical methods, examples and case studies using SAS, *SAS Institute Inc. Cary North Carolina, USA*

[13]. Frakes W., Baeza-Yates R. (2004), CHAPTER 8: STEMMING ALGORITHMS, Information Retrieval: Data Structures & Algorithms, (Retrieved from: http://dns.uls.cl/~ej/daa_08/Algoritmos/books/book5/chap08.htm)

[14]. S.Vijayarani, et.al. , Preprocessing Techniques for Text Mining - An Overview .*International Journal of Computer Science & Communication Networks*,5(1),7-16, http://www.ijcscn.com/Documents/Volumes/vol5issue1/ijcscn2015050102.pdf

[15]. M.F. Porter, 1980. An algorithm for suffix stripping, Program, 14(3) pp 130−137 http://www.cs.odu.edu/~jbollen/IR04/readings/readings5.pdf

[16]. Atharva Joshi et al, Modified Porters Stemming Algorithm. *(IJCSIT)International Journal of Computer Science and Information Technologies*, Vol. 7 (1) , 2016, 266-269, www.ijcsit.com