Rule Based Morphological Variation Removable Stemming Algorithm

Maheswari S, K. Arthi

Abstract*: Sentiment analysis is a technique to analyze the people opinion, attitude, sentiment and emotion towards any particular object. Sentiment analysis has the following steps to predict the opinion of a review sentences. The steps are preprocessing, feature selection, classification and sentiment prediction. Preprocessing is the main important step and it consists of many techniques. They are Stop word Removal, punctuation removal, conversion of numbers to number names. Stemming is another important preprocessing technique which is used to transform the words in text into their grammatical root form and is mainly used to improve the retrieval of the information from the internet. It is applied mainly to get strengthen the retrieval of the information. Many morphological languages have immense amount of morphological deviation in the words. It triggered vast challenges. Many algorithms exist with different techniques and has several drawbacks. The aim of this paper is to propose a rule based stemmer that is a truncating stemmer. The new stemming mechanism in this paper has brought about many morphological changes. The new rule based morphological variation removable stemming algorithm is better than the existing other algorithms such as New Porter, Paice/Lovins and Lancaster stemming algorithm.

Keywords* : *Preprocessing, Stemming, Index Compression Factor, Word stemmed Factor.

I. INTRODUCTION

In this digital world finding information from the internet is the main activity. In internet the information are indexed as documents. The relevant documents are retrieved and listed out in respect of given user query. There are many steps involved in this process. The one of the main step is preprocessing. . In preprocessing several steps are used for structuring the text and extracting features. The steps are tokenization, stop word removal and pos tagging and parsing. Tokenization is a method which is used to tokenize the file content into individual word. Removing stop words reduces the dimensionality of term space. The words which do not give the meaning of the documents such as articles preposition and pro-nouns are removed. Stemming is the method which is used to identify the root/stem of a word. For example the words joined, joining, join all can be stemmed to the word 'join'. The main work of this method is to

- Minimize the number of words
- It is to save the time and memory
- It removes various suffixes.

There are some problems in stemming. They are under stemming and over stemming. If the two words belong to the same conceptual group, and are converted to the same stem then there is no problem. Otherwise they are converted to different stem then it is Under Stemming. If the words belong to the different conceptual group and remain distinct after stemming then there is no problem, otherwise converted to same stem then it arise Over Stemming. Designing a rule based morphological variation removable stemming algorithm is often a question of finding the perfect balance between these two extremes. This paper aims to propose a rule based morphological variation removable algorithm that cascade under the truncating stemmer category and analyze its performance with the existing stemmers namely Porter and Lancaster.

II. RELATED WORKS

Stemming algorithm is classified into three groups. They are

- Truncating
- Statistical
- Mixed

All the above methods have their own way of finding the stem of the word variants. Truncating is the way to removing suffixes or prefixes of a word. Under truncating there are 4 different stemmers. Similarly the statistical type having 3 types of stemmers and mixed type is classified into 3 types of mixed stemmers and each have some merits and demerits. The following Fig.1 represents all the types of stemming algorithms.

Revised Manuscript Received on November 15, 2019

S Maheswari, Assistant Professor in the Department of Computer Science, Bishop Heber College, Trichy,

Dr. K. Arthi MCA.,M.Phil.,Ph.D, Assistant Professor in Department of Computer Applications, Government Arts College, Coimbatore.

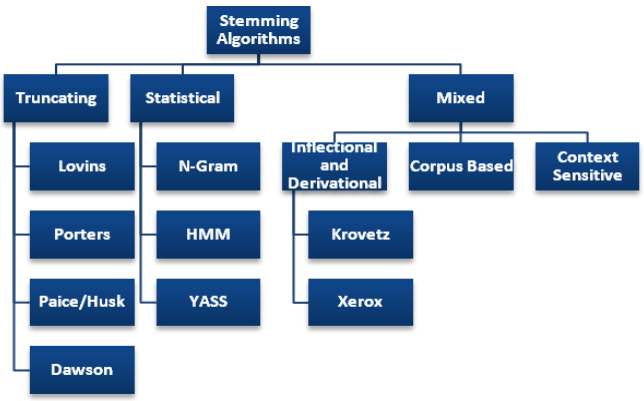

Fig. 1 Types of stemming algorithm

A. *Truncating Algorithms*

1) Lovins Method:

i) The Lovins Stemmer was proposed by Lovins in 1968.

ii) It removes the longest suffix from a word. It always removes a maximum of one suffix from a word.

iii) Removes the longest suffix of double letter.

iv) **Advantages:** Iit is very fast and can handle the removal of the double letters in words like 'getting' being transformed to 'get' and also handles irregular plurals.

v) **Drawback:** The Lovins approach is it is time consuming and many suffixes are not available in the table of endings. It is highly unreliable.

2) Paice/Husk Method:

It is an Iterative algorithm. One table containing about one hundred and twenty rules indexed by last letter of a suffix. Each rule specifies either a deletion or replacement of an ending. It is simple and every iteration taking care of both deletion and replacement as per the rule applied.

i) Advantages: Simple and every iteration taking care of both deletion and replacement as per the rule applied.

ii) Disadvantage: It is very heavy algorithm and over stemming may occur.

3) Porter Method:

Most commonly used truncation stemmers. It removes affixes from a word over a number of iterations until all the rules/conditions are considered. Less error rate.

i) **Limitations:** The stems produced are not always real words.

ii) It has at least five steps and sixty rules and hence is time consuming.

4) Dawson Method:

i) Covers more suffixes than Lovins

ii) Fast in execution

iii) **Limitations:**Very complex

iv) Lack a standard implementation.

B. *Statistical Stemmers*

1) N-Gram Method:

i) Words are conflated to their stem using a string-similarity approach.

ii) N-gram is a set of n consecutive characters extracted from a word. Words that are similar will have a high proportion of n-grams in common hence they can be converted to same stem. It is language independent.

iii) **Limitations:** Not time efficient

iv) Require significant amount of space for creating and indexing the n-grams.

v) Not a very practical method.

2) HMM Method:

i) Based on the concept of the Hidden Markov Model.

ii) Finite set of automata.

iii) Transitions between states are ruled by probability functions.

iv) Based on unsupervised learning and language Independent.

v) **Limitations**: A complex method for implementation.

vi) Over stemming may occur

3) YASS : Method:

i) Stands for Yet Another Suffix Striper.

ii) Statistical as well as corpus based

iii) Clusters are created using hierarchical approach and distance measures.

iv) Can be used for any language without knowing its morphology

v) **Limitations:** Difficult to decide a threshold for creating clusters

vi) Require significant computing power.

C. *Mixed Stemming*

1) Krovetz/K-STEM Method:

i) It heavily depends upon the contents of its dictionary. Its conflation might end up being conservative.

i) **Limitations:** For large documents, this stemmer is not efficient

ii) Inability to cope with words outside the lexicon

iii) Does not consistently produce a good recall and precision.

iv) Lexicon to be created in advance.

4) Xerox Method:

i) Linguists at Xerox Corporation created a lexical database for English which helps to identify the base word using morphological analysis of the word in the lexicon. It works well for a large document.

ii) Removes the prefixes where ever applicable

iii) All stems are valid words.

iv) **Limitations:** Inability to cope with words outside the lexicon.

v) Not implemented successfully on languages other than English. Over stemming may occur.

vi) Dependence on the lexicon makes it language dependent.

5) Corpus Based Method:

i) Conflation classes are automatically modified to overcome problems in Porter algorithm.

ii) The significance of word form co-occurrence can be determined by the statistical measure given by Em(a,b)=nab/(na+nb).

iii) **Limitations:** Potentially keep away from making conflations that are not appropriate for a given corpus and the result is a real word stem.

iv) It is complex and processing time increases.

III. LITERATURE REVIEW

In the digital world, there are vast amount of information stored in shapeless text format. This shapeless text cannot be used for further processing by computer. Hence specific preprocessing methods and algorithms are required in order to extract useful patterns. The Text Mining normally to the method of extracting interesting information and knowledge from the unstructured data. Text mining is the process of discovering information in text documents.

- Four different affix removal stemmers were analyzed [1]. Porter, Lovin's stemmer, Paice and Krovertz Stemmer were taken for experimentation. The strength and computational time of each algorithm was calculated. According to the experimentation, it was observed that the paice algorithm performed while comparing the ICF value. Execution time taken for Lovin‟s stemmer is faster than other stemmers. But all these four stemmers are not able to generate correct stem or root words.
- Improved porter's algorithm [2] was proposed by taking prefixes into an account. It also has dealt with the enhancement of text pre-processing technique.
- An automatic word stemming system for Hausa language [3] was proposed the modified porter's algorithm to fit Hausa morphological rules. The accuracy was up to 73.8% for implementation with 2573 words extracted from four different articles from Hausa leadership newspaper.
- A new context free stemmer was proposed [4] which was a grouping of traditional rule based system with string similarity approach. This hybrid algorithm was language dependent algorithm. This algorithm was tested on Nepali language which is based on Devanagari script. The accuracy obtained from this hybrid algorithm is 70.10% which was higher than the traditional rule based approach.
- An improved version of the original porter stemming algorithm [9] was proposed for the English language. Error counting method was used to evaluate the proposed stemmer. The performance of the stemmer was computed by calculating the under stemming and over stemming errors.
- A new effective light stemmer algorithm [5] was developed to over the high percentage of error caused by other stemming algorithms. The new technique truncates the word infixed in addition to the prefixes and suffixed based on simple rules.
- Stefano et al. [8] discussed automatic learning methods of linguistic resources for stop words removal.
- Wahiba et al.[9] proposed new stemmer for rectifying the limitations of porter stemmer algorithm.
- Giridhar et al. [11] conducted a prospective study of stemming techniques in web documents.
- Toms et al. [12] is explained High Precision Stemmer (HPS) methods. HPS is not easy to decide a threshold for creating clusters and needed significant computing power.
- Venkatsudhakarareddy et al. [13] imparted stemming technique applied to information extraction using RDBMS.
- Sandeep et al. [14] have considered the strength of affix removal stemmers and also discussed comparative analysis of affix removal stemming algorithm accuracies.

IV. PROPOSED APPROACH

The Rule Based Morphological Variation Removable Stemming algorithm decreases different morphological variation to their basic rules. Rule based stemmers are used to make over the alternative word forms into their stems or base forms by using certain predefined language-specific rules. Rule based stemmers sometimes take up additional linguistic resources like dictionaries to conflate morphologically related words. Stemming is used to allow matching of queries and documents in keyword-based information retrieval systems. This assumes that morphological variants of words have related semantic interpretation and can be considered as equivalent for the purpose of information retrieval applications. The following figure shows the general outlook of the proposed system. The Rule Based Morphological Variation Removable Stemming algorithm rectifies the drawbacks of earlier stemming algorithms.

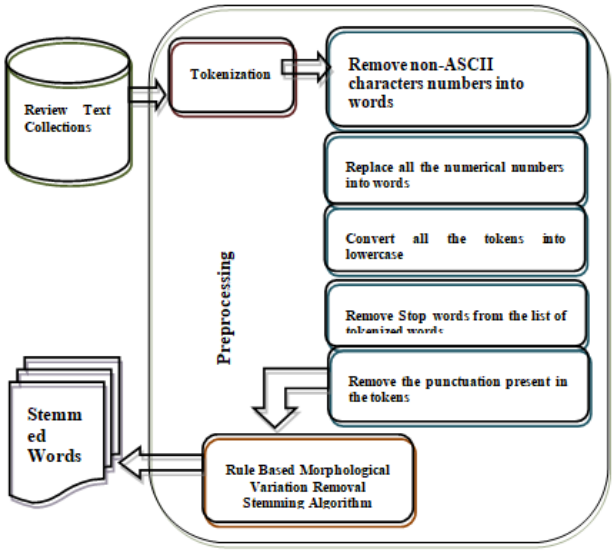

Fig. 2 Overview of Proposed System

The above Fig. 2 represents the architecture of the proposed system. Text Mining starts with group of documents, which would retrieve a particular document and preprocess it by checking format and character sets. Preprocessing is the process to make better the unstructured documents into a structure format. Dataset is an formless dataset of documents which are pre processed using the following three rules:

A. Tokenize the file into respective separate tokens using space as the delimiter.
B. Removing the stop word which does not communicate any meaning.
C. Use stemmer algorithm to find out the stem of the word with the common root word.

The Rule Based Morphological Variation Removable Stemming Algorithm:

Step1: Start the Process

Step2: Read text from the collections

Step3: Start preprocessing

Step3.1: tokenize the text using space as the delimiter

Step3.2: Remove non-ASCII characters from the list of tokenized words

Step3.3: Convert all the tokens into lowercase

Step3.4: Remove the punctuation present in the tokens

Step 3.5: Replace all the numerical numbers into words

Step 3.6: Remove Stop words from the list of tokenized words.

Step3.7: Apply stemming rules in list of words

Step3.8: Finally store the root/stemmed words in a document

Step 4: Repeat the step1 for read the next text until to reach last text in a document.

Step 5: Stop the Process

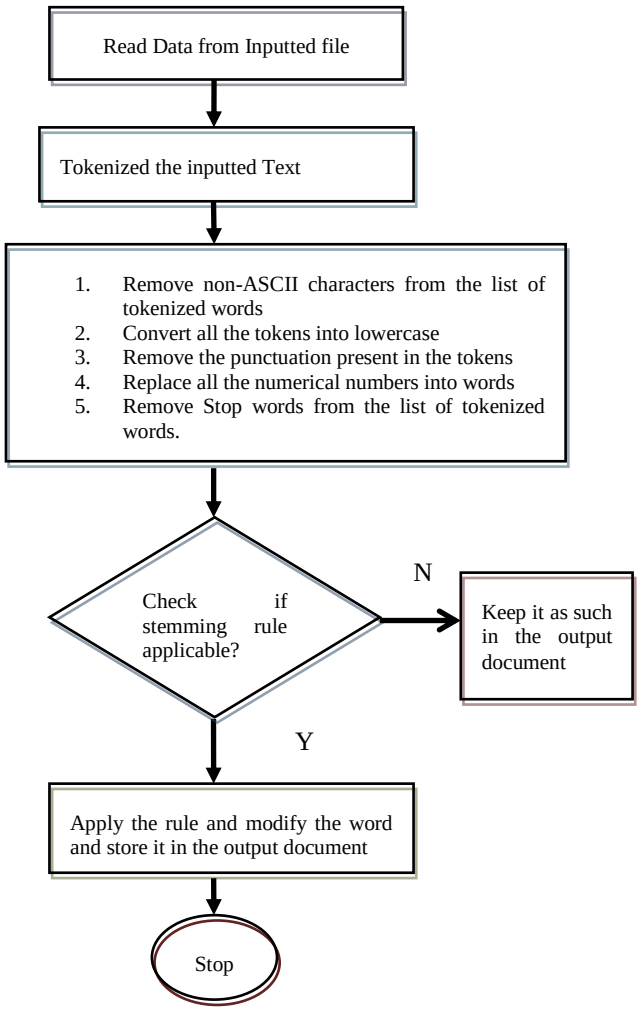

Fig. 3 Data Flow Diagram of Rule Based Morphological Variation Removable Stemming algorithm

V. EXPERIMENTAL RESULT

To evaluate the performance of the stemmer, the new algorithm and other existing algorithms has been applied to the sample vocabulary downloaded from the web site KAGGLE. The data set contains lots of review details. The initial preprocessing steps have been applied to the dataset and tokenize the sentences into words. This token contains correct and incorrect words. The new stemmer has been developed using Python. The new stemmer algorithm is based on rule based strategies. The existing algorithms such as Porter, Lancaster and the new stemmer algorithm has applied in same dataset and calculated the performance metrics of stemming algorithms.The performance metrics are list out in the following Table 2. Thus from the table the new stemmer algorithm is the best as compare with other two earlier stemming algorithms. The strength and accuracy of the stemmer are evaluated by using the following measuring criteria:

i) Index Compression Factor (ICF):

The index compression factor represents a percent by which a collection of distinct words is reduced by stemming. The large number of words stemmed will give the greater strength of the stemmer. This is calculated as :

$$ICF = ((N - S)/S) * 100$$

where
N- Number of distinct words before stemming
S- Number of distinct stems after stemming.

ii) Word Stemmed Factor (WSF):

It is the percentage of words that have been stemmed by the stemming process out of the total words in a sample. Improved the number of words stemmed, greater the strength of the stemmer. Minimum threshold for this factor should be 50%.

$$WSF = (WS/TW) * 100$$

iii) Correctly Stemmed Words Factor (CSWF):

It is the percentage of the number of words stemmed. Higher the percentage of this factor, higher will be the accuracy of the stemmer. Minimum threshold for his factors should ne 50%.

$$CSWF = \left(\frac{CSW}{WS}\right) * 100$$

iv) Average Words Conflation Factor (AWCF):

This indicates the average number if variant words of different conflation group that are stemmed correctly to the root words. To calculate AWCF, first to find out the number of unique words after conflation as:

$$\boldsymbol{NWC = S - CW}$$

where,
CW- Number of correct words not stemmed. Thus Word conflation Factor is obtained as:

$$AWCF = \frac{CSW - NWC}{CSW} * 100$$

Greater the percentage of AWCF, greater will be the accuracy of the stemmer.

The result of the proposed algorithm and existing most noticeable aggressive stemmers referred in this paper are shown in Table – I and the visualization of the result is represented in Fig. 4.

Table –I Performance metrics of Existing and Rule Based Morphological Variation Removable Stemming Algorithms

Metrics/Stemmers	Porter	Lancaster	RBMVRS
Total Number of Words (TW)	1464	1464	1464
Before Stemming(BS)	728	728	728
After Stemming(AS)	642	602	681
Word Stemmed(WS)	1432	1398	1433
Correct Stemmed Words(CSW)	1017	667	1288
Not Stemmed Words	543	485	619
Index Compression Factor(ICF)	13.55	20.93	7.04
Word Stemming Factor(WSF)	97.81	95.50	97.88
Correct Stemmed Word Factor(CSWF)	71.01	47.71	89.88
Average Word Conflation Factor(AWCF)	46.60	27.29	51.94

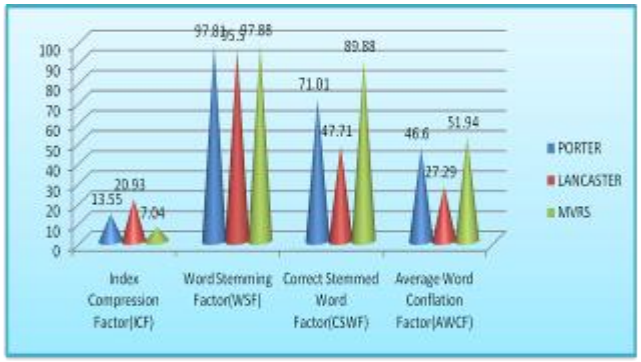

Fig. 4 Performance metrics of Existing and Rule Based Morphological Variation Removable Stemming Algorithms

- From the above result the index compression factor is observed to be less for the proposed stemmer.
- The word stemmed factor calculated for all the three stemmers. The efficiency of all the stemmers were good.
- The correctly stemmed factor is 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.
- The average conflation factor of the proposed stemmer is high. This higher value proves that the efficiency of the stemmer.
- The higher value of correctly stemmed factor and average conflation factor indicates that the less over and under stemming errors rates.

VI. CONCLUSION

This paper deals with pre-processing of the text data before subjecting the text to classification or validation analysis. Junk data may affect the accuracy of knowledge gained from the review data. Considering that fact, the review data is pre-processed. It is tokenized using NLTK tokenize and then stop word are removed using dictionary based approach. Finally, a new Rule Based Morphological Variation Removable Algorithm is used to stem the data and the stemmed words are subjected to evaluation.

The experimental result obtained satisfies the new Rule Based Morphological Variation Removable Algorithm as a stemming algorithm which can stem with less error compared with the existing algorithms. Preprocessed review data is then classified using classification algorithm which is taken as the future direction of this work.

REFERENCES

1. Rupan Gupta, Dr. Anjali Ganesh Jivani, "Empirical Analysis of Affix Removal Stemmers", International Journal of Computer Technology and Applications, Volume 5, Issue 2, pp 393-399, 2014.
2. R. Jayanthi Ms. C. Jeevitha, "An Approach for Effective Text Pre-Processing Using Improved Porters Stemming Algorithm", - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 7, pp 797-807, July 2015.
3. Muazzam Bashir, Azilawati Binti Rozaimee, Wan Malini Binti Wan Isa, "A Word Stemming Algorithm for Hausa Language", IOSR Journal of Computer Engineering (IOSR-JCE), Volume 17, Issue 3, pp 25-31, Ver. VI (May – Jun. 2015).
4. Chiranjibi Sitaula, "A Hybrid Algorithm for Stemming of Nepali Text", Intelligent Information Management, Volume 5, pp 136-139, 2013.
5. Sohair Al hakeem, Ghazi Shakah, Belal abu Saleh, Nisreen Jaber Thalji, "Developing an Effective Light Stemmer for Arabic Language Information Retrieval", International Journal of Computer and Information Technology, Volume 05 – Issue 01, pp 55-59, January 2016.
6. Francesco Colace, Massimo De Santo, Luca Greco and Paolo Napoletano, "Text classification using afew labeled examples", Computer in Human Behavior 30(2014)689-697, Elsevier, 2013.
7. B.Ramesh, J.G.R.Sathiaseelan, "A Theoretical Stuudy on Advanced Techniques in Pre-Processing and Text Classification" International Journal of Data Mining and Emerging Technologies, Vol.5, No.1,pp.6-10, 2015.
8. Stefano Ferilli, Floriana Esposito and Domenico Grieco, "Automatic Learning of Lingustic Resources for Stopword Removal and Stemming from Text", Procedia Computer Science 38 (2014) 116-123, Elsevier, 2014.
9. Wahiba Ben AbdessalemKaraa,"A new stemmer to improve information retrieval", International Journal of Network Security & Its Applications, July 2013.
10. AlperKursatUysal and SerkanGunal, "The impact of preprocessing on text classification", Information Processing and Management 50(2014) 104-112, Elsevier, 2013.
11. GiridharN.S,Prema K.V and N.V SubbaReddy,"A Prospective Study of Stemming Algorithms for Web Text Mining", Ganpat University Journal of Engineering & Technology,Vol-1,Issue-1,Jan-Jun-2011.
12. Tomas Brychcin, Miloslav Konopik. "HPS: High precision stemmer". Information Processing and Management 51 pp.68–91, 2015.
13. VenkatSudhakaraReddy.Ch and Hemavathi.D, "Information extraction using RDBMS and stemming algorithm", International Journal of Science and Research (IJSR), April 2014
14. Sandeep R.Sirsat, Vinay Chavan and Hemant S.Mahalle, "Strength and Accuracy Analysis of Affix Removal Stemming Algotithms", International Journal of Computer Science and Information Technologies, Vol. 4(2), 2013, 265-269

AUTHORS PROFILE

S Maheswari, has completed Master of Computer Applications in 2000 and M.Phil in 2013. She worked as an IT Trainer at TRECSTEP for 8 years. She is working as an Assistant Professor in the Department of Computer Science, Bishop Heber College, Trichy, since 2013. She is pursuing Ph.D. at Bharathiar University, Coimbatore. She has cleared State Eligibility Test (SET) in 2018. She has fifteen years of experience in software development. She has cleared Microsoft Database Administration Certificate exam. She has presented papers in International Conferences and has published papers in reputed journals. Her area of specialization is Data Mining and in particular Web Mining.

Dr. K. Arthi MCA.,M.Phil.,Ph.D, is working as an Assistant Professor in Department of Computer Applications, Government Arts College, Coimbatore. Her area of specialization is Soft computing, Evolutionary Algorithms, Fuzzy cognitive maps. She has more than 19 years of teaching experience. She got her Ph.D research degree in Anna University, Chennai. She is the life member of ISTE. She has published many papers. She has successfully guided a number of M.Phil and Ph.D scholars. At present, she is guiding both M.Phil and Ph.D scholars in Bharathiar University and Karpagam University. Her research area include, Data Mining, Soft Computing.