ENHANCING THE STEMMING ALGORITHM IN TEXT MINING

A Dissertation Submitted

By

Anil Kumar Soni

Registration No. 11111310

To

Department of Computer Science & Engineering

In partial fulfillment of the Requirement for the
Award of the Degree
Of

Master of Technology in Computer Science & Engineering

Under the guidance of

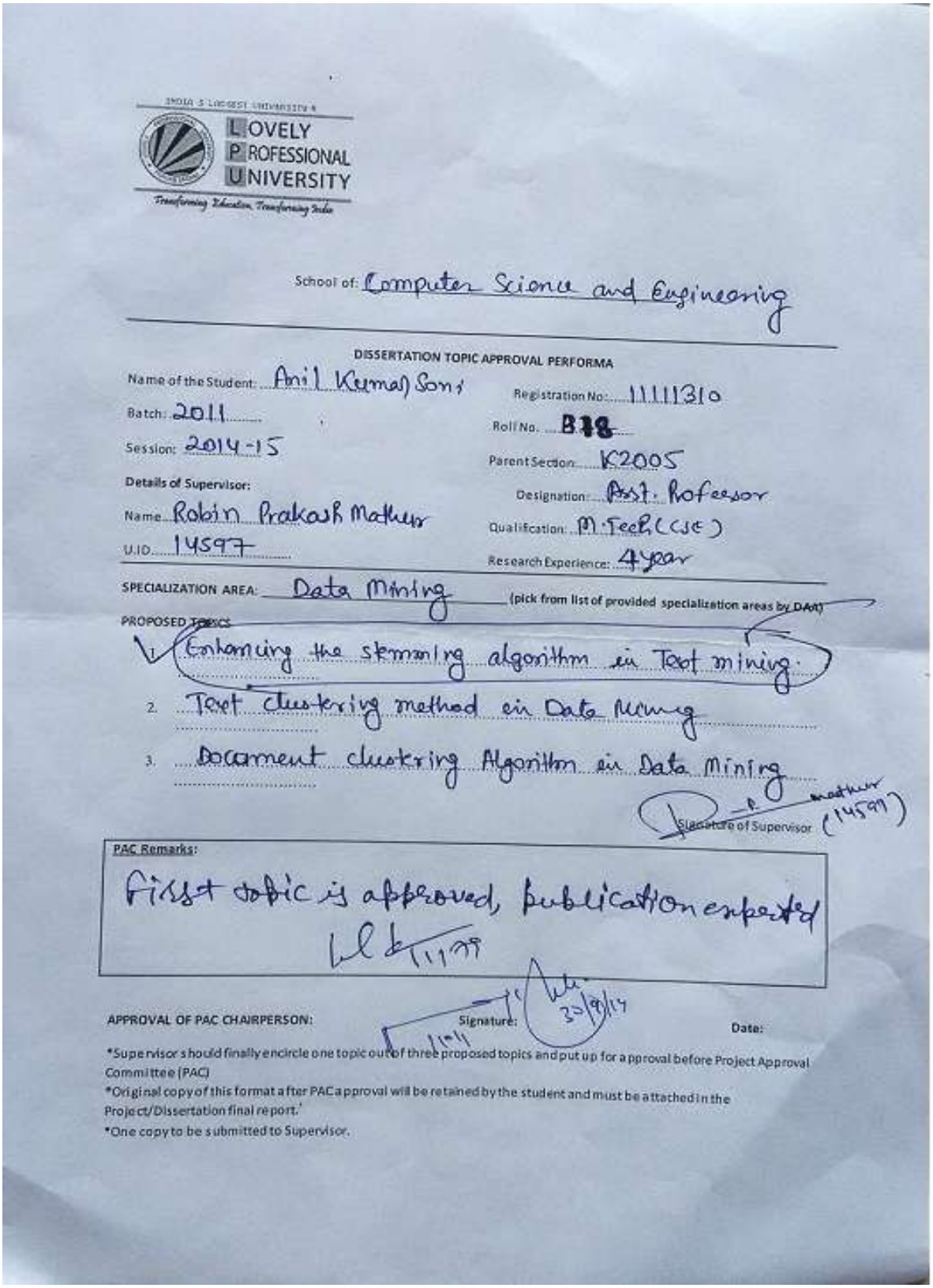
Robin Prakash Mathur

Asst. Professor

Department of CSE, LPU

May 2015

PAC APPROVAL FORM



INDIA'S LARGEST UNIVERSITY

LOVELY PROFESSIONAL UNIVERSITY

Transforming Education, Transforming India

School of: Computer Science and Engineering

DISSERTATION TOPIC APPROVAL PERFORMA

Name of the Student: Anil Kumar Soni

Registration No: 11111310

Batch: 2011

Roll No. B38

Session: 2014-15

Parent Section: K2005

Details of Supervisor:

Designation: Asst. Professor

Name: Robin Prakash Mathur

Qualification: M.Tech (CSE)

U.ID: 14597

Research Experience: 4 year

SPECIALIZATION AREA: Data Mining (pick from list of provided specialization areas by DAA)

PROPOSED TOPICS

1. Enhancing the stemming algorithm in Text mining.
2. Text clustering method ein Data Mining
3. Document clustring Algorithm in Data Mining

Signature of Supervisor R. Mathur (14597)

PAC Remarks:

First topic is approved, publication expected

APPROVAL OF PAC CHAIRPERSON: Signature: Date: 30/9/14

*Supervisor should finally encircle one topic out of three proposed topics and put up for approval before Project Approval Committee (PAC)

*Original copy of this format after PAC approval will be retained by the student and must be attached in the Project/Dissertation final report.'

*One copy to be submitted to Supervisor.

ABSTRACT

In this research, we have worked on the stemming algorithm which is used in text mining. The text mining is the process which is used to fetch the knowledge from textual database. Most of the organization used the large amount of data for analysis. It can be performed by the text mining. Pre-processing is required to be performed before the text mining process. It is divided into three parts which are Tokenization, Stop words removals and Stemming process. Tokenization is based on splitting the sentence into multiple tokens or words. Stop words removal is used for avoid the indexing useless words. Stemming is based on transform the words into their base form or root form. The initial process of text mining is pre-processing technique. We can improve the performance of text mining using enhancement of the stemming algorithm. The aim of stemming is to provide the effective result for text mining.

In this research we have proposed an algorithm for stemming named as TPLMSA. The TPLMSA (Two Phase Longest Match Stemming Algorithm) is based on the longest match principal. Our work is focused on enhancing the stemming algorithm so as improve the overall performance of text mining.

CERTIFICATE

This is to certify that **Anil Kumar Soni** has completed M.Tech dissertation titled **Enhancing the Stemming Algorithm in Text Mining** under my guidance and supervision. To the best of my knowledge, the present work is the result of her original investigation and study. No part of the dissertation has ever been submitted for any other degree or diploma.

The dissertation is fit for the submission and the partial fulfillment of the conditions for the award of M.Tech Computer Science & Engg.

Date: _____ Signature of Advisor

Robin Prakash Mathur

ACKNOWLEDGEMENT

"Keep away from people who try to belittle your ambitions. Small people always do that, but the really great make you feel that you too, can become great". (Mark Twain)

We take this opportunity to express our sincere thanks and deep gratitude to all those people who extended their wholehearted co-operation and have helped us in completing this dissertation successfully.

First of all, we owe sincere thanks to our **Mr. Rajiv Sobti (HOS)** and **Mr. Dalwinder Singh (HOD)** for her appreciations and her support all the time. Without her support this dissertation would not have taken up the shape.

We are highly indebted and grateful to Mr**. Robin Prakash Mathur (Dissertation Mentor)** for his strict supervision, constant encouragement, inspiration and guidance, which ensure the worthiness of our work. Working under him was an enrich experience. We express our sincere thanks to him for his cooperation, encouragement and valuable suggestion.

We would also like to thank our parents for guiding and encouraging us throughout the duration of the dissertation.

In all we found a congenial work environment in **Lovely Professional University** and this completion of the dissertation will mark a new beginning for us in the coming days.

Anil Kumar Soni (11111310) Dated: _____

Place: Lovely Professional University

DECLARATION

I hereby declare that the dissertation entitled, **Enhancing the Stemming Algorithm in Text Mining** submitted for the M.Tech Degree is entirely my original work and all ideas and references have been duly acknowledged. It does not contain any work for the award of any other degree or diploma.

Date: _____

Anil Kumar Soni

11111310

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1
INTRODUCTION

In today era, the data growth rate increase day by day and the huge amount of data is unstructured [Luying LIU, Jianchu KANG, Jing YU, Zhongliang WANG (2005)]. Most of the organization performs the analysis on textual data [Jayaraj Jayabharathy and Selvadurai Kanmani (2014)]. Text analysis or text mining is performed after the text pre-processing technique. The text pre-processing techniques include the Tokenization, Stop word removal and stemming [C.Ramasubramanian, R.Ramya (2013)].

Our work is based on the stemming algorithm. It is used for transforming the multiple words into their root word [Tuomo Korenius, Jorma.Laurikkala, Kalervo.Jarvelin, Martti Juhola (2004)]. In this section, we have discussed the concept of data mining, text mining and pre-processing technique for the text mining. Finally we have focused on the stemming algorithm. In this research we proposed the new algorithm TPLMSA Two Phase Longest Match Stemming Algorithm for the stemming. The main objective of the research is based on enhancing the stemming algorithm. The enhancement of stemming algorithm it leads to improve the overall performance of text mining.

1.1 DATA MINING

The data mining is the process which is used for extract the potential and useful information from the huge amount of data [C.Ramasubramanian, R.Ramya (2013)]. The objective of data mining is to extract the knowledge from the database. It can be used by many organizations for predict and discover useful information [Liu HaiTao, Cong Jin (2013)]. It is a proactive strategy for business and industry to establish the growth. Using Data Mining, an organization can take effective and good business decisions for improve the marketing, CRM (customer relationship management), advertising, etc and these decisions that will help in companies growth [Jiawei Han and Micheline Kamber (2006)].

It includes functional modules for tasks such as classification, prediction, characterization, correlation analysis, association and, evolution analysis, cluster analysis, outlier analysis [C.Ramasubramanian, R.Ramya (2013)]. The following figure shows the steps of the data mining.

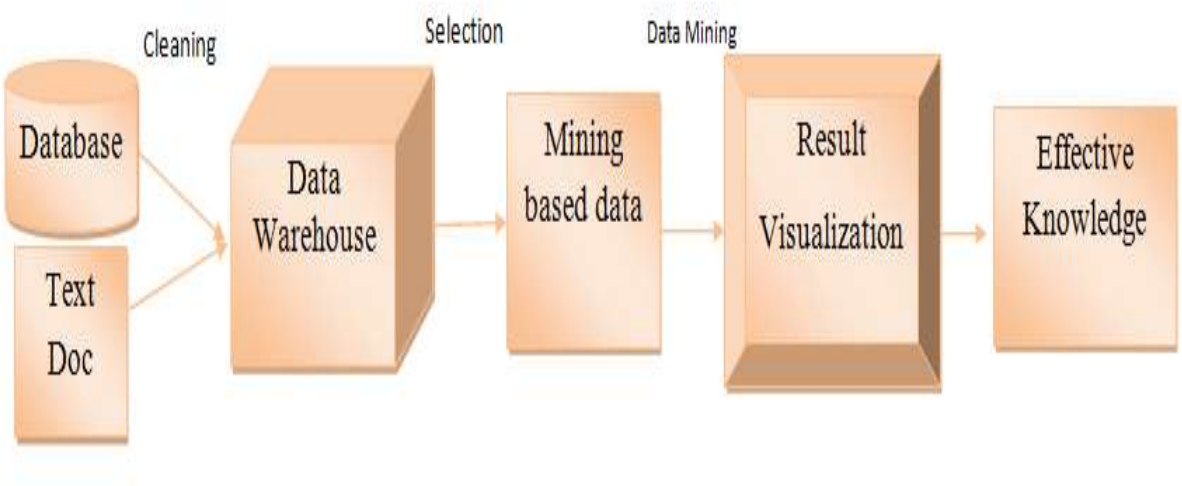


Figure 1Process of Data Mining

1.2 TEXT MINING

Text mining is the technique which is used for find the knowledge from text based database [Yanping Lu, Shengrui Wang, Shaozi Li, Changle Zhau (2011)]. It is the process of extracting interesting knowledge or information from unstructured textual document or database [Zhong, Ning, Li, Yuefeng, & Wu, Sheng-Tang (2010)] [Yongzhe shi, Wei-Qiang Zhang, Jia Jiu, Michael T Johnson (2013)].

It is based on the unstructured data. The textual database includes the huge collection of text documents from different sources [Jiawei Han and Micheline Kamber (2006)]. The text document sources are following listed [Anindya Ghose, Panagiotics G. Ipeirotis (2010)] [Jayaraj Jayabharathy and Selvadurai Kanmani (2014)]

- Digital Libraries which means online library.
- The research paper collected from different journals.
- Education Books from different department.
- The News Articles from different news papers
- The Web Pages contents for any search [Daniel Ramage, Christopher D. Manning, Susan Dumais (2011)].
- The Email Messages from mail services, etc.

The most of the Business sector, Government sector, institutions, and other industry are stored the data in form of text database [Jayaraj Jayabharathy and Selvadurai Kanmani (2014)]. The text data is based on the unstructured format. To extracting useful knowledge or information from the unstructured sources, we required text mining technique.

Which are following listed.

- Text summarization
- Text classification
- Text clustering
- Text information extraction
- Text visualization

Text clustering or Document clustering is the process of text extraction and fast text knowledge retrieval from the huge amount of textual database or document database [Tao Liu, Shengping Liu, Zheng Chen, Wei- Ying Ma (2003)] [Sun Kim, W John Wilbur (2012)]. It is closely related to data clustering [Luying LIU, Jianchu KANG, Jing YU, Zhongliang WANG (2005)] [Liu HaiTao, Cong Jin (2013)].

Basically clustering is the process of grouping the similar text from the large amount of text data [Jayaraj Jayabharathy and Selvadurai Kanmani (2014)]. The outlier detection can be performed by clustering.

The outlier values are different from the cluster value. Clustering is unsupervised learning which means it not having the predefined classes [Jiawei Han and Micheline Kamber (2006)].

The following figure shows the concept of text clustering.

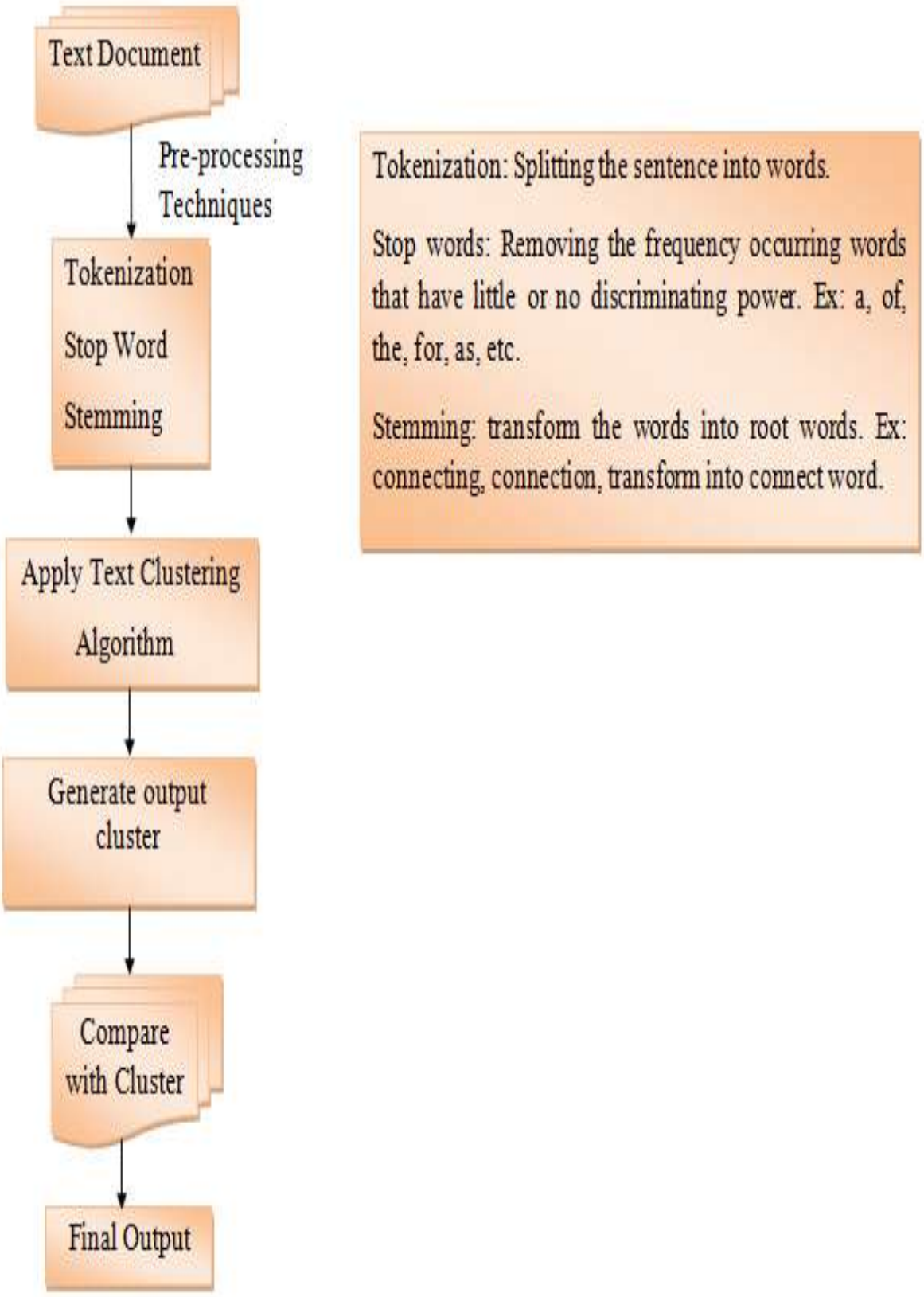


Figure 2Flow Chart of text Clustering

1.3 PRE-PROCESSING TECHNIQUE OF TEXT MINING

Pre-processing is technique through which we can manipulate our data for mining. In structured data we can perform data mining. Data mining is also having pre-process technique which is called ETL (Extract, Transform, and Load). Same as the data mining, the text mining also used the pre-process technique for effective result.

In text mining, the pre-processing is divided into three parts which are following: [Tuomo Korenius, Jorma.Laurikkala, Kalervo.Jarvelin, Martti Juhola (2004)]

- Tokenization
- Stop words removal
- Stemming

Tokenization is identifying the keywords for representing documents. It is the process of splitting the sentences into many separate tokens. For example, "this is the dissertation-I report for M.Tech" is split as: this/is/the/dissertation-I/report/for/M.Tech [Julie Beth Lovins (1968)].

Stop words removal are used for removing the useless word. In other words, the stop words removal is the word which has less useful information related to the text document [Ms. Anjali Ganesh Jivani (2011)]. For example: an, a, the, of, for, with, as, about, all, so, etc [Jayaraj Jayabharathy and Selvadurai Kanmani (2014)].

Now the final pre-processing technique is stemming which means a multiple words may share the same word stem [Peter Willet (2006)].

For example the multiple words which have drugs, drugged and drug are showing the same stem word drug. It is the way of removing the affixes in words and produce new word that is base (root) word. It is the process of transform the words into base form.

For example: connection, connecting, connected words transformed into connect word.

The algorithms for stemming are proposed by Porter 1998, Lovins 1968 and s-removal Harman 1991 [Peter Willet (2006)].

Stemming is the pre-processing technique in text mining. Text mining used for access the relevant information from multiple textual documents. Hence the initial process of text mining is pre-processing technique.

If pre-processing technique is produce the effective results then we save the memory space and time requirement for performing the mining in textual database [C.Ramasubramanian, R.Ramya (2013)].

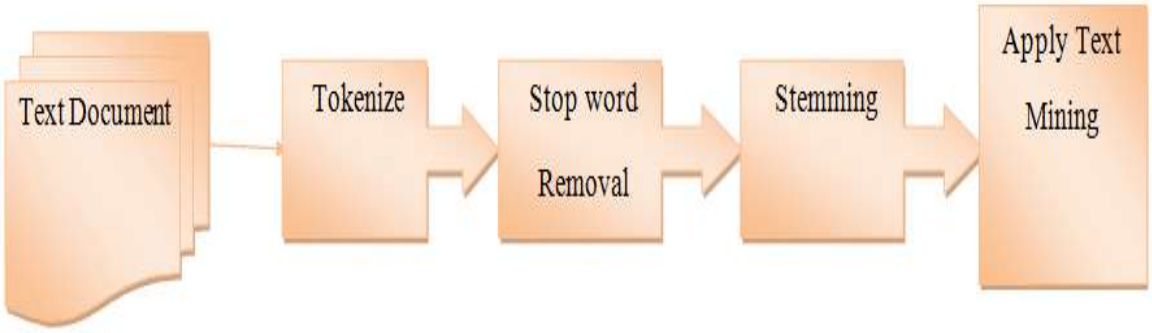


Figure 3Flow Chart of Pre-processing of text mining

Chapter 2

REVIEW OF LITERATURE

The Author Jayaraj Jayabharathy and Selvadurai Kanmani discuss about the document clustering and propose three new algorithms for document clustering [**Jayaraj Jayabharathy and Selvadurai Kanmani (2014)**].

Increase in the number of documents in the corpuses like News groups, government organizations, and internet and digital libraries; have led to greater complexity in categorizing and retrieving them. The Author proposed three dynamic document clustering algorithms which are Term frequency based maximum resemblance document clustering (TMARDC), Correlated concept based maximum resemblance document clustering (CCMARDC) and Correlated concept based fast incremental clustering algorithm (CCFICA) are proposed. From the above three proposed algorithms the TMARDC algorithm is based on term frequency, whereas, the CCMARDC and CCFICA are based on Correlated terms (Terms and their Related terms) concept extraction algorithm.

The Document clustering process is very useful for find the useful information from the textual database. The document clustering was finding for improving the precision or recall in information extraction systems and effective process to finding the similarity of the document.

The automatic generation of cluster hierarchy is based on document clustering. For example, when we search the string on browser it returns the thousands of pages related to that query, then it is difficult to finding the relevant document from the multiple document.

The Vector Space Model is the base concept of the every text clustering method. The VSM basically used for text clustering and the classification. The VSM is based on the concept of vector; the vector is collection of word from the document.

The Author C. Rama Subramanian, R. Ramya is discussing the pre-processing step in the text mining for the porter stemming algorithm [**C.Ramasubramanian, R.Ramya (2013)**].

The pre-processing is the initial process of the text mining. If our pre-processing steps are good then we are save the space as well as time by improvement of porter algorithm. The main aim of stemming algorithm is transform the text words into their root form.

The popular stemming algorithm is porter but it stills some drawbacks of handling the named entity. The name entity is the noun. Ex: Delhi

The Author Ms. Anjali Ganesh Jivani focus on the comparative study of stemming algorithm [**Ms. Anjali Ganesh Jivani (2011)**].

Stemming is a pre-processing step in text mining for extraction of knowledge from various textual documents. The main purpose of stemming algorithm is to reduce different grammatical form to its root form. The process of stemming is usually done by removing any attached suffixes and prefixes from index term before the actual assignment of the term to the index.

The stemming algorithm is used for the find the stem words from the multiple words without take care the part of speech POS. The lemmatization deals with 'lemma' of a word which involves reducing the word after understanding the POS.

Ex: stemming

- introduction, introducing, introduce- introduc
- gone, going, goes, went -go

Ex: Lemmatization

- introduction, introducing, introduce- introduce
- gone, going, goes, went -go

The Author Peter willett is focus on the porter stemming algorithm for the stem words [**Peter Willet (2006)**]. The first porter algorithm for stemming was proposed in 1980 which was used for the English language.

The first stemming algorithm was developed by Lovins 1968 for information retrieval application. They work on basis of the dictionary which contain the all the suffixes like ing, ation, SES, etc. The Lovins algorithm is the backbone of the all other stemming algorithm like Lennon 1981, Porter 1980, 2005. During the stemming process it works on concept of the word suffixes match with the suffixes dictionary if the suffixes match then we remove it and find the stem word.

In 2006 the new porter algorithm was developed which was use the snowball. The snowball is the programming language. The snowball provides the rules and regulation of stemming algorithm.

The Author Luying LIU, Jianchu KANG, Jing YU, Zhongliang WANG discuss about the text clustering for the selection of unsupervised information [**Luying LIU, Jianchu KANG, Jing YU, Zhongliang WANG (2005)**].

Text clustering is the important technique of the text mining. Feature Extraction and Feature Selection are techniques which are used in text clustering. The text categorization is done by the feature extraction. The Author discusses the four methods for the feature extraction. The methods are document frequency (DF), term contribution (TC), TVQ, and Term Variance (TV).

These methods are basically based on the Vector Space Model. The main concept of VSM is that combination of words having in text database is not important for the mining.

Now the Feature extraction is the concept of extracts useful information from the textual database using some mapping concept. The drawback of feature extraction method is if the result value do not correct then its cluster match is not good.

The Author Tuomo Korenius, Jorma Laurikkala, Kalervo Jarvelin, Martti Juhola describe the concept of lemmatization and stemming process for the Finnish textual document [**Tuomo Korenius, Jorma.Laurikkala, Kalervo.Jarvelin, Martti Juhola (2004)**].

In this paper, the author focus on the how to perform the word transformation into their base form for the document clustering in Finnish language. The Finnish language is the language which spoken by the Finland people. The word transformation has two methods stemming and lemmatization which is used for the convert words into root form.

The stemming method not used the concept of search key value. Stemming is also helpful reducing the words size. On the other hand lemmatization based on the lemma word which means base form is defined.

The Author Julie Beth Lovins is the first researcher which design and developed the first stemming algorithm for reducing the word into their root form [**Julie Beth Lovins (1968)**].

The basic functionality of the stemming is to transform the words into the base form of the word which is called the stem word. Stemming is useful for computation linguistic and information retrieval work.

The main problem in stemming algorithm is variation in spelling of the word. The stemming algorithm is used for maximize the importance of the word terms. The following figure shows the example of stemming.

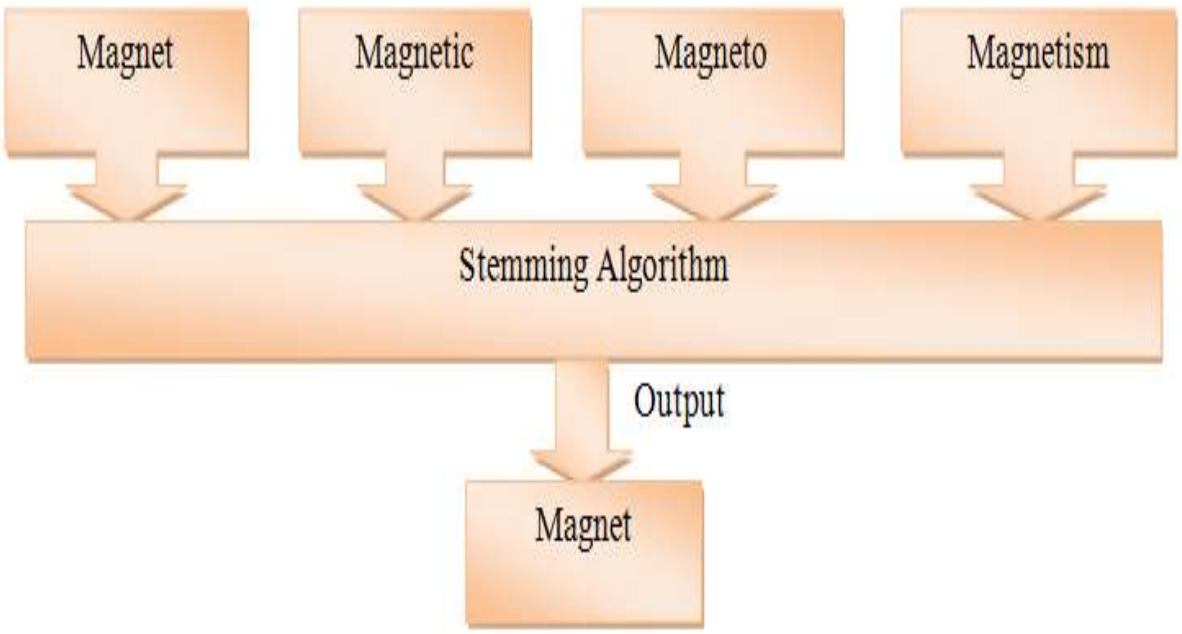


Figure 4Example of stemming pre-processing

The stemming algorithm having two approaches: the first approach is the stemming algorithms which fetch the stem word by removing the suffixes of the word. The suffixes remove by the matching the suffixes with list which is present in the system. The second approach is checking the spelling of the stem words.

The type of stemming algorithm is based on two main principles: iteration and longest match. The iteration principles is usually based on simply recursive procedure which delete the word strings with match to each other classes one at a time. The longest match is based on the principle of only one order class.

The working of this algorithm is determining the affixes then compiled and give ordered according their length. The longer match combination is not found so we take the shorter one as stem words. The all affixes combination is the first disadvantage of the longest match principle and the second one is it take the more space.

Chapter 3

PRESENT WORK

In this chapter, we divide into three sections. The section 3.1 we explain the problem formulation, in section 3.2 we explain the objective of the research and in section 3.3 we explain the methodology for the research.

3.1 PROBLEM FORMULATION

In our work, we are focusing upon the text mining which comes under the data mining field. The text mining is based on the deriving of knowledge from textual database or unstructured database.

The text database is the unstructured database which includes the database which is having textual format for example books, digital library, web pages, news articles, research paper, email message, etc. There are several technique used for the text mining which are text summarization, text clustering, text classification, text visualization, etc.

The text mining is useful analysis in the field of enterprise business intelligence, e-discovery, national security, scientific discovering, sentiment analysis, search and information access, social media monitoring.

Our research work problem is enhancing the stemming algorithm in text mining. We decide the research problem on the basis of general area of interest that is data mining. Initially, the problem was started in a broad general way that is text mining. When we study about the text mining then we find out the stemming step during pre-processing. The stemming is pre-processing step before the performing text mining in textual database. Stemming is used for transforming the words into their base form or root form. For example we have takes the three words which are following: connecting, connection, connective are transform into connect root word. Stemming pre-process used before the text mining, it save the time and space during the mining and generate the effective mining result.

In the research work we have proposed the new algorithm for the stemming. The proposed algorithm is TPLMSA Two Phase Longest Match Stemming Algorithm. It is based on the longest match principal in term of two ways. In first phase we match the word from left to right and in second phase we match the word from right to left. Then we

finally generate the effective result for the stemming in final phase. The details working of the algorithm we have discuss into the section 3.3 which is methodology section.

3.2 OBJECTIVE

The main objective of the research is focus on enhancing the stemming algorithm. The enhancement of stemming algorithm it leads to improve the overall performance of text mining. The pre-processing process is required to be performed before text mining.

The pre-processing is divided into three steps:

- Tokenization
- Stop words
- Stemming

The Splitting of sentence into many words or tokens is called the tokenization.

The Stop words are word which is not useful for the mining purpose so we need to remove such words form the textual database. The stop words like a, as, an, for, the, of, on, etc. Stemming is the concept through which we transform the words into their base form or root form.

Our Research works is focuses to enhance the stemming algorithm for performance improvement in text mining. Our aim in this research is to produce the effective result for the stemming process.

3.3 METHODOLOGY

In methodology section we have discuss the research methodology, Flowchart for TPLMSA Two Phase Longest Match Stemming Algorithm, Algorithm for TPLMSA and development tools.

3.3.1 RESEARCH METHODOLOGY

3.3.1.1 Defining Research Problem

Our research work problem is enhancing the stemming algorithm in text mining. We decide the research problem on the basis of general area of interest that is data mining. Initially, the problem was started in a broad general way that is text mining. When we study about the text mining then we find out the stemming step during pre-processing. The stemming is pre-processing step before the performing text mining in textual database. Stemming is used for transforming the words into their base form or root form.

For example we have takes the three words which are following: selection, selecting, select are transform into select root word.

3.3.1.2 Literature Survey

The literature survey is the process of study the problem in deep way. When the broad area of problem is formulated text mining then we focuses the literature survey in details. We find out during the text clustering it is necessary to perform the stemming step for performing the text mining on textual database. The literature survey information we got the information from various journals, conference preceding, books, etc.

3.3.1.3 Formulation of Hypothesis

The formulation of hypothesis we gain the idea from discussions with dissertation mentor, experts, and colleagues about the problem, its origin and objective. Our research work hypotheses based on try to enhance the stemming algorithm. In dissertation we will try to enhance the stemming algorithm in term of effective result.

3.3.1.4 Research design

Once the research problem has been formulated then we need to design the research in respect to sample. The aim of research design is to create the conceptual structure of the research.

3.3.1.5 Collect the data and Execution

For conducting the research we need to gather the data from multiple textual databases for performing the stemming. Once the data is collect then we start the execution of the research work.

3.3.1.6 Analysis of Research and Hypothesis Testing

After the execution we analyze the research work whether it is appropriate or not then we perform the hypothesis testing.

During the analysis we analyze the stemming algorithm in term of effective results produce by the algorithm.

3.3.1.7 Reporting

Once the all step are completed then we create a report for research. Report is the documentation which contains the overall information about the dissertation.

3.3.2 FLOWCHART FOR TPLMSA

The following flowchart shows the concept of TPLMSA two phase longest match stemming algorithm.

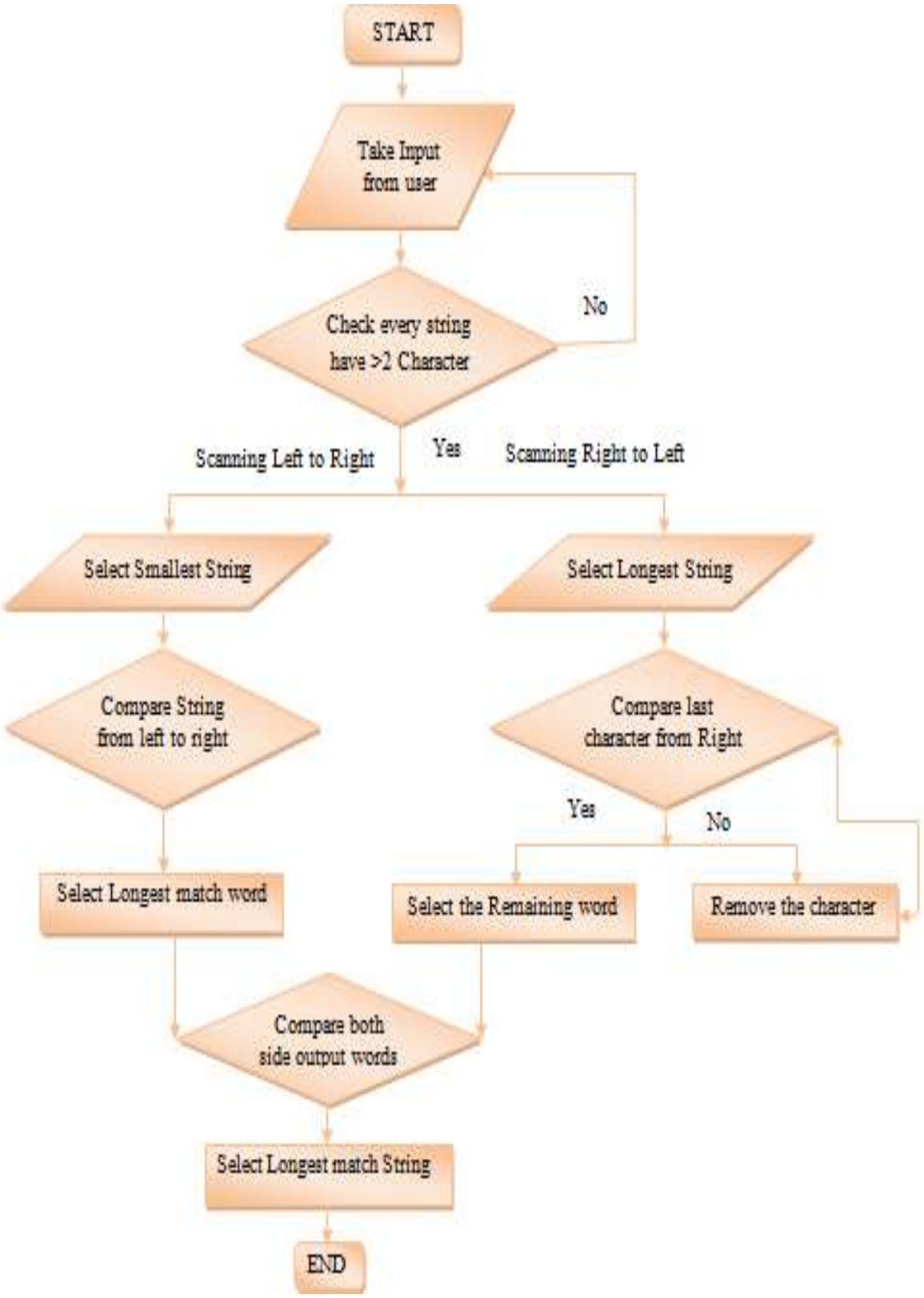


Figure 5 flowchart of the TPLMSA algorithm

In the Research Methodology we propose the new algorithm for stemming. It is used for the transforming the multiple word into their base form. When we perform the mining techniques like text clustering then we need cluster for the miming. The pre-processing output is not effective then our result will be affected. So we achieve effective cluster value through improve the stemming step in the text mining.

In this research we propose the TPLMSA Two Phase Longest Match Stemming Algorithm for the stemming.

3.3.3 ALGORITHM FOR TPLMSA

Now we discuss the working of TPLMSA two phase longest match stemming algorithm.

Procedure TPLMSA (Words, SelectedStr, LHSString, LHSResult, First, Largestring, Second)

Descriptions

Let Words= String for stemming

Let SelectedStr= String of minimum Length

Let LHSString= Array for words

Let LHSResult= Result for first phase

Let First= parameter for final from first

Let largestring= String of maximum Length

Let second= Variable for final right to left scan string

Begin

// count the character of each input string

1. Check every string having > 2 character

// process for first phase (scanning and compare the input string from left to right)

2. FIRST PHASE (left to right)

   i. Select smallest string from among them: SelectedStr=small
   ii. Allocate the remaining string into the array LHSString
   iii. Compare string from left to right

```
For (j=0; j<LHSString.Lenght; j++)
If (LHSString [ j ] == SelectedSts [ i ])
    LHSResult[k] = LHSString [ i ]
    k ++
Else
    Break
```

iv. Select the LHSResult string from left to right scanning and store in first.

// process for second phase (scanning and compare the input string from right to left)

3. SECOND PHASE (right to left)

i. Select longest string from among them lorgestring=long
ii. Allocate the remaining string into array words
iii. Compare string from right to left

```
M=1
For (I = 0; I < largestring.Lenght; I ++)
  For (j= 0; j<words.Lenght; j++)
        If(words[words.Lenght-m]== largestring [ largestring.Lenght-m ] )
    str = str + largestring
    Break
  m ++
```

iv. Select the str from right to left scanning and store into second.

// final process for the effective result

4. Compare the first and second result then select the longest match string among them for the final result.

End procedure

The TPLMSA stemming algorithm is divided into two phase. In the first phase we compare the words from left to right. In this phase we select the smaller words and then compare with rest other. In second phase we select the longest word and compare last character of the longest word with the remaining words if the character match then select the second output if not match then eliminate the last character from the words and compare again. After the successful receiving of first and second result we need to compare again for the final effective result.

3.3.4 DEVELOPMENT TOOLS (VISUAL STUDIO 2012)

An algorithm TPLMSA for stemming is developed in the C# language. The platform for development we used the visual studio 2012. Visual studio is an IDE (Integrated development environment), which was developed by Microsoft. It is basically used to develop the web application, computer program and web services.

The visual studio having the strong features which is code editor supporting the Intellisense and code refactoring. Visual studio supports the many programming languages. It includes C, C++, C#, VB.NET, M, PYTHON, RUBY and HTML, JAVASCRIPT, CSS etc. The Express Edition provide by Microsoft at no cost.

3.3.4.1 Feature

3.3.4.1.1 Code Editor

The code editor provides the interface for the user to write the program in easy way. Microsoft visual studio code editor includes the facility of code completion (Intellisense) for variables, queries, functions etc. and syntax highlighting.

3.3.4.1.2 Debugger

Visual studio debugger support the both source level debugger and machine level debugger. The debugger can allow to user set the breakpoints in the program which can helps the programmer to find out the error and check the all variable values during the execution.

3.3.4.1.3 Designer

Visual studio supports the designer tool according to development application. The user developed the window application then they used the window form designer. The visual

studio also supports the web designing using ASP.NET, HTML, CSS, and JAVASCRIPT.

3.3.4.1.4 Class Designer

The class designer is used for the user to modify and change the classes' member and function using the UML diagram.

3.3.4.1.5 Data Designer

The data designer supports the designing of queries using graphically. It also supports the manipulation of database schema using graphically which include type table primary and foreign key constraints.

3.3.4.1.6 Object browser

The object browser is the collection of the namespace and the class library browse for the .NET frameworks.

3.3.4.1.7 Solution Explorer

The development of the application we need the code file and other resources. The solution explorer is the collection of the code file and the other resources. The basically working of the solution explorer is to manage and browse the file in the solution.

3.3.4.1.8 Data Explorer

The data explorer is basically used for manage and controlling the database Microsoft SQL server. It includes the creation of table, function, procedure using TSQL queries. The data explorer also supports the debugging and Intellisense.

3.3.4.1.9 Server Explorer

The server explorer supports to manage the database connection instance. It's also support the running window services.

3.3.4.1.10 Properties Editor

The properties editor tools are useful for designing the form during the development. It includes the all properties variable objects classes, form and web pages. The properties can set by the GUI interface.

3.3.4.2 Supported Products

3.3.4.2.1 Microsoft Visual C++

Microsoft visual C++ is used for write and compiles the code of C and C++ languages. It supports the ISO C standards. We can also use the Visual C++ for windows API.

3.3.4.2.2 Microsoft Visual C#

The Microsoft Visual C# is used for the C# language in .net frameworks. It supports the development of the windows application. The Microsoft Visual C# includes the Visual Studio Class designer, Forms designer, and Data designer.

Chapter 4

RESULTS AND DISCUSSIONS

Our research work expected outcome is depending upon the effective result by the program during the stemming algorithm. The stemming algorithm is used in text mining. Before the text mining it is necessary to avoid the unnecessary and repeated words from the textual database for saving the space and time. It is basically used for transformed the words into their root form. The following figure 6 shows the memory size for stem word.

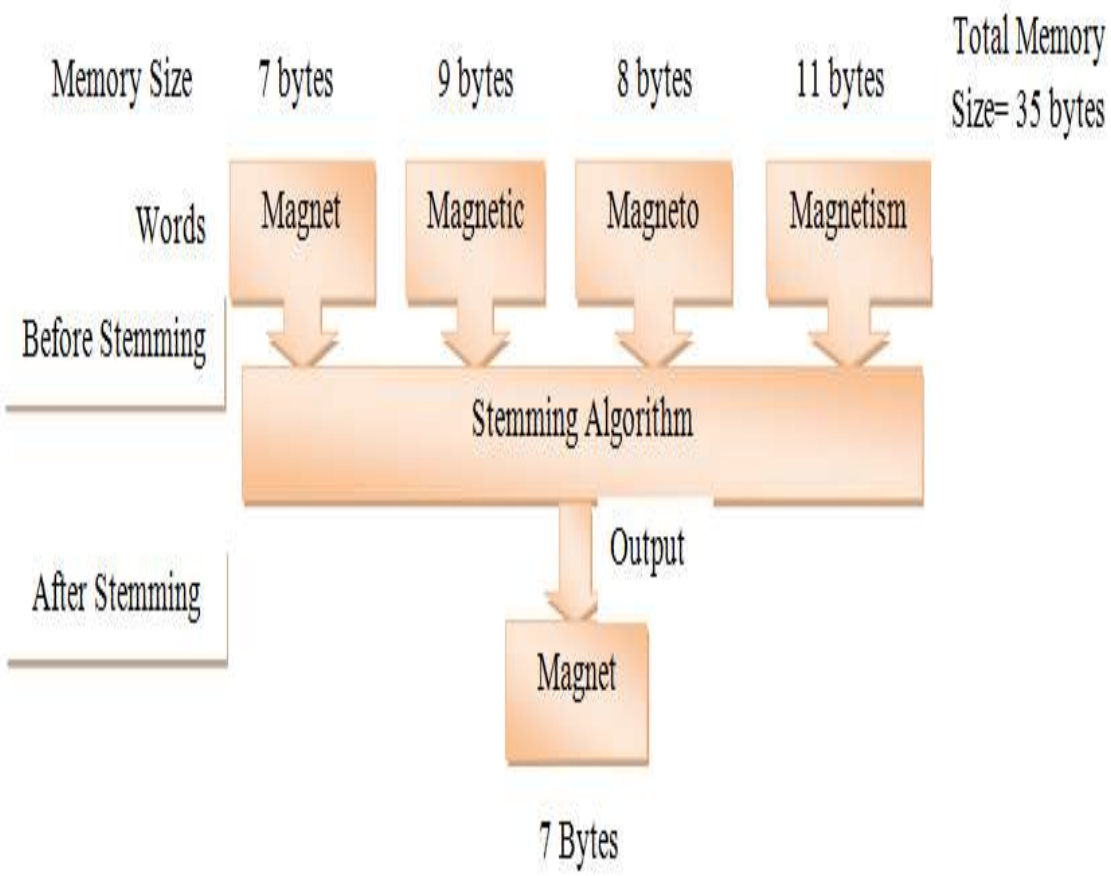


Figure 6Example memory sizes for stem word

In the existing Lovins algorithm they generate the stemming output on the basic of some principal. The first principal is stemming the words using longest match. The second

principal is after the stemming compare the result with some rule. The existing algorithm sometimes not gives the effective result. It is also take the more memory space for the result.

To overcome this problem they used the concept of revision of program, which take the extra process for the stemming result. So we proposed new algorithm for stemming TPLMSA Two Phase Longest Match Stemming Algorithm. It is based on two phase longest match principal. In the first phase, we match the shorter word with remaining words from left to right. In the second phase, we match the longest words with remaining words from right to left.

Now we take the some input from Lovins Dataset and compare with the proposed TPLMSA algorithm.

Table 1Result of TPLMSA on Lovins Dataset

INPUT	EXISTING OUTPUT	TPLMSA OUTPUT
Magnesia	Magnes	Magnesi
Magnesite	Magnesit	
Magnesian	Magnes	
Magnesium	Magnesium	
Magnetic	Magnes	Magnet
Magnet	Magnes	
Magneto	Magnes	
Magnetically	Magnes	
Magnetism	Magnes	
Magnetite	Magnetit	
Magnetitic	Magnetit	
Magnetizable	Magnes	
Magnetization	Magnes	
Magnetize	Magnes	
Magnetostrictive	Magnetostrict	

Magnetostriction	Magnetostrict	magneto
Magneton	Magneto	
Magnetomotive	Magnetomot	
Magnetos	Magneto	
Magnetometry	Magnetomotr	
Magnetometric	Magnetomotr	
Magnetometer	Magnetomoter	
Metal	Met	Metal
Metallic	Met	
Metallically	Metall	
Metalliferous	Metallifer	
Metallize	Metal	
Metallurgical	Metallurg	
Metallurgy	Metallurg	
Induction	Induc	induc
Inductance	Induc	
Induced	Induc	
Angular	Angl	Ang
Angle	Angl	

Now we show the execution of the Lovins datasets using the TPLMSA.

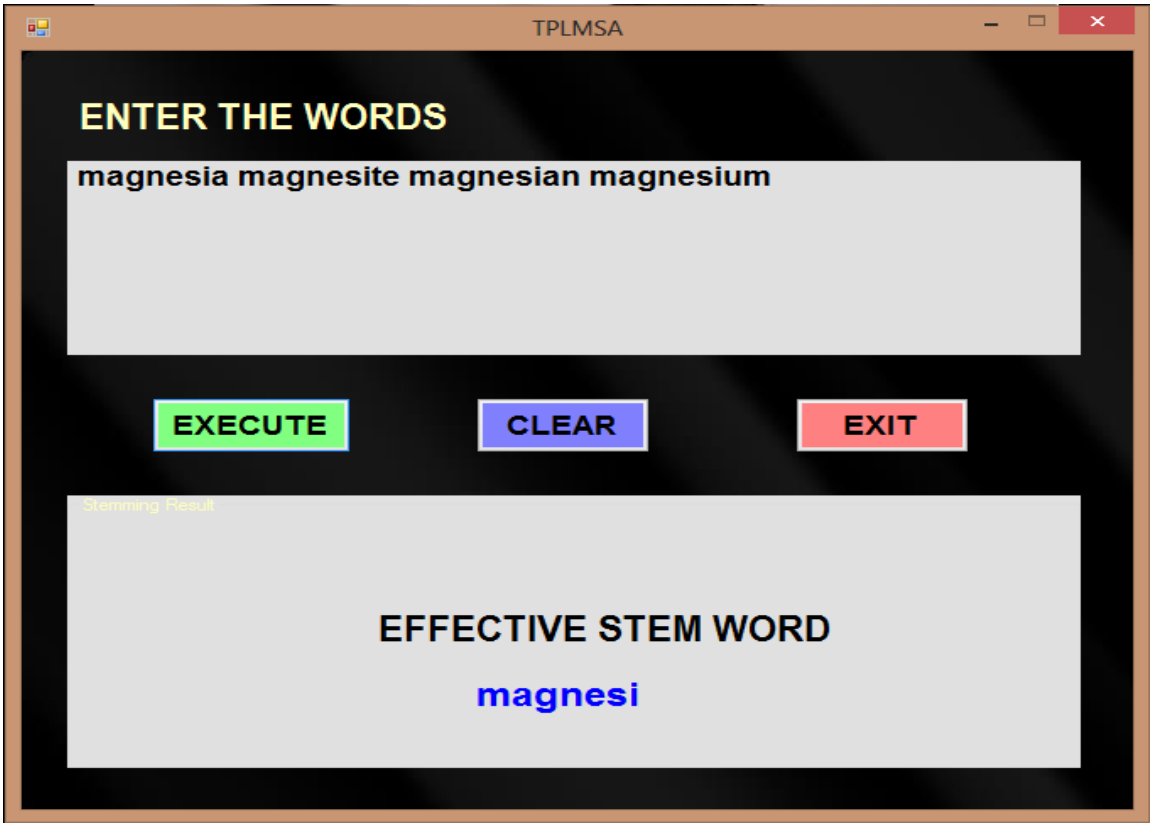


Figure 7 Execution Result of magnesi output

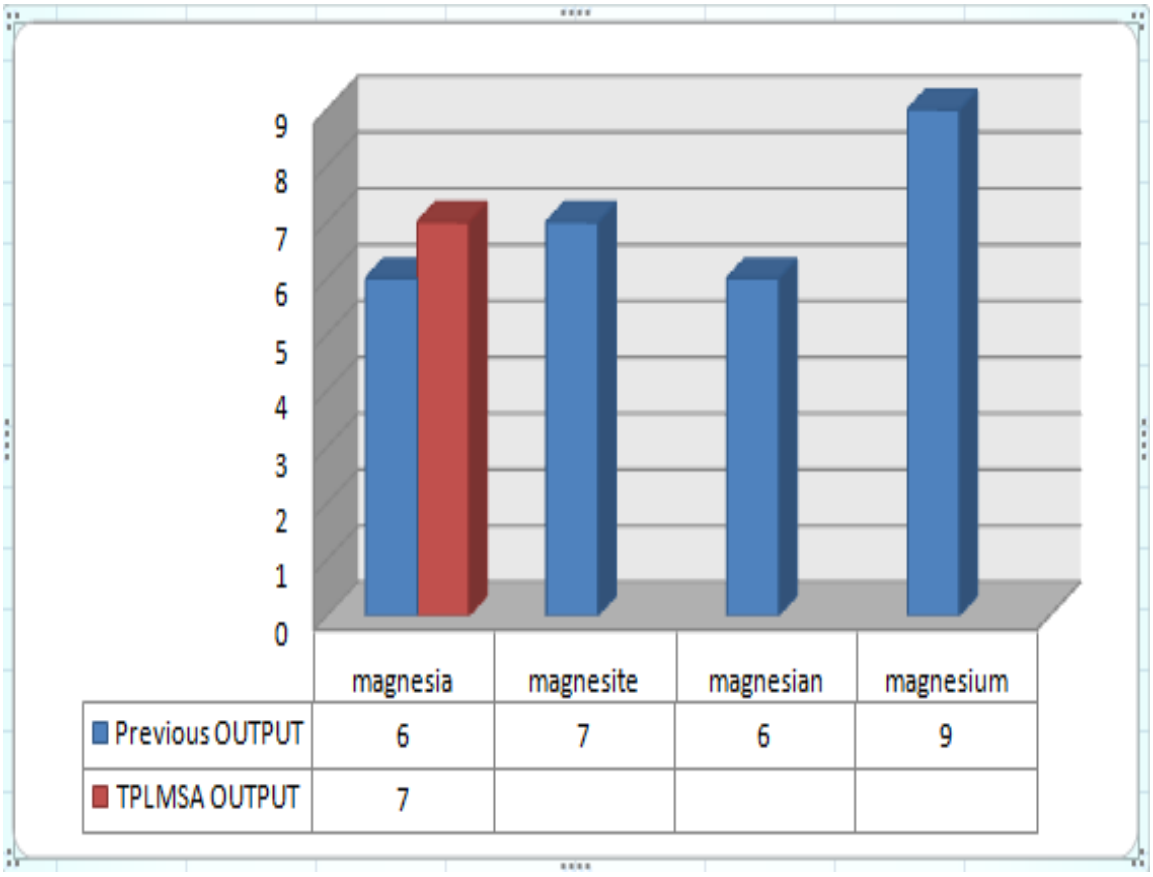


Figure 8 Graph Comparison of magnesi output

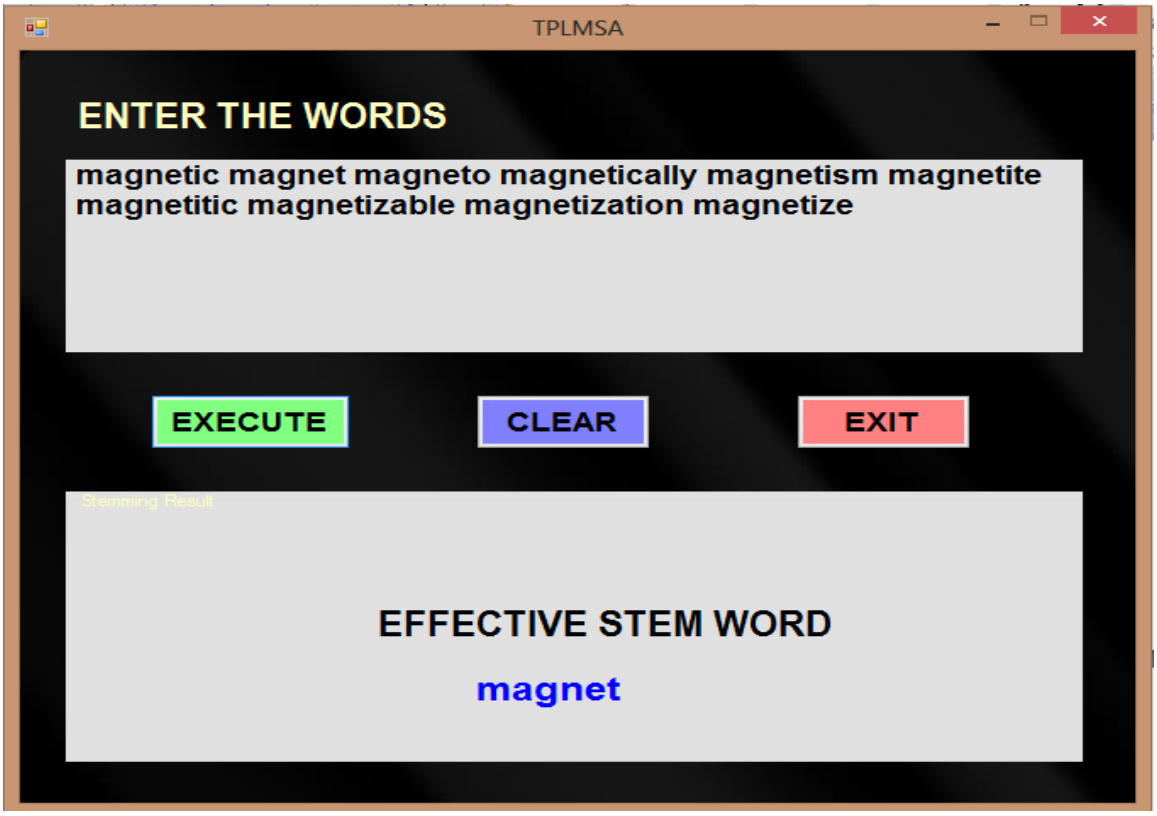


Figure 9 Execution Result of magnet output

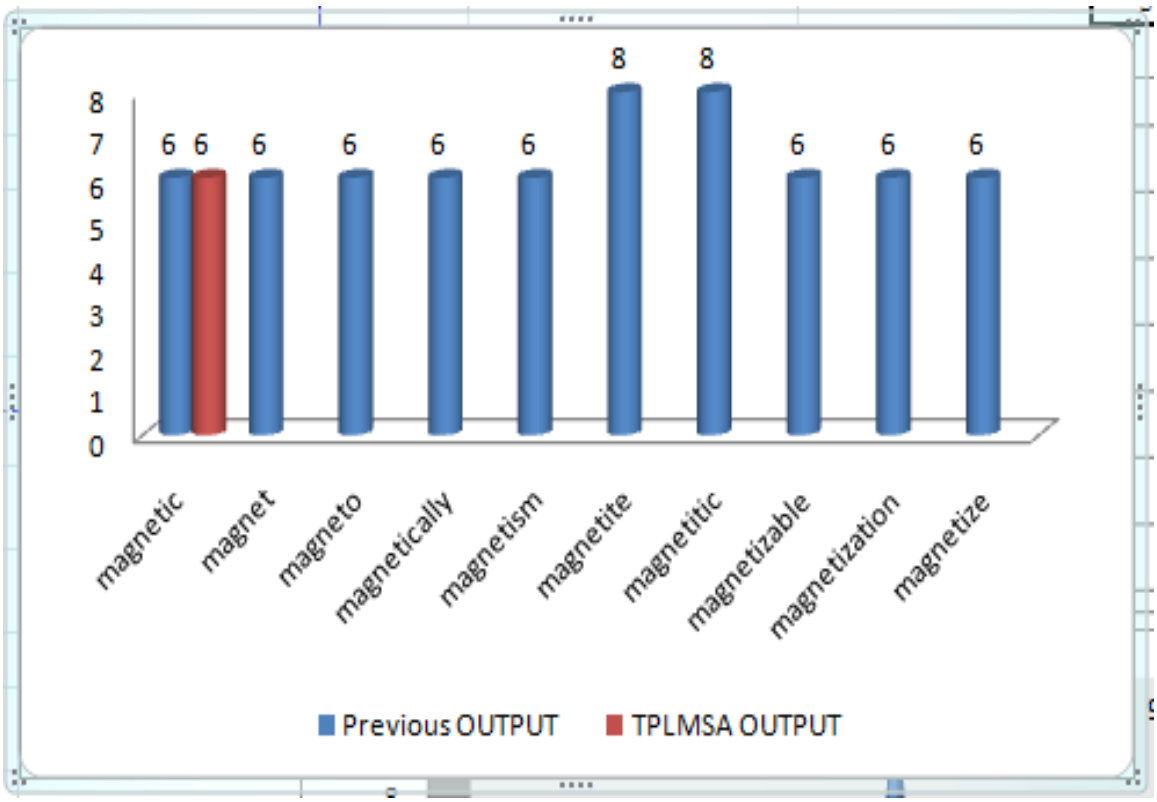


Figure 10 Graph Comparison of magnet output

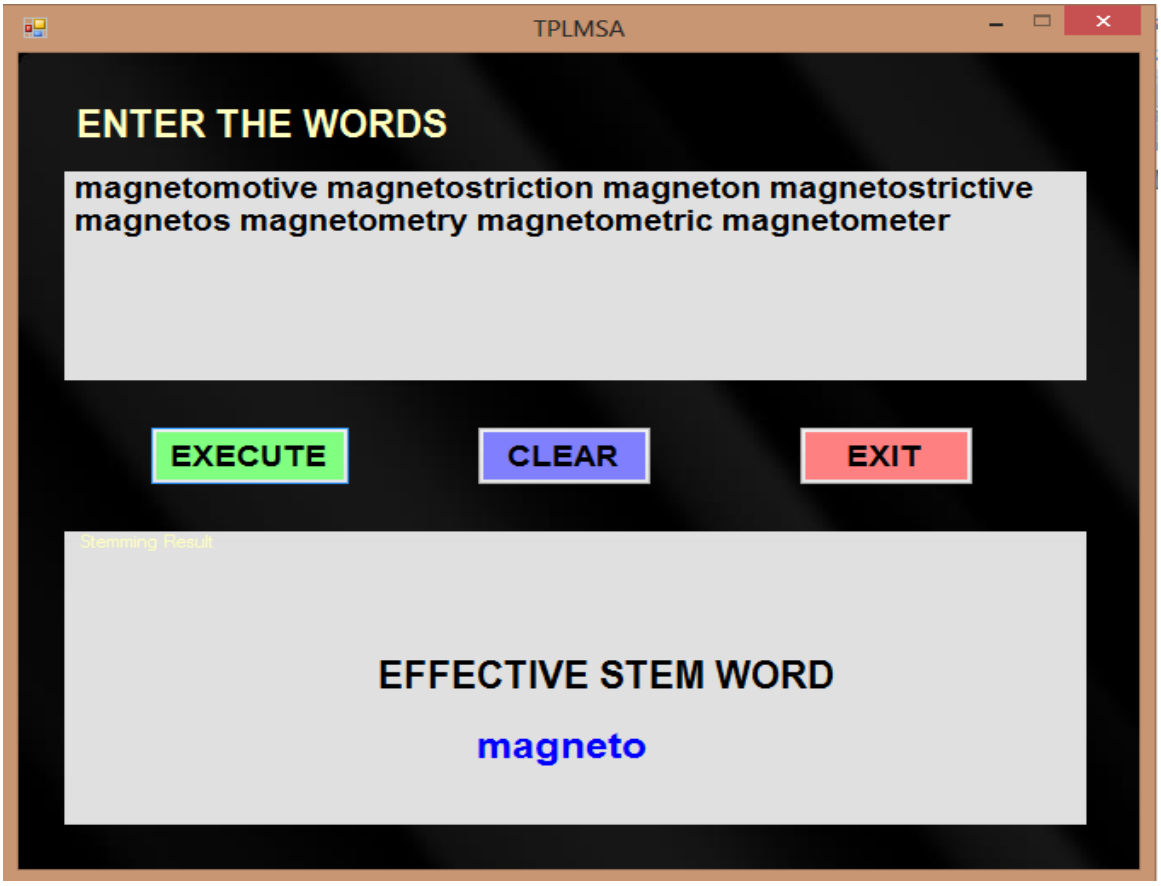


Figure 11Execution Result of magneto output

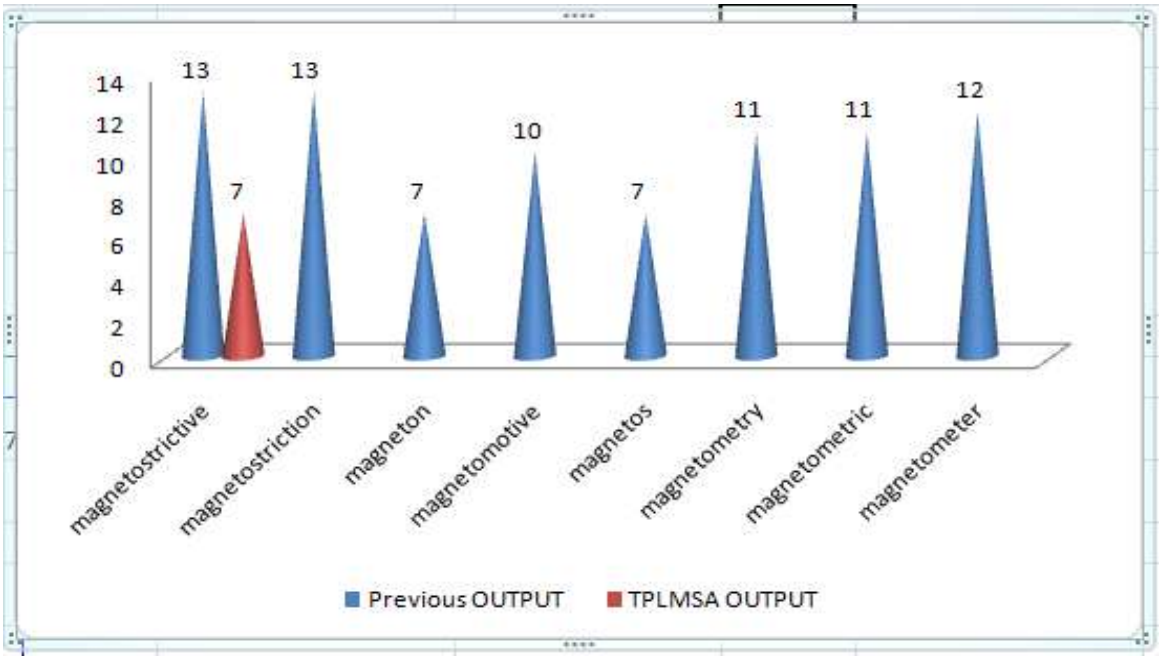


Figure 12 Graph Comparison of magneto output

Figure 13 Execution Result of metal output

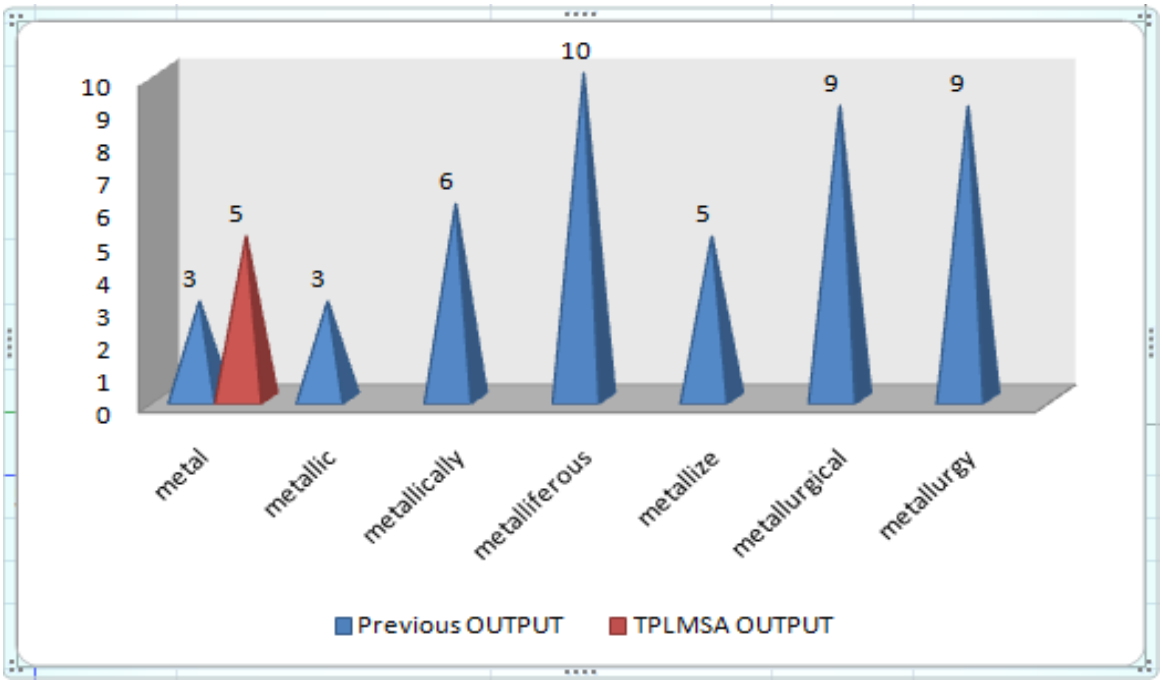


Figure 14 Graph Comparison of metal output

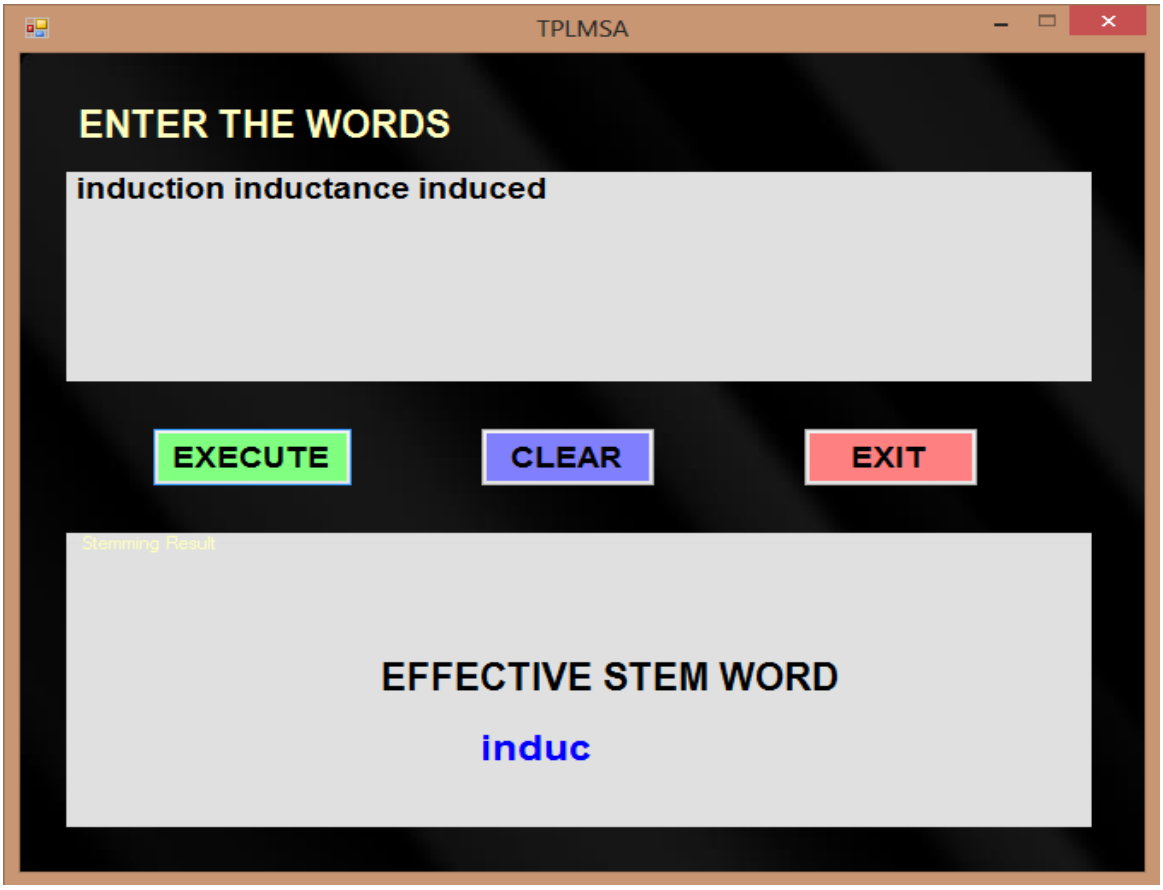


Figure 15 Execution Result of induc output

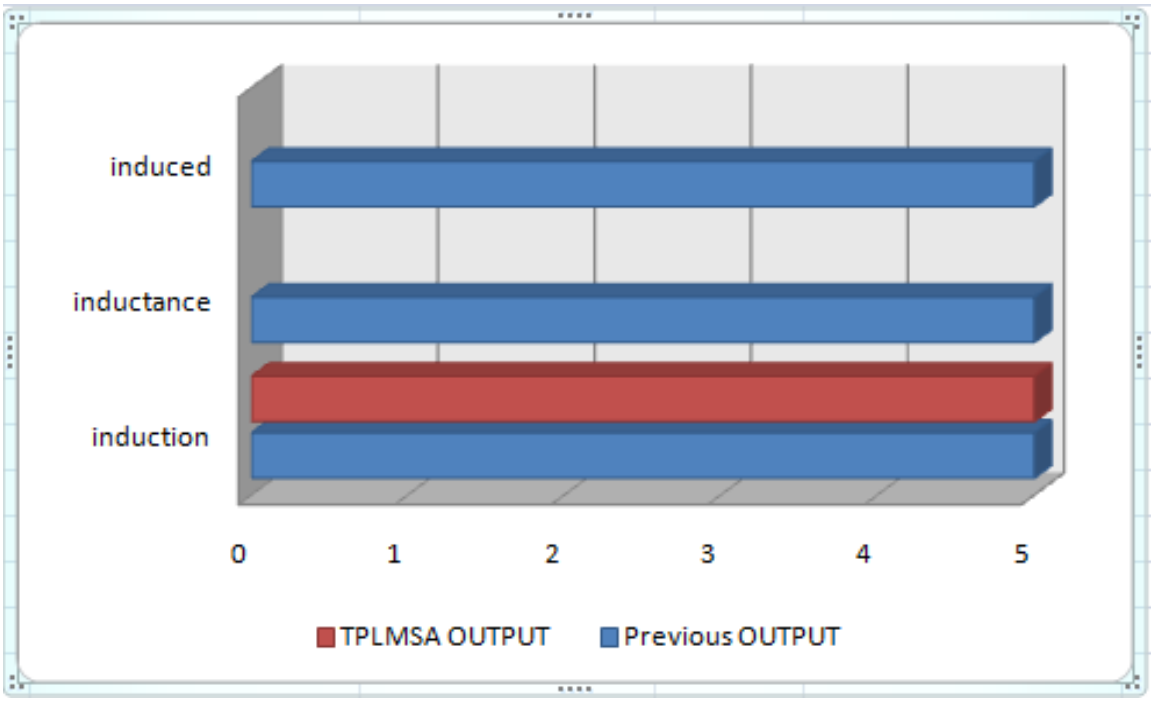


Figure 16 Graph Comparison of induc output

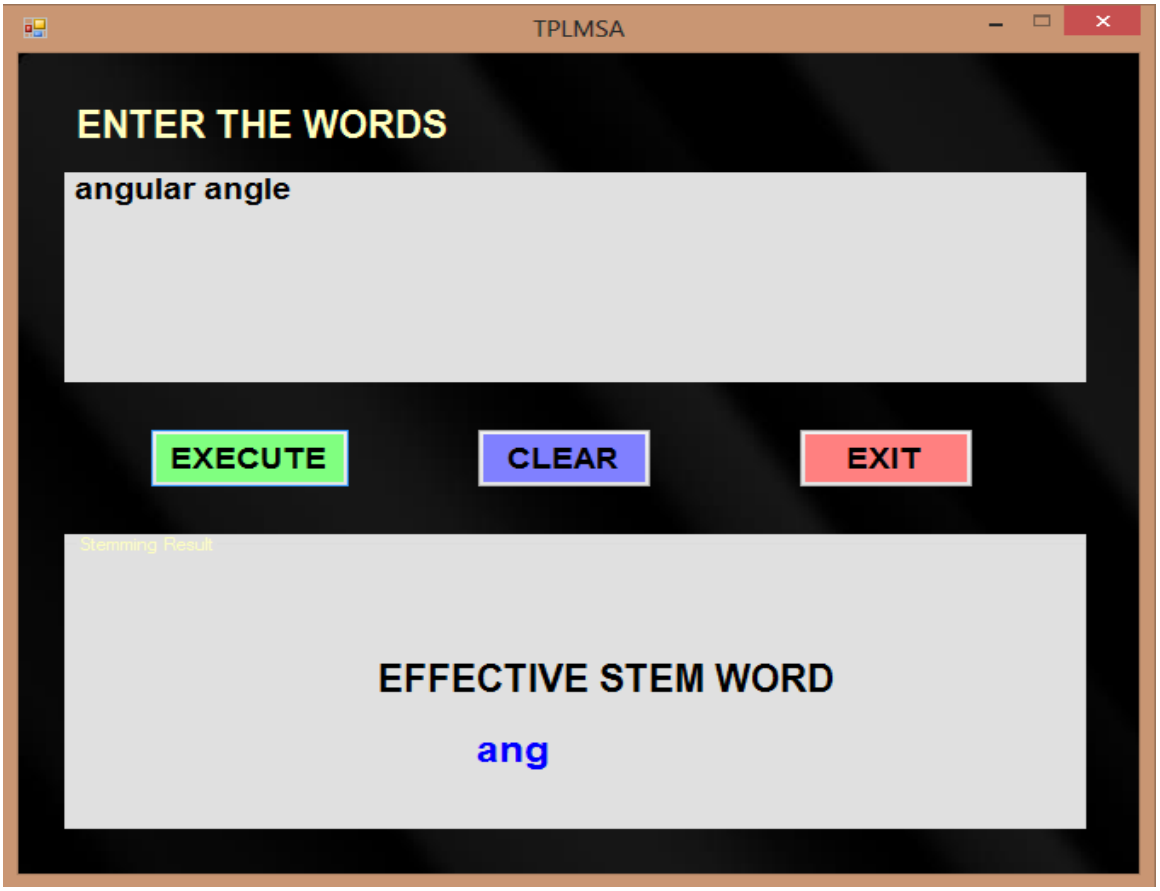


Figure 17 Execution Result of ang output

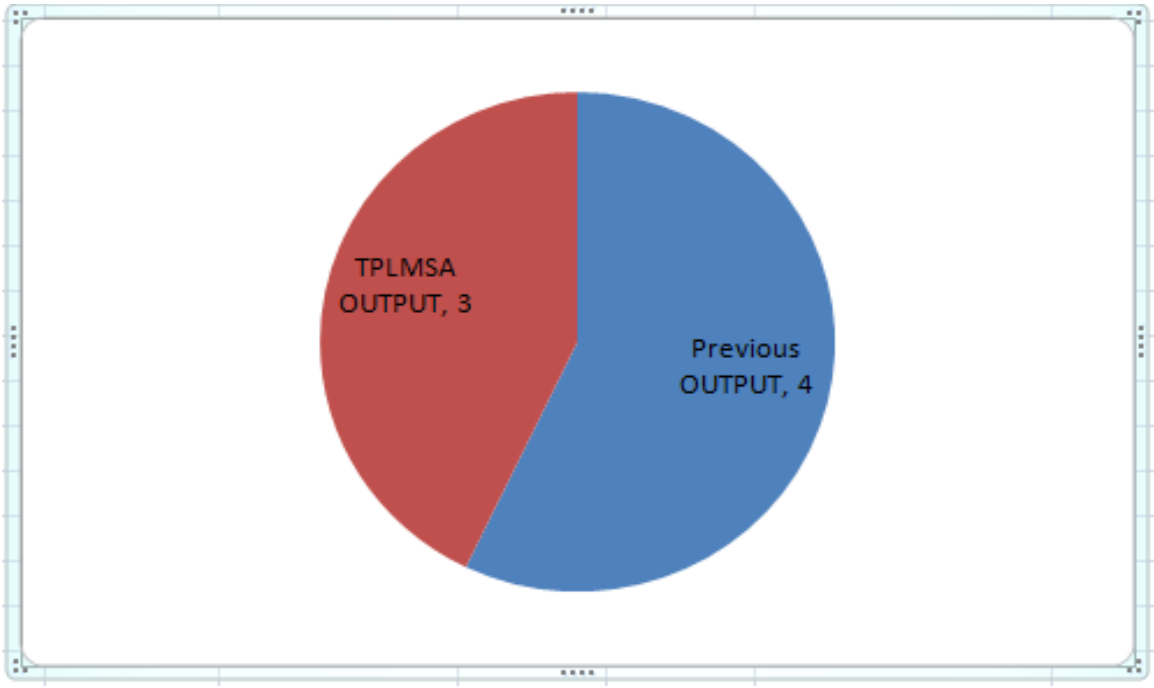


Figure 18 Graph Comparison of ang output

Now we further take the dataset from Kodimala, Savitha, "Study of stemming algorithms" (2010).

Table 2 Result of TPLMSA on Kodimala Dataset

INPUT FROM KODIMALA DATASET	OUTPUT OF TPLMSA
Shoes	Shoe
Shoed	
Thresholds	Threshold
Threshold	
Values	Value
Valued	
valuex	
seen	See
See	
Sees	
Kings	King
Kingdom	
Aeronautics	Aeronautic
Aeronautical	
substitutes	Substitute
Substituted	

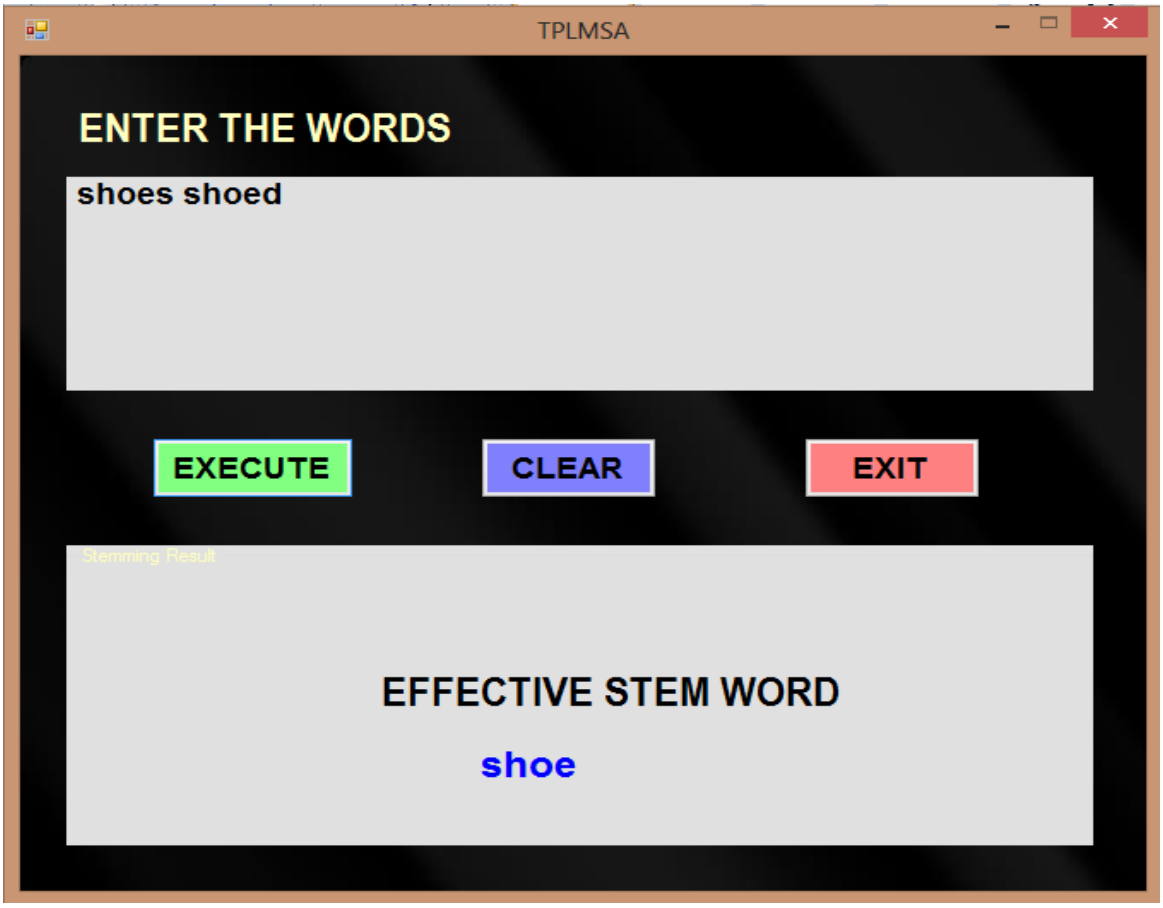


Figure 19 Execution Result of shoe output

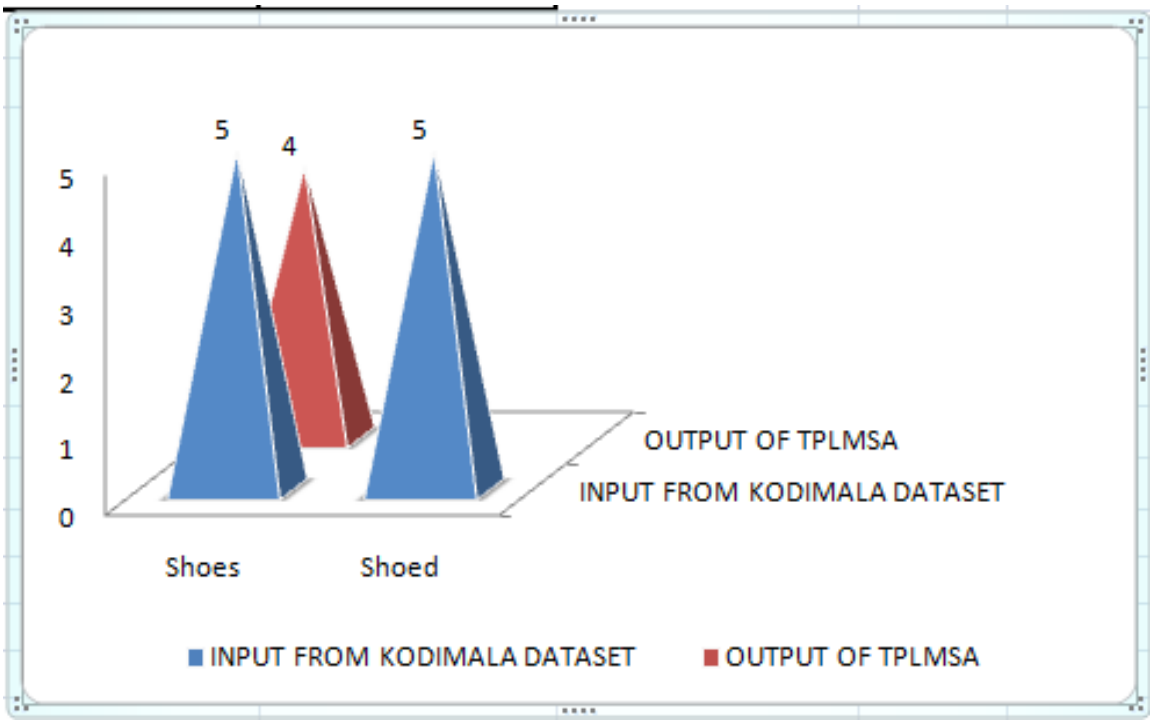


Figure 20 Graph Comparison of shoe output

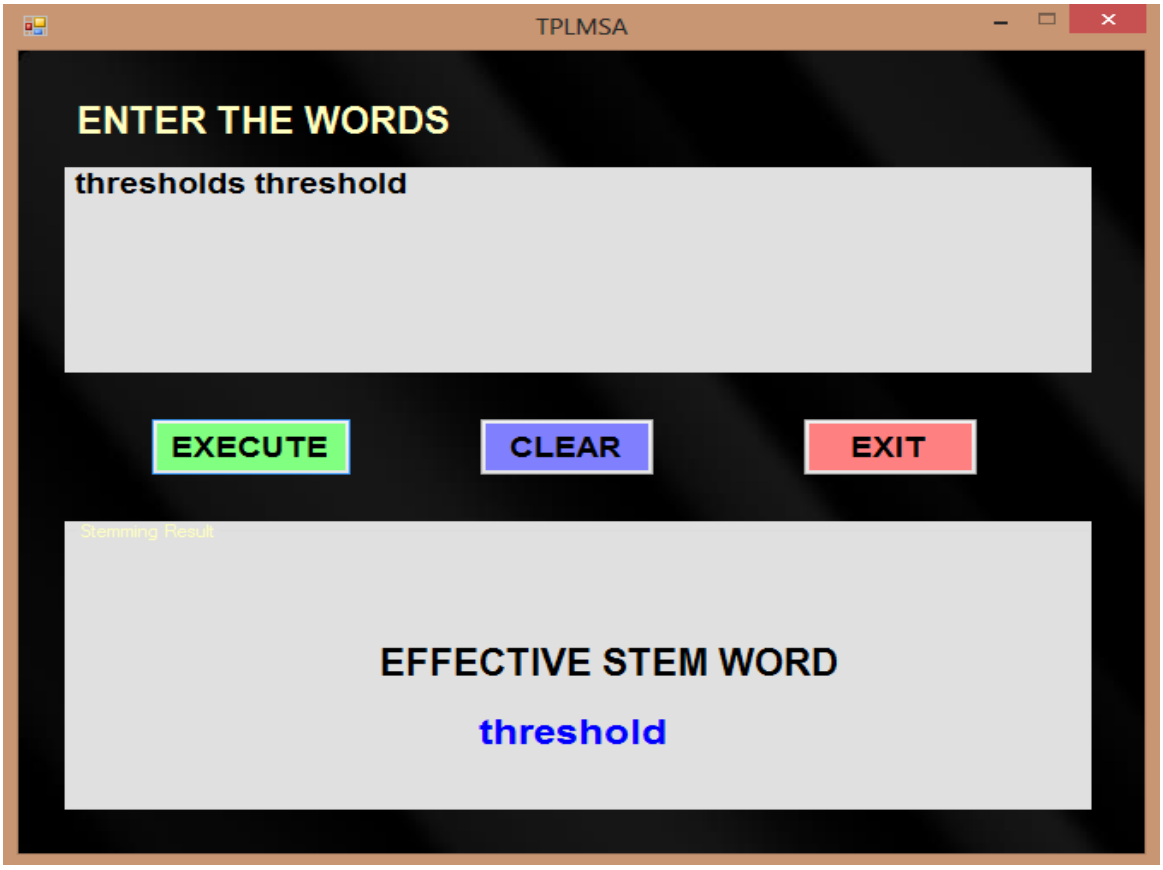


Figure 21 Execution Result of threshold output

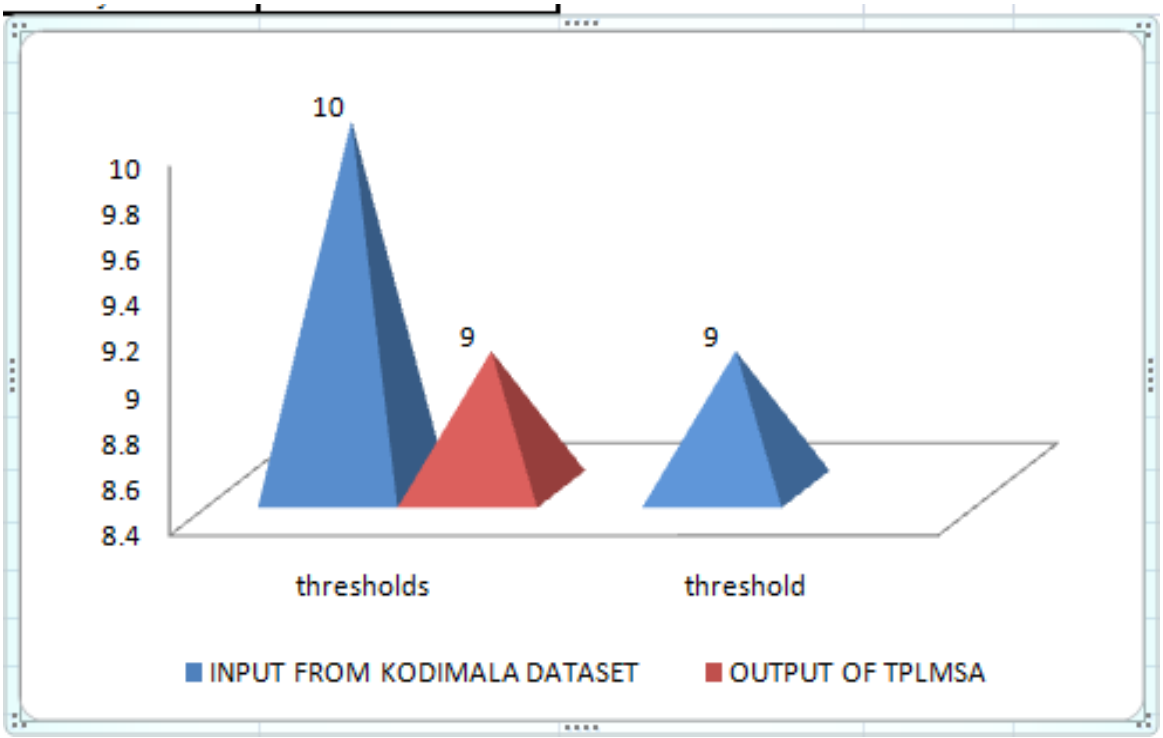


Figure 22 Graph Comparison of threshold output

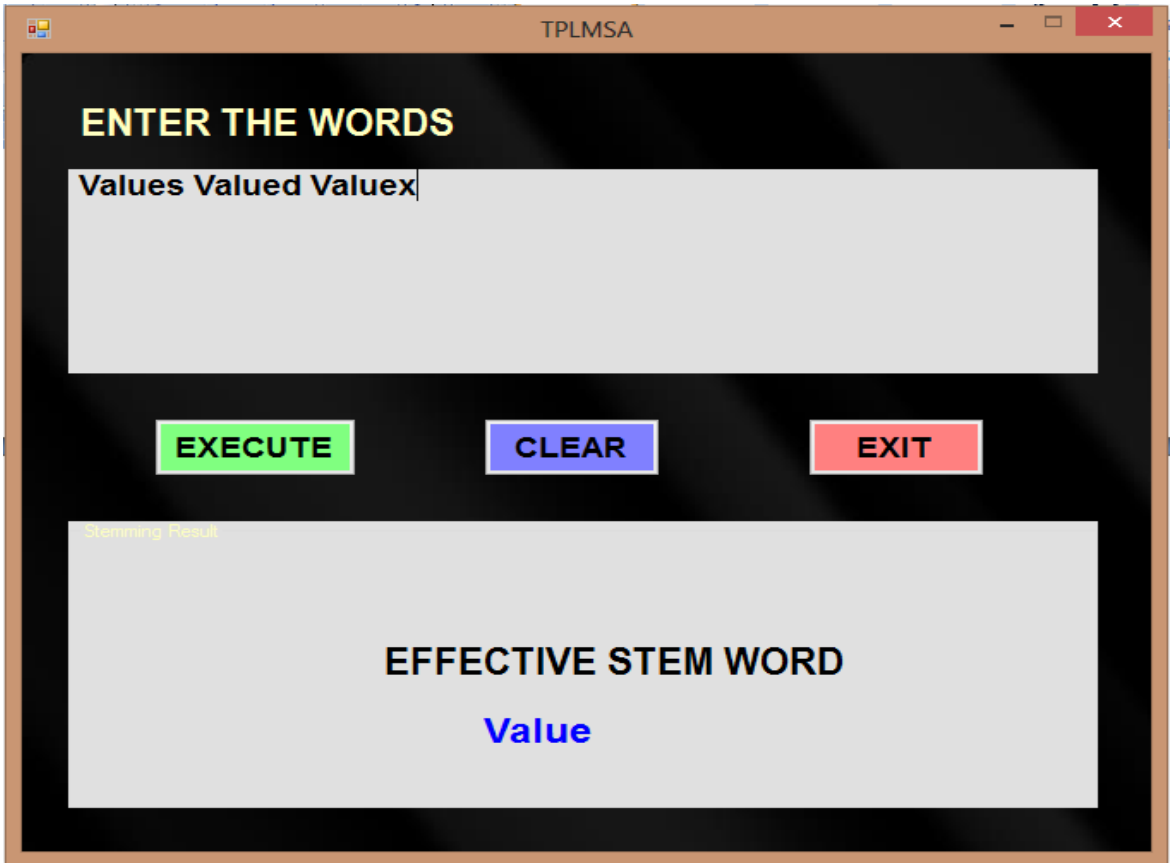


Figure 23 Execution Result of value output

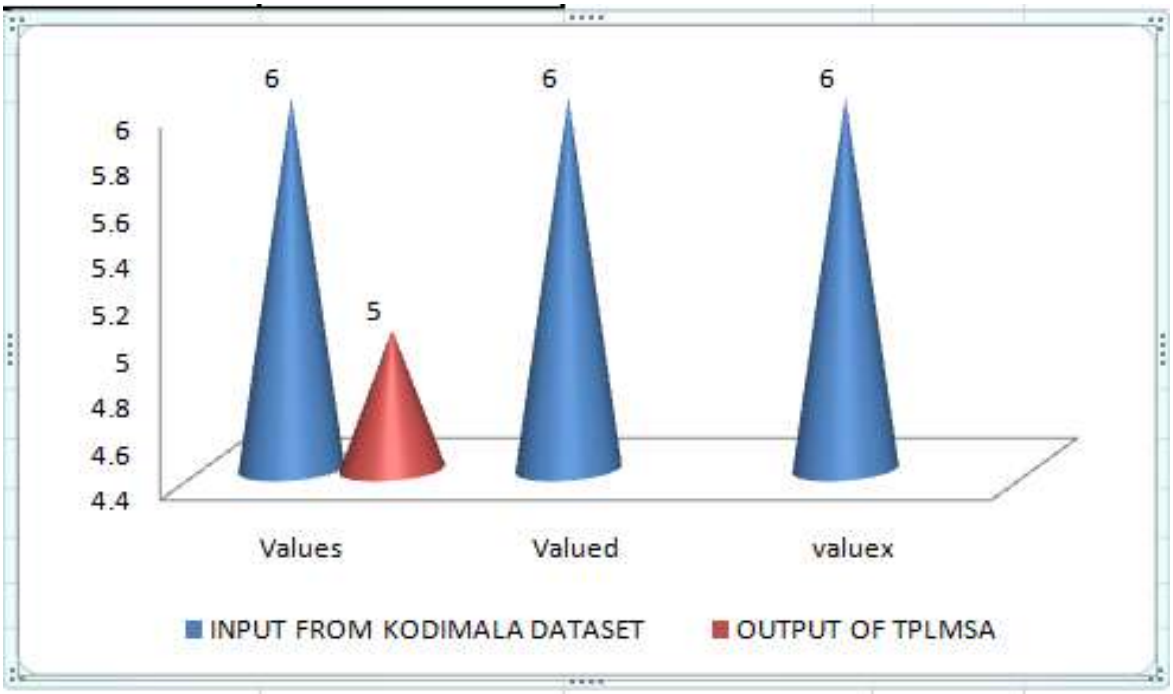


Figure 24 Graph Comparison of value output

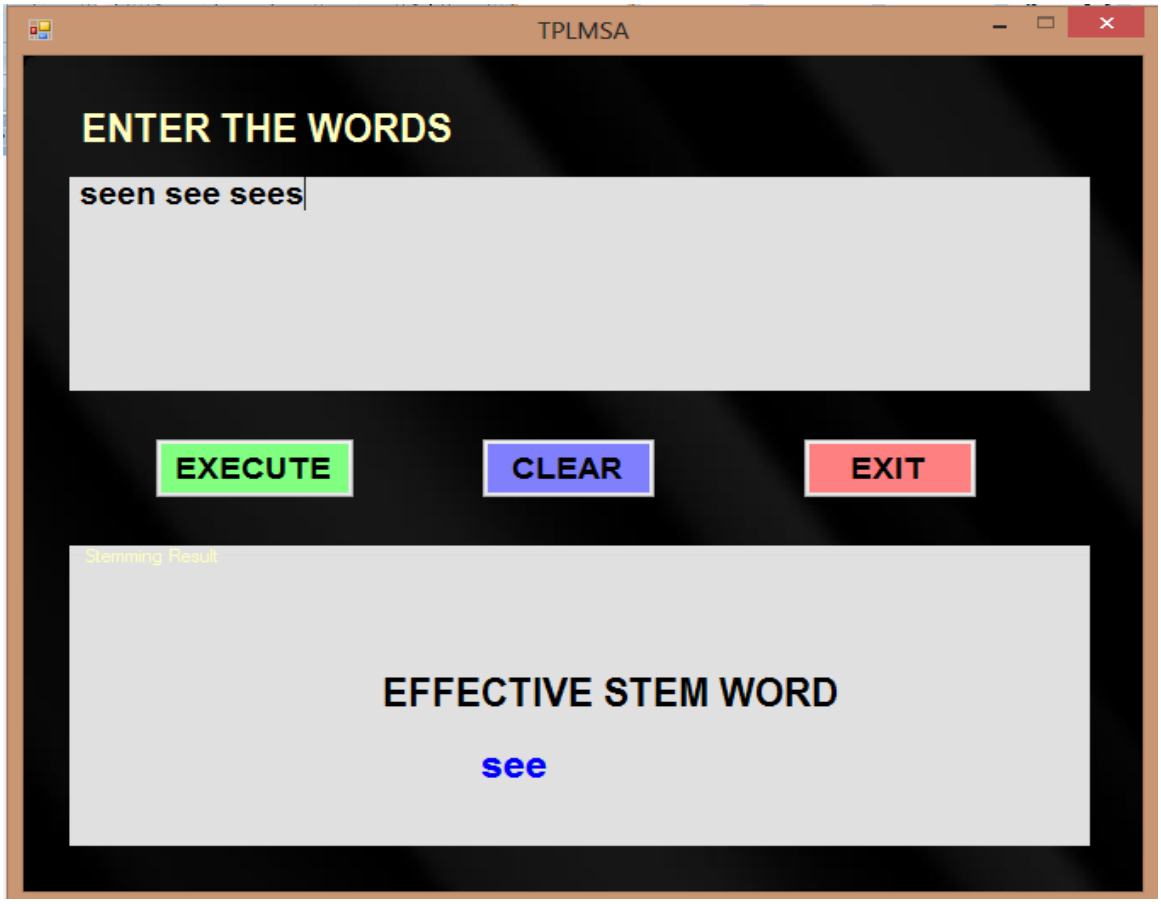


Figure 25 Execution Result of see output

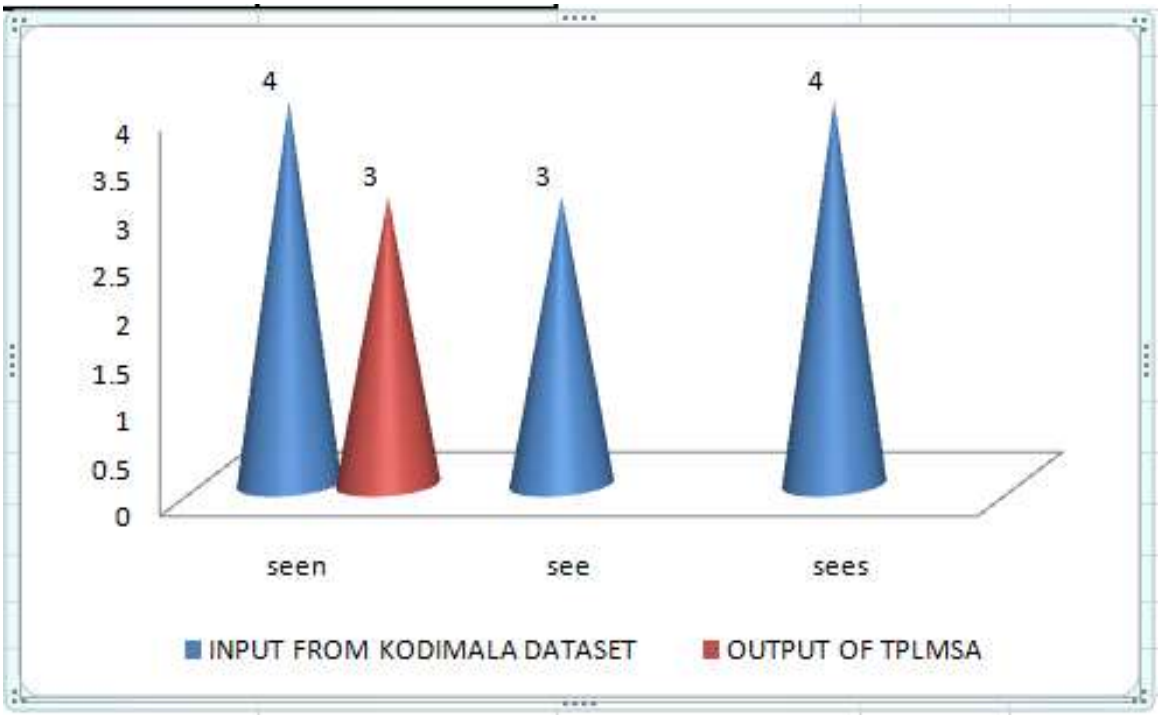


Figure 26 Graph Comparison of see output

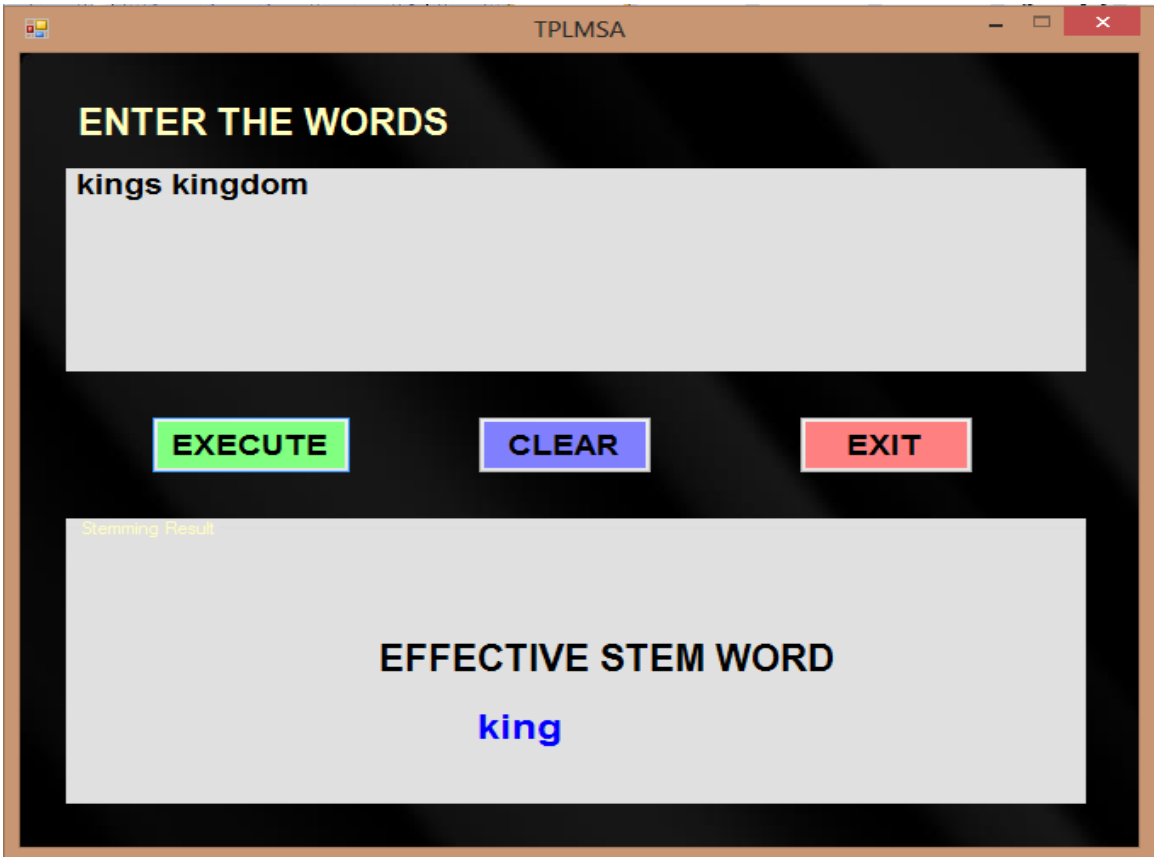


Figure 27 Execution Result of king output

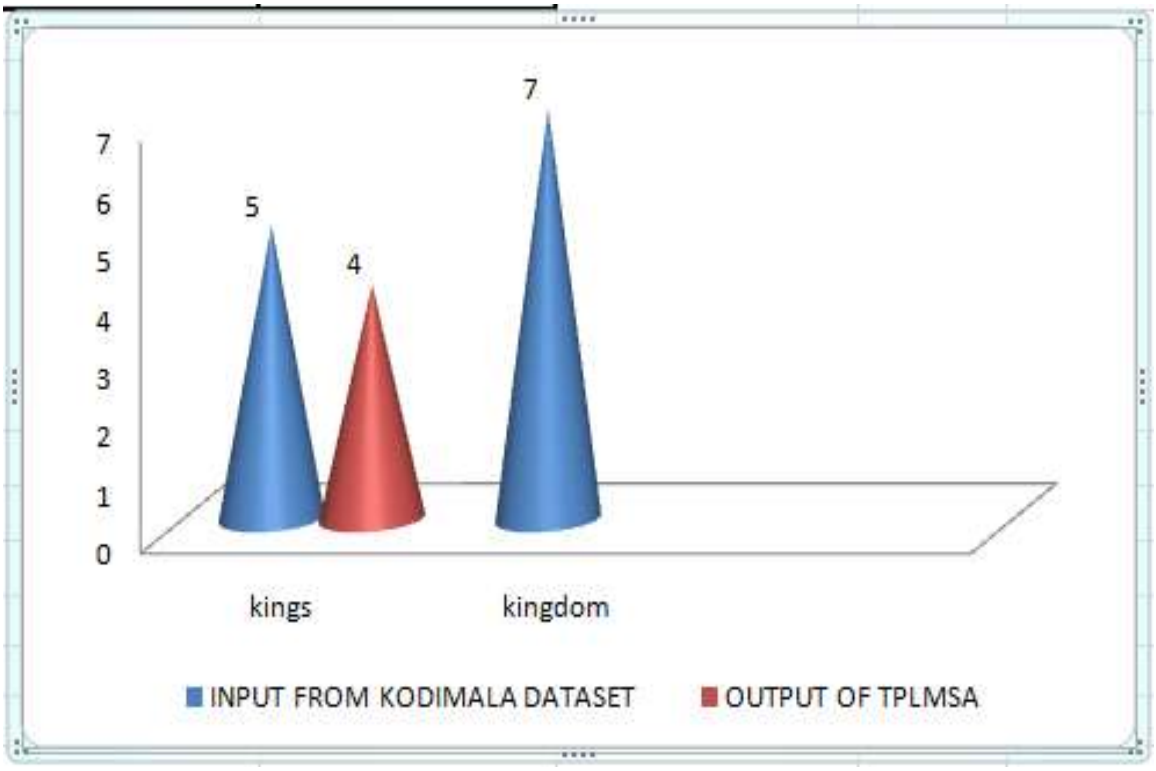


Figure 28 Graph Comparison of king output

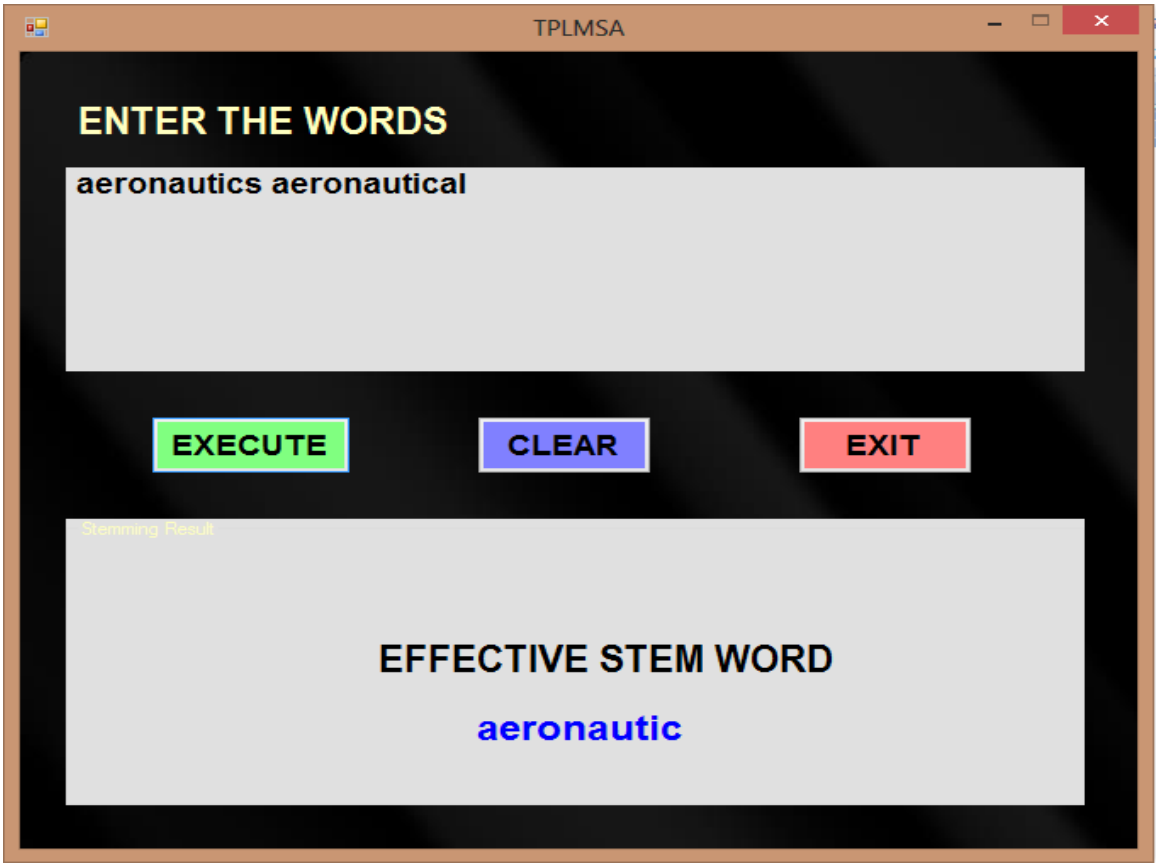


Figure 29 Execution Result of aeronautic output

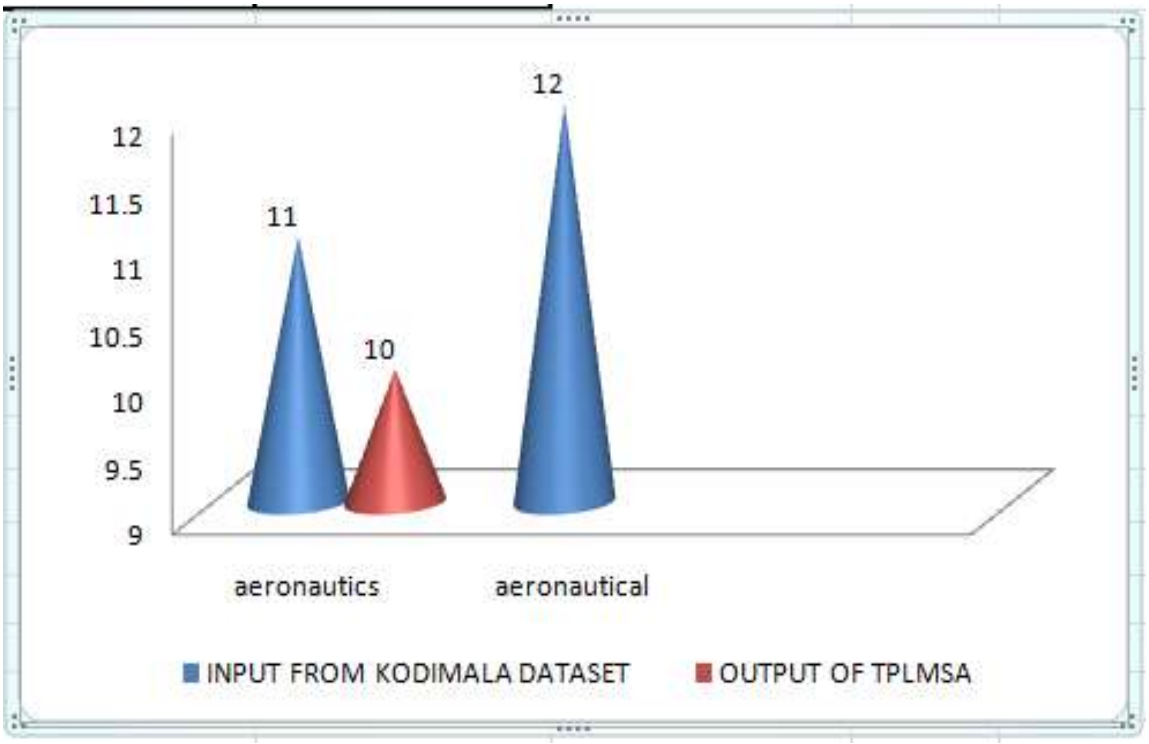


Figure 30 Graph Comparison of aeronautic output

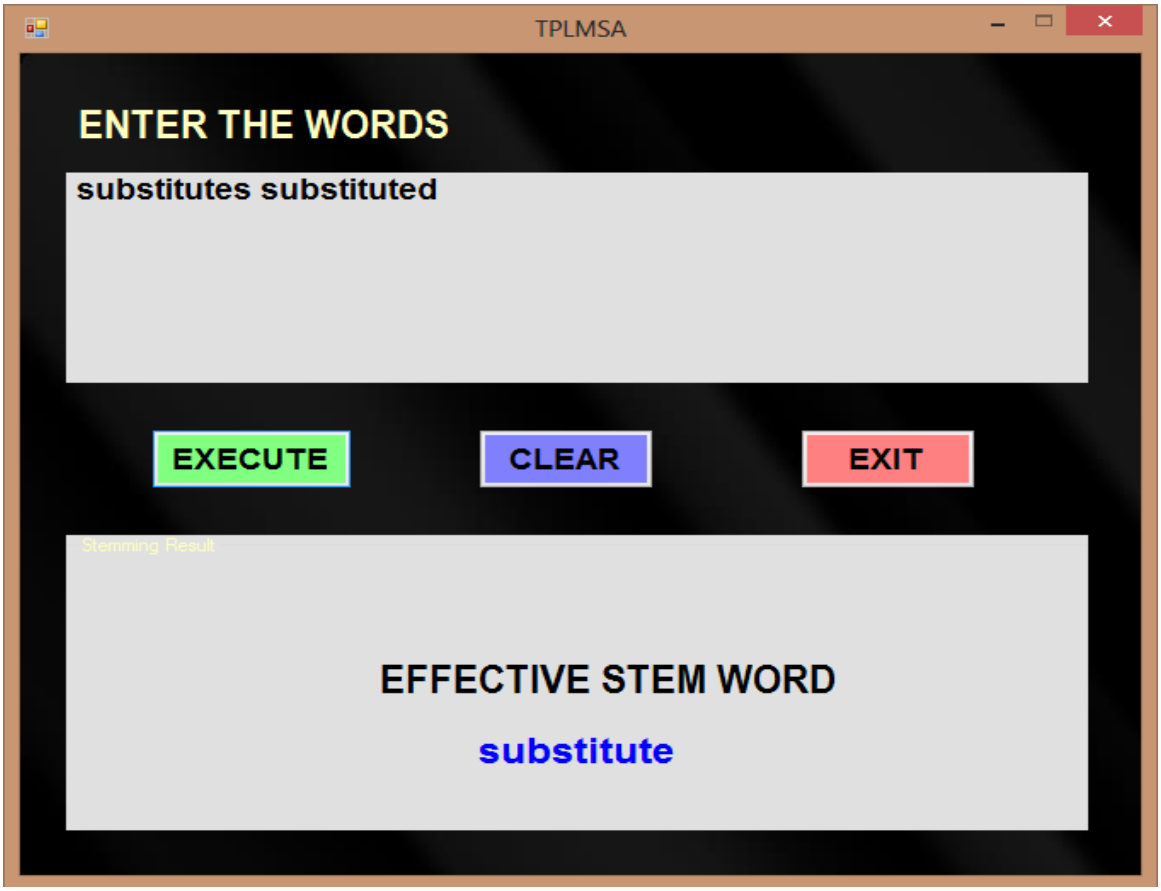


Figure 31Execution Result of substitute output

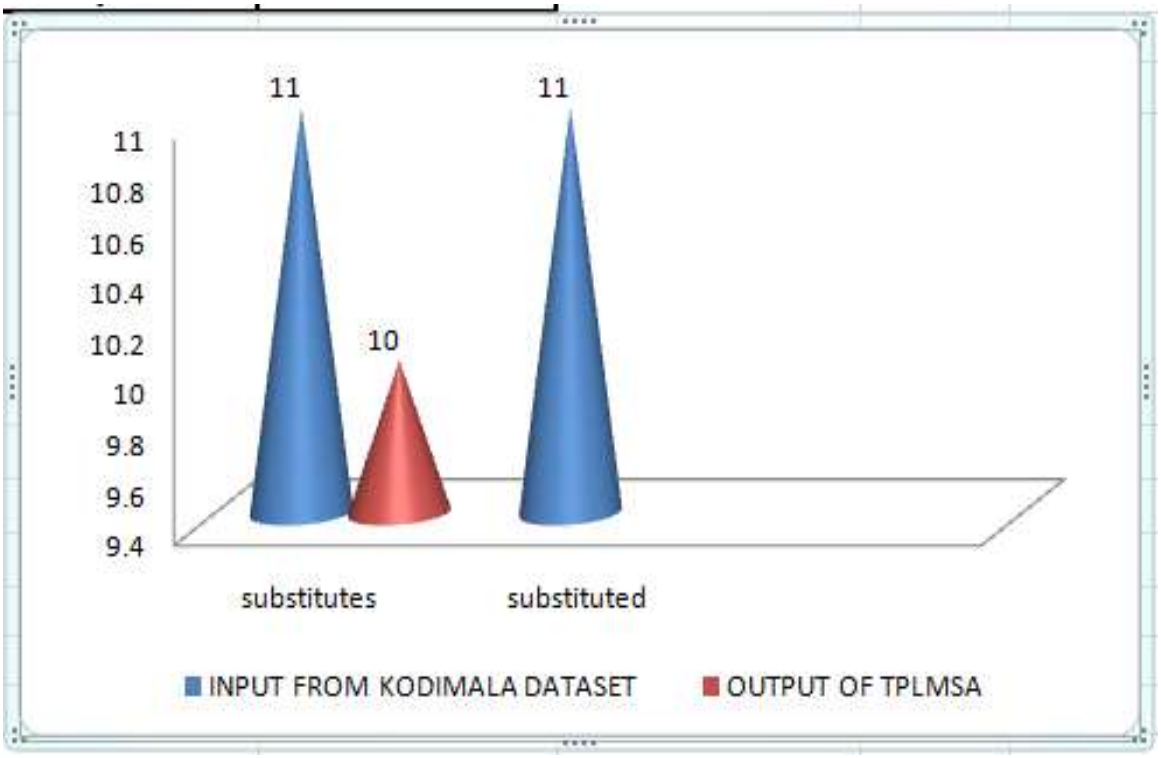


Figure 32 Graph Comparison of substitute output

We can also take the example from C.Ramasubramanian, R.Ramya Effective Pre-Processing Activities in Text Mining using Improved Porter's Stemming Algorithm.

Table 3 Result of TPLMSA on C. Ramasubramanian Dataset

INPUT	OUTPUT OF TPLMSA
Materially	Material
Materialize	
Material	
Materialization	
Materialise	
Materiality	

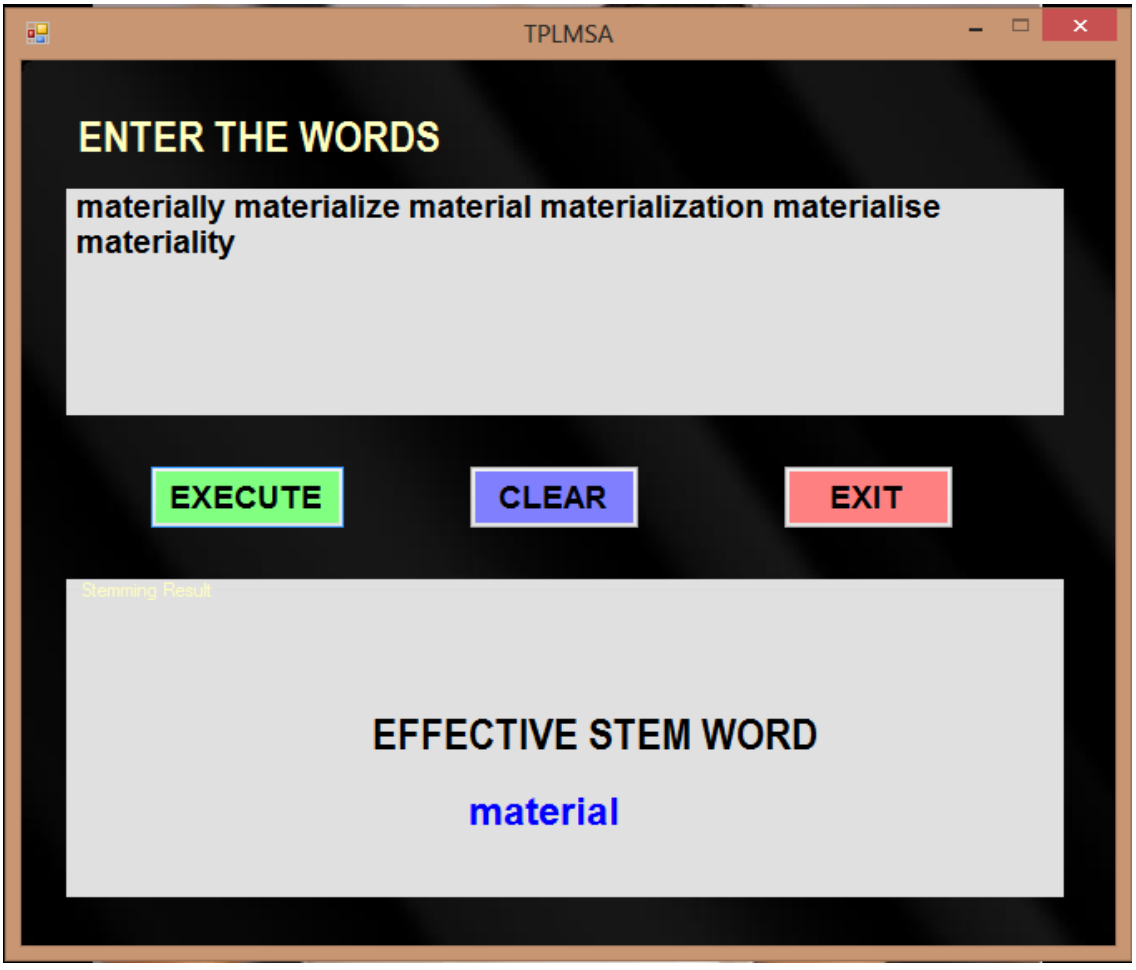


Figure 33 Execution Result of material output

Finally we take the some input from many research paper and website to check the execution result for the TPLMSA.

Table 4 Result of TPLMSA on Dataset

INPUT	OUTPUT OF TPLMSA
introduction introduces introducing introduced	introduc
connection connecting connectify connected	connect
converting converts convert	convert
selection selecting selected	select
producer production	Produc
consumed consumption	Consum
longest longer long	Long
bibliographically bibliographic bibliographics	Bibliographic
done doing does	Do
material materially materialize materialization materialise materiality	Material
drugs drugged drug	Drug
magnet magnetic magneto	Magnet
developing developed developing	Develop
processing processes process	Process
proposed proposing	Propos
programming programs programmer programmers	Program
document documenting documents	Document
functional functionality functioning	Function
transforming transformation transformed	Transform

Chapter 5

CONCLUSION AND FUTURE SCOPE

The data mining is used for analysis purpose in different sector. The aim of data mining is extract the knowledge from huge amount of data. Now a day the data is structured and unstructured format. The structured data analysis performed by the data mining but the unstructured data analysis performed by the text mining.

Text mining is the concept through which we extract of useful information from textual database or unstructured data. The text mining techniques are text clustering, text classification, text summarization, etc.

Before performing the text mining we need to pre-process the text data. The pre-processing of textual database is divided into three step; Tokenization, Stop Words removal, Stemming. Tokenization is used for the splitting the sentences into many words or tokens. The stop words is used for avoid the indexing unless words by removing the words like: as, a, of, the, so, is, are, etc. Stemming is used for the transformed the words into their root form. For example: similar, similarity similarly would be transformed onto similar word.

Our research work is focuses on the enhancing the stemming algorithm so that we can also improve the overall performance of text mining.

Text mining is the concept through which we can extract the useful information from textual database or unstructured data. Before performing the text mining we need to pre-process the text data. Our work based on the enhancing the stemming algorithm which leads to improve the overall performance of text mining. In this research we proposed new algorithm for stemming TPLMSA Two Phase Longest Match Stemming Algorithm. It is based on the longest match principal on the basic of left to right and right to left. The TPLMSA is used for the transform the most similar words into their root form. It produces the effective result for text mining.

In future we will try to implement the TPLMSA with the mining techniques which will leads to improve the overall performance of text mining.

REFERENCES

Anindya Ghose, Panagiotics G. Ipeirotis (2010), Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics, IEEE transactions on knowledge and data engineering 2010, 1041-4347/10.

C.Ramasubramanian1, R.Ramya (2013), Effective Pre-Processing Activities in Text Mining using Improved Porter's Stemming Algorithm, IJARCCE Vol. 2, Issue 12, December 2013.

Daniel Ramage, Christopher D. Manning, Susan Dumais (2011), Partially labeled topic models for interpretable text mining, KDD'11, August 2011, San Diego, California, ACM 978-1-4503-0813-7/11/08.

Jayaraj Jayabharathy and Selvadurai Kanmani (2014), Correlated concept based dynamic document clustering algorithms for newsgroups and scientific literature, Springer 2014 open access, http://www.decisionanalyticsjournal.com/1/1/3.

Jiawei Han and Micheline Kamber (2006) Data Mining: Concepts and Techniques, Elsevier Inc.

Julie Beth Lovins (1968), Development of Stemming Algorithm, Mechanical translation and computational linguistics, vol.11, nos.1 and 2, March and June 1968.

Kodimala, Savitha, "Study of stemming algorithms" (2010). UNLV Theses/Dissertations/Professional Papers/Capstones. Paper 754.

Liu HaiTao, Cong Jin (2013), Language clustering with word co-occurrence networks based on parallel texts, Chinese Science Bulletin, 2013, 58:1139-1144, doi: 10.1007/s11434-013-5711-8, published with open access at Springerlink.com.

Luying LIU, Jianchu KANG, Jing YU, Zhongliang WANG (2005), A Comparative study on unsupervised feature selection method for text clustering, IEEE 2005-0-7803-9361-9/05

Ms. Anjali Ganesh Jivani (2011), a comparative study of stemming algorithm, Int. J. Computer technology application, vol 2 (6), Nov-Dec 2011, ISSN: 2229-6093.

Peter Willet (2006), the Porter stemming algorithm: then and now, Electronic library and information system, 40 (3). pp. 219-223. ISSN 0033-0337.

Sun Kim, W John Wilbur (2012), Thematic clustering of text document using em-based approach, Journal of Biomedical Semantics 2012, 3(suppl 3):s6, doi: 10.1186/2041-1480-3-s3-s6.

Tao Liu, Shengping Liu, Zheng Chen, Wei- Ying Ma (2003), An evaluation on feature selection for text clustering, Proceedings of Twentieth International Conference on Machine Learning (ICMP-2003), Washington DC, 2003.

Tuomo Korenius, Jorma.Laurikkala, Kalervo.Jarvelin, Martti Juhola (2004), Stemming and Lemmatization in the Clustering of Finnish Text Documents, ACM 1-58113-000-0/00/0004, November 8–13 2004, Washington DC, USA.

Yanping Lu, Shengrui Wang, Shaozi Li, Changle Zhau (2011), Particle swarm optimizer for variable weighting in clustering high dimensional data, Machine Learning (2011) 82: 43-70, doi: 10.1007/s10994-009-5154-2, Springer open access.

Yongzhe shi, Wei-Qiang Zhang, Jia Jiu, Michael T Johnson (2013), RNN language model with word clustering and class based output layer, EURASIP Journal on audio, speech and music processing 2013, 2013:22, A Springer Open Journal.

Zhong, Ning, Li, Yuefeng, Wu, Sheng-Tang (2010), Effective pattern discovery for text mining, IEEE Transactions on Knowledge and Data Engineering.

WEBSITES

http://en.wikipedia.org/wiki/C_Sharp_(programming_language)

http://www.tutorialspoint.com/csharp/csharp_strings.htm

http://www.tutorialspoint.com/csharp/csharp_strings.htm

https://msdn.microsoft.com/en-us/library/dd492171.aspx

http://en.wikipedia.org/wiki/Microsoft_Visual_Studio

http://en.wikipedia.org/wiki/Stemming

http://en.wikipedia.org/wiki/Text_mining

http://en.wikipedia.org/wiki/Vector_space_model

APPENDIX

LIST OF ABBREVIATIONS

TPLMSA Two Phase Longest Match Stemming Algorithm

CRM Customer Relationship Management

ETL Extract, Transform, and Load

TMARDC Term frequency based maximum resemblance document clustering

CCMARDC Correlated concept based maximum resemblance document clustering

CCFICA Correlated concept based fast incremental clustering algorithm

VSM Vector Space Model

POS Part of speech

DF Document frequency

TC Term Contribution

TV Term Variance

ASP.NET Active Server Page .NET

HTML Hypertext Markup Language

CSS Cascading Style Sheet

UML Unified Modeling Language

GUI Graphical User Interface