A Rule Based Stemmer

R. Cynthia Monica Priya, J.G.R. Sathiaseelan

Abstract*: The present digital world generates enormous amount of data instantaneously. The need to effectively mine knowledge seems to be the need of the hour. Sentiment Analysis, a part of web content mining which is a subpart of web mining has gained momentum in the field of research. It analyses the opinion of variety of people all over the world. Sentiment Analysis encompasses preprocessing, feature selection, classification and sentiment prediction. Preprocessing is an important process and it deals with many techniques. Stop word removal, punctuation removal, conversion of numbers to number names are some of the basic techniques. Stemming is yet another important preprocessing technique that reduces the different words form to its root. There are basically three types of stemmers namely truncating, statistical and hybrid. The aim of this paper is to propose a rule based stemmer that is a truncating stemmer. It deals with rules for truncation and replacement. The data given as input passes through a series of rules. If the condition specified gets satisfied then the associated rule gets executed otherwise the input is checked with the next rule and the process continues further. The result of execution is stemmed words. The performance of the proposed rule based stemmer is compared with the existing stemmers under the same rule based category namely Porter and Lancaster. Various metrics have been used for evaluation. The observations reveal the fact that the proposed stemmer out performs the Porter and Lancaster stemmers in terms of correctly stemmed words factor and shows a good average conflation factor and lesser over stemming and under stemming errors.

Keywords: metrics, preprocessing, stemmer, stemming

I. INTRODUCTION

The ever growing rate of data generation has created an emergent need to collect the available data and extract knowledge from it. In today's busy world people use the online platform to meet out almost all their needs [1]. The usage of internet has become the part and parcel of life. Online buying and selling is at its peak. This scenario is kindling the spirits of people to look at the opinion of others while making purchases. Plenty of online reviews are available. The product reviews are helpful in creating an understanding about the various attributes of a product. The availability of online reviews alone does not serve the needs of people. There arises a need to analyze the available reviews and provide graded ratings. Sentiment analysis is a powerful tool that helps in analyzing the opinions of people. The step after data collection is data preprocessing. The collected data is preprocessed through methods like stop word removal, punctuation removal, and then fed into the proposed stemming algorithm. The stemming process reduces the various grammatical forms of a word like its adjective, noun, verb, adverb to its base form [2].

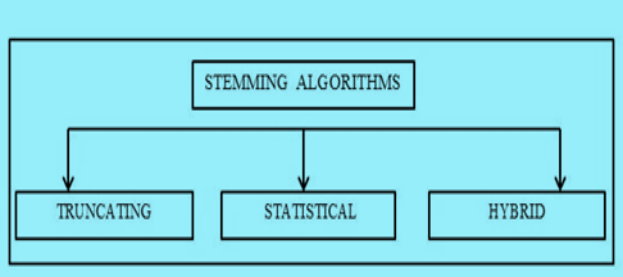

Fig.1. Types of stemming algorithms

Three categories of stemmers are available. They are truncating stemmers, statistical and hybrid stemmers. Porter, Lovin, Lancaster and Dawson stemmer are some of the truncating stemmers. The statistical stemmers include Hidden Markov Model Stemmer, n-gram stemmer and Yet Another Suffix Stripper [3]. This paper aims to propose a rule based stemming algorithm that falls under the truncating stemmer category and analyze its performance with the existing stemmers namely Porter and Lancaster.

II. LITERATURE REVIEW

Hussein et.al [4] brought to light that 85% of the total time in data analysis is dedicated to preprocessing. Three algorithms for tokenization, cleaning and stop word removal have been proposed. Porter's algorithm has been used for stemming. The execution time for each step has been calculated by changing the data size .It has been concluded that the execution time increases as the data set is increased. Vijayarani et.al [5] provided a detailed analysis on the various steps in preprocessing namely stop word removal and stemming. The need of preprocessing the collected data is justified. The stop word elimination methods namely classic method, Z-Method, the mutual information method, term based random sampling have been surveyed. A detailed study on the various types of stemmers namely truncating, statistical stemmers, mixed stemmers have been made. Jasmeet et.al [6] analyzed the various types of stemmers available for Indian languages and their applications have been discussed in detail.

Brajendra et.al [7] studied the types of stemming errors available in both rule-based and statistical category.Ruba Rani et. al [8] provided a detailed evaluative report of the various stemming algorithms by comparing four affix removal and three statistical stemming algorithms and discussed their limitations. Giulio et.al [9] compared the various preprocessing steps. They have evaluated the performance enhancement with a basic cleaning algorithm and a process namely stemming, stopword removal, negation, emoticon and a dictionary.

Revised Manuscript Received on October 15, 2019

R.Cynthia Monica Priya, Department of Computer Science, Bishop Heber College , Tiruchirapalli, India. cynmonpri@gmail.com

Dr.J.G.R. Sathiaseelan, Department of Computer Science, Bishop Heber College , Tiruchirapalli, India. jgrsathiaseelan@gmail.com

The results produced, reveal the fact that stemming improves accuracy. Akrivi et.al [10] stated that data preprocessing is an important step in sentiment analysis .Three different datasets have been used for analysis on four different classifiers. The results show that pre-processing operations greatly influence the quality of classification. Sundar Singh et.al [11] stated that stemming improves the performance of information retrieval systems. Porter stemmer algorithm that follows iterative approach has been considered to be the most widely used one. The authors have reported five errors and provided solutions for the same.

Atharva et.al [12] studied the various stemming algorithms on three broad categories namely truncation stemmers, statistical stemmers and inflectional/derivational stemmers. The drawbacks of Porter's algorithm have been stated and few improvements have been suggested and carried out. The results have shown an increase in efficiency. Ramalingam et.al[13] enhanced the Porter stemmer algorithm by applying certain rules. Add rule and replace rules have been proposed and the two metrics namely precision and recall values have been calculated. The results showed an increased precision.

Mubashir Ali et.al [14] proposed a rule based stemmer for Urdu language. Infix stripping rules have been developed to cope up with the infix stemming in addition to the basic prefix and postfix techniques. The longest match rule is applied in situations where two rules are applicable for the same text. Abdul et.al [15] suggested an integrated approach for Urdu stemming by combining the affix stripping, template matching and table lookup techniques. The observations made showed that the proposed algorithm has higher accuracy and compression rate.

Hunaida et.al [16] proposed a hybrid approach by combining a root based and light stemmer for finding the sentiment words in Arabic language. The results obtained showed higher performance and accuracy. Durairaj et.al [17] developed a modified Porter stemmer algorithm and evaluated using the metrics words stemmed factor and the correctly stemmed words factor. The observation revealed that the proposed stemmer performed better than Dawson and Lovin. Priya et.al [18] proposed a rule based stemmer and analyzed its performance and concluded that the proposed stemmer produced less under stemming and over stemming errors. Vairaprakash et.al [19] made a performance analysis for three of the common stemming algorithms namely Porter, Lovin and Paice/Husk. The different metrics for evaluating accuracy and performance have been considered. The results revealed Paice/Husk to be the stronger than Lovin and Porter the weakest. Kasthuri et.al [20] developed a language independent stemmer and reported that the accuracy and speed of the proposed stemmer was higher.

III. METRICS USED FOR EVALUATION

The developed stemmer needs to be gauged to prove its worth. The performance of the stemmer is gauged through various metrics like word stemmed factor, mean number of words, index compression factor, correctly stemmed words factor and average conflation factor [18, 19, 20].

A. Index Compression Factor(I) is the percentage of the ratio between the difference in the number of distinct words before stemming and the number of distinct stems after stemming to the number of distinct stems after stemming.

I=(BS-AS)/BS*100

where,

BS denote the number of distinct stems before stemming

AS denote the number of distinct stems after stemming

B. Word Stemmed factor (W) is the percentage of the ratio of stemmed words to the total word count.

W=WS/WC*100

where,

WS denotes the total number of words stemmed

WC denotes the total number of words in the sample

C. Mean Number of words (M) is the ratio between the number of distinct words before stemming and the number of distinct words after stemming.

M=BS/AS

D. Correctly Stemmed Words Factor(C) is the percentage of the ratio count of the correctly stemmed words to the total number of stemmed words.

C=CS/WS*100

where,

CS denotes the total number of correctly stemmed words

WS denotes the total number of words stemmed

E. Average Words Conflation Factor(A) is the percentage of the ratio of the difference between the number of correctly stemmed words and the number of distinct words after conflation to the number of correctly stemmed words.

To find N, we have to calculate the difference between the number of distinct words after stemming and the number of correct words not stemmed.

N=AS-CW

Where CW is the number of correct words not stemmed

Then,

A=(CS-N)/CS*100

IV. PROPOSED METHODOLOGY

The online reviews collected from the web are passed through the noise remover that involves cleaning of stop words, ASCII, punctuation, lower case conversion and number name conversion. The partially preprocessed data is passed through a tokenizer and the tokens thus obtained are fed to the suffix analyzer. The analyzer deals with two categories of rules, namely truncation rules and truncation and replacement rules. Truncation rules deal with trimming alone whereas truncation and replacement rules aim at trimming along with substitution. The output of the analyzer is a list of stemmed words.

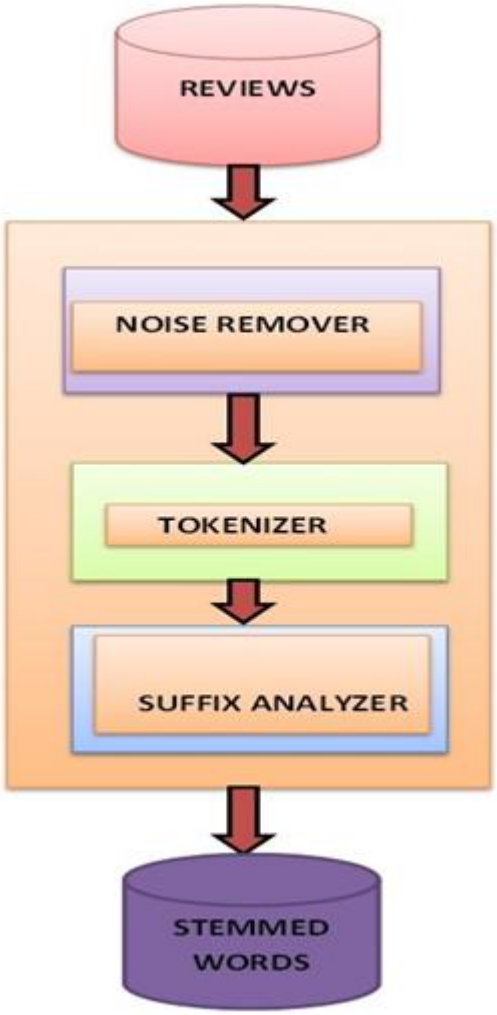

Fig. 2. Methodology Diagram

V. RESULTS AND DISCUSSION

The performance is monitored through various metrics like word stemmed factor, mean number of words, index compression factor, correctly stemmed words factor and average conflation factor. The same dataset was fed through the existing algorithms and the results obtained are tabulated.

Table- I: Analysis of Stemmers

Stemmer Analysis	Porter	Lancaster	Proposed
Total Number of Words(WC)	2008	2008	2008
Total Number of Stemmed Words(WS)	1949	1985	1957
Words Stemmed Factor(W)	97.06	98.85	97.46
Number of Distinct Words Before Stemming(BS)	808	808	808
Number of Distinct Words After Stemming(AS)	683	644	703
Mean Number of words(M)	1.18	1.25	1.15
Index Compression Factor(I)	15.47	20.30	13.00
Correctly Stemmed Words(CS)	1521	1099	1589
Incorrectly Stemmed Words(IS)	428	886	368
Correctly Stemmed Words Factor(C)	78.04	55.37	81.20
Correct Words not Stemmed(CW)	59	23	51
Number of Distinct Words after Conflation(N)	624	621	652
Average Conflation Factor(A)	58.97	43.49	58.97

The online reviews about a branded tablet were collected and 2008 tokenized words were considered for analysis. The following observations were made:

A. Word Stemmed Factor

The word stemmed factor signposts the stemmed words among the available words. The word stemmed factor calculated for all the three stemmers are well above the threshold value i.e.50%.The efficiency of all the stemmers were good.

B. Mean number of words and index compression factor

The mean number of words is approximately equal for all the stemmers. The index compression factor is observed to be less for the proposed stemmer. Both these gauging factors indicate the storage requirements.

C. Correctly stemmed words and incorrectly stemmed words

Both the parameters are indicators of the strength of the stemmers. It is desirable to have maximum value in the case of number of correctly stemmed words and the minimum value for the number of incorrectly stemmed words. The table depicts that the proposed stemmer meets the above said desirable results. Porter shows a higher value than Lancaster in terms of the number of correctly stemmed words and lower value for the number of incorrectly stemmed words.

D. Correctly Stemmed Word Factor

The number of correctly stemmed words is observed to be the maximum for the proposed stemmer. It is clear that the higher the correctly stemmed word factor, lesser the stemming errors. Hence, the proposed stemmer has less over-stemming and under-stemming errors.

E. Correct words not stemmed

The stemmed words produced include three categories of words namely correctly stemmed, incorrectly stemmed and correct words not stemmed. The first two have been discussed above. The number of correct words not stemmed is moderate for the proposed stemmer.

F. Number of Distinct Words after Conflation

The number of distinct words after conflation is more or less equal in the case of Porter and Lancaster stemmers. The number is relatively higher for the proposed stemmer that indicates the minimum rate of over stemming and under stemming errors.

G. Average Conflation Factor

The average conflation factor is equal for both the proposed stemmer and Porter stemmer. The higher value obtained proves the efficiency of the stemmer in word conflation.

H. Over Stemming and Under Stemming Errors

The correctly stemmed word factor and average conflation factor are high for the proposed stemmer. These parameters are indicators of less over and under stemming error rates.

VI. CONCLUSION AND FUTURE WORK

The importance of stemming which is an important preprocessing step has been understood through the analysis. The proposed stemmer has outperformed the existing stemmers namely Porter and Lancaster in terms of correctly stemmed words factor and average conflation factor. The index compression factor, the mean number of words and words stemmed factor are observed to be moderate. The higher correctly stemmed word factor and distinct words after conflation are clearly proving that the over-stemming and under-stemming error rate is relatively lesser than the existing rule based stemmers namely Porter and Lancaster. In future, the performance can be further enhanced by addition of more rules and the similar work can be extended towards other languages.

REFERENCES

1. R. Cynthia Monica Priya and J.G.R.Sathiaseelan, "An Explorative Study on Sentiment Analysis," IEEE, 2017.
2. A Pappu Rajan, "Web Sentiment Analysis," International Journal of Applied Research, 2016, pp. 563-566.
3. Anvitha Hedge and Mrs. Savitha K Shetty, "A Study on Stemming Algorithms," IJETST-Vol.02, Issue 05, May 2015 , ISSN 2348-9480, pp. 2361-2364.
4. Hussein K. Al-Khafaji and Areej Tarief Habeeb, "Efficient Algorithms for Preprocessing and Stemming of Tweets in a Sentiment Analysis System," e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 19, Issue 3 , 2017, pp. 44-50.
5. S. Vijayarani, Ms. J. Ilamathi and Ms. Nithya, "Preprocessing Techniques for Text Mining - An Overview," International Journal of Computer Science & Communication Networks, Vol. 5(1), 7-16, ISSN: 2249-5789, 2015.
6. Jasmeet Singh and Vishal Gupta, "A systematic review of text stemming techniques," Artificial Intelligence Review, Volume 48, Issue 2, 2017, pp. 157–217
7. Brajendra Singh Rajput and Dr. Nilay Khare, "A survey of Stemming Algorithms for Information Retrieval," e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 3, pp. 76-80, 2015.
8. S.P.Ruba Rani, B.Ramesh, M.Anusha and Dr. J.G.R.Sathiaseelan, "Evaluation of Stemming Techniques for Text Classification," IJCSMC, Vol. 4, Issue. 3, March 2015, pp. 165 – 171.
9. Giulio Angiani, Laura Ferrari, Tomaso Fontanini, Paolo Fornacciari, Eleonora Iotti, Federico Magliani, and Stefano Manicardi, "A Comparison between Preprocessing Techniques for Sentiment Analysis in Twitter," 2016,InKDWeb
10. Arkivi Krouska, Christos Troussas and Maria Virvov, "The effect of preprocessing techniques on Twitter Sentiment Analysis," IEEE, 2016
11. Sundar Singh and R K Pateriya, "Enhanced Suffix Stripping Algorithm to Improve Information Retrieval," International Journal of Computer Sciences and Engineering, Volume 3, Issue 8, 2015, pp. 115-119.
12. Atharva Joshi, Nidhin Thomas and Megha Dabhade, "Modified Porter Stemming techniques on Twitter sentiment analysis," International Journal of Computer Science and Information Technologies, Vol. 7 (1) ,2016, pp. 266-269
13. Ramalingam Sugumar and M. Rama Priya, " Improved Performance Of Stemming Using Efficient Stemmer Algorithm For Information Retrieval," Journal of Global Research in Computer Science, Volume 9, No.5, May 2018.
14. Mubashir Ali, Shehzad Khalid , M. Haneef Saleemi , Waheed Iqbal, Armughan Ali and Ghayur Naqvi, " A Rule based Stemming Method for Multilingual Urdu Text," International Journal of Computer Applications, Volume 134 ,2016.
15. Abdul Jabbara , Sajid Iqbalb, Adnan Akhunzadaa and Qaisar Abbas, "An improved Urdu stemming algorithm for text mining based on multi-step hybrid approach, " Journal of Experimental & Theoretical Artificial Intelligence, 2018.
16. Hunaida Awwad, Adil Alpkocak, "Using Hybrid-Stemming Approach to Enhance Lexicon-based Sentiment Analysis in Arabic," International Conference on New Trends in Computing Sciences, 2017 , pp.229-234.
17. M. Durairaj, A. Alagu Karthikeyan, "Modified Porter's Algorithm for Pre-Processing Academic Feedback Data," International Journal of Pure and Applied Mathematics Volume 118 No. 18 , 2018, pp. 3009-3015
18. Priya Govindarajan and Ravichandran K.S,"Modified Stemmer for a Medical System – Evaluated using Predefined Metrics," pp. 1746-1750, IEEE 2017.
19. Vairaprakash Gurusamy and Subbu Kannan,
20. " Performance Analysis: Stemming Algorithm for the English Language," IJSRD - International Journal for Scientific Research & Development, Vol. 5, Issue 05, ISSN (online): 2321-0613,2017.
21. Dr. M. Kasthuri, Dr. S. Britto Ramesh Kumar, "PLIS: Proposed Language Independent Stemmer Performance Evaluation," International Journal of Advanced Research in Computer Science & Technology, Vol. 5, Issue 4 ,2017,pp.91-96

AUTHORS PROFILE

Mrs. R. Cynthia Monica Priya has completed Master of Computer Applications in 2008 and M.Phil in 2011.She has passed all the degree programs with distinction from Bharathidasan University. She is working as an Assistant Professor in the Computer Science Department, Bishop Heber College, Tiruchirappalli, since 2012. She has cleared State Eligibility Test in 2016 and National Eligibility Test in 2019. She is pursuing Ph.D in Computer Science from Bharathidasan University. She has presented papers in International Conferences and has published a number of papers in reputed journals. Her area of specialization is Data Mining and in particular Web Mining. E-mail: cynmonpri@gmail.com

Dr. J.G.R. Sathiaseelan M.Sc., Ph.D is the Head of Computer Science Department, Bishop Heber College, Tiruchirappalli. He has three decades of teaching experience. He has presented many research papers in International Conferences and has published more than fifty research papers in reputed journals. He has authored a book entitled, "Programming In C#.Net". It was published by PHI Learning, New Delhi in 2009.He has successfully guided a number of M.Phil and Ph.D scholars. At present, he is guiding both M.Phil and Ph.D scholars in Bharathidasan University. His research areas include Web Services Security, Data Mining, Image Processing and Internet of Things. E-mail: jgrsathiaseelan@gmail.com