Text Summarization

Sonali Agarwal
Department of Information Technology,
Vidyavardhini's College of Engineering and Technology,
India,

Pranita Redkar
Department of Information Technology,
Vidyavardhini's College of Engineering and Technology,
India

Aditi Gaur
Department of Information Technology,
Vidyavardhini's College of Engineering and Technology,
India

Prof. Swati Varma
Asst. prof.
Dept of Information Technology V.C.E.T. Vasai, India

Abstract- - **Due to availability of high volume of information and electronic documents on web, it becomes difficult for a human to study, research and analyze this data. The main idea and the major concept of the summarization enables humans to read the summary of a huge amount of text quickly and decide whether to further dig into details. Extractive summaries generates a brief by extracting proper set of sentences from a document or multiple documents by deep learning. For better efficiency, the summarization procedure is manipulated by Restricted Boltzmann Machine (RBM) algorithm by removing redundant sentences.RBM consist of three layers input, hidden and output layer. The input is uniformly distributed in the hidden layer for operation and after operations summary is obtained as output.**

Index Terms – Summarization, Restricted Boltzmann Machine (RBM),Deep Neural Networks(DNN).

I. Introduction

From about many years, synopsis are generated by humans manually..In the intervening years , the heap of text data is increasing gradually by the means of internet and other sources. To overcome the problem of information overloading , summarization is essential .Abstract generated by human is time consuming and tedious. Thus there is a need for automatic summarization to save time and to get quick results. Auto Summarization is defined as the art of condensing large text documents into few lines of summary. Summarization can be classified as follows:

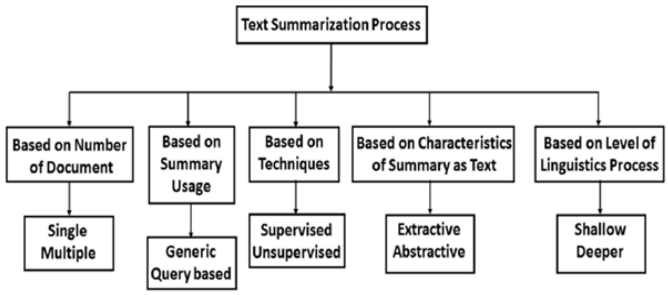


Figure 1:- Types of text summarizarion

Text summarization can be classified on the basis of different criteria. Synopsis based on construction, extractive summaries pick important sentences from the document based on certain conditions and display it to the user as they are, abstractive summaries gives reconstructed summary which is not exactly the same as the original document. On the number of sources for the summary, single document summary which is produced from single document, multi-document summary which is obtained from multiple documents. Trigger based abstracts can be of two types, generic based summary, present the summary in concise manner as the main topic of the data and query based summarization gives synopsis as an answer to the query given by the user. Also based on the important details of the synopsis it can be classified as indicative which gives information to the user whether the document should be read or not, and the informative abstract provide all the relevant information to represent the original document. The most challenging problem of auto summarization is to provide information that is relevant to user's topic of interest. To accomplish this approach is proposed using Deep neural network (DNN) along with a provision to give seed word to the summarizer so that it summarizer the documents which include the seed word, so that the user gets all the relevant information related to the search thus saving a lot of time. DNN provides the semantic space for the sentence so that the sentences with meaningful semantics can be extracted thereby reducing the data redundancy.

II. MOTIVATION

Now days more and more information is available through internet and other sources. To handle these data more efficiently we need a tool for extracting proper set of sentences from the given documents. Summarization of text is essential to get the important information while dealing with large collection of documents. With the advent of World Wide Web information has become intrinsic part of our life. To remember the details of every information is not possible for human mind. Therefore summarization of text documents plays a very important role in information gathering. In this study we are using deep learning Algorithm for the summarization task. Deep learning is the emerging field of machine learning, which is used to solve problems of number of computer science domain like image processing, robotics, motion. Recently it is also used in domain of Natural language processing with very encouraging results. An algorithm is deep if its input is passed through several of nonlinearity's before being output most modern learning algorithms including svm and naive ayes classifier are shallow. Here we are using the Restricted Boltzman Machine to extract the top most feature word of text.

III.LITERATURE SURVEY

3.1 Text Summarization Based on Fuzzy Logics

A Design of a Fuzzy logic system usually involves selecting membership function and fuzzy rules. The performance of the fuzzy logic system will directly affect by the selection of fuzzy rules and membership functions[1]. The four main components of the Fuzzy Logic were: fuzzifier, inference engine, defuzzifier, and the fuzzy knowledge base. In the fuzzifier section, snappy inputs are translated into linguistic values, using a membership function, which is to be used to the input linguistic variables. After fuzzification, to derive the linguistic values, the inference engine refers to the rule base containing fuzzy IF THEN rules[2]. Finally, the defuzzifier converts the output linguistic variables from the inference to the final crisp values using membership function which represents the final sentence score. The output membership function in the defuzzfication step is divided into three membership functions: Unimportant, Average, Important, Which is used to convert the inference engine result into a crisp output to obtain a final score for each sentence[5]. Here fuzzy centroid method is used to calculate the score for each sentence in a document, which is obtained by using generalized triangular membership function which depends on the three parameters a, b, and c. where the parameter a and c are left and right most feet of a triangle and b is the peak of a triangle[4]. Based upon the sentence features and knowledge base the output is obtained as a value from zero to one for each sentence. Such obtained value shows the degree of importance of the sentences in the final summary.

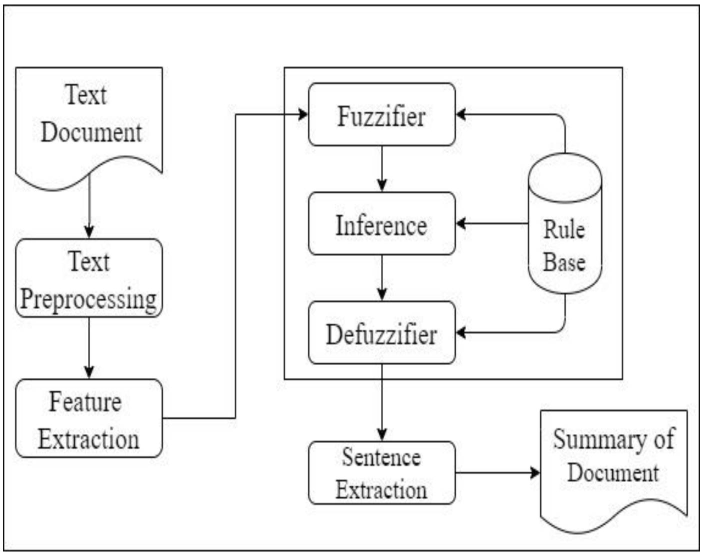


Figure 2:- Fuzzy Logic System Architecture.

3.2 .Text Summarization Based on Neural Networks

Restricted Boltzmann Machine RBM is a stochastic neural network (that is a network of neurons where each neuron has some random behavior when activated). It consists of one layer of visible units (neurons) and one layer of hidden units[3] . Units in each layer have no connections between them and are connected to all other units in other layer as shown below:

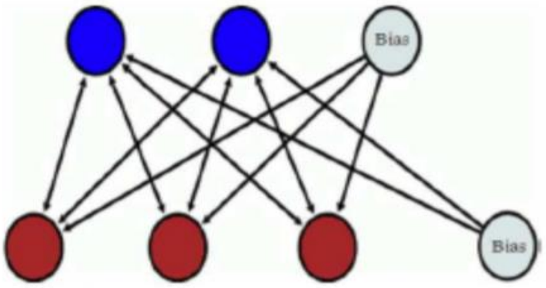


Figure 3:- Layers of restricted boltz-mann machine.

Connections between neurons are bidirectional and symmetric. This means that information flows in both directions during the training and during the usage of the network and those weights are the same in both directions. First the network is trained by using some data set and setting the neurons on visible layer to match data points in this data set[7]. After the network is trained we can use it on new unknown data to make classification of the data which is known as unsupervised learning. During text summarization the text document is preprocessed using various prevalent preprocessing techniques and then it is converted into feature matrix defined over a vocabulary of words. This feature matrix each row will work as a input to our RBM[3]. Based on the structured matrix, the proposed text summarization algorithm uses the fuzzy classifier to assign class labels for the sentences, in order to compute the relevance of each sentence based on the rule selector. The rules are then divided into corresponding sentences and the sentences are then used to form the new feature matrix[7]. After getting the set of top priority word from the RBM the input query, sentence vector and high priority word output is compared to generate the extractive summary of the text document.

.

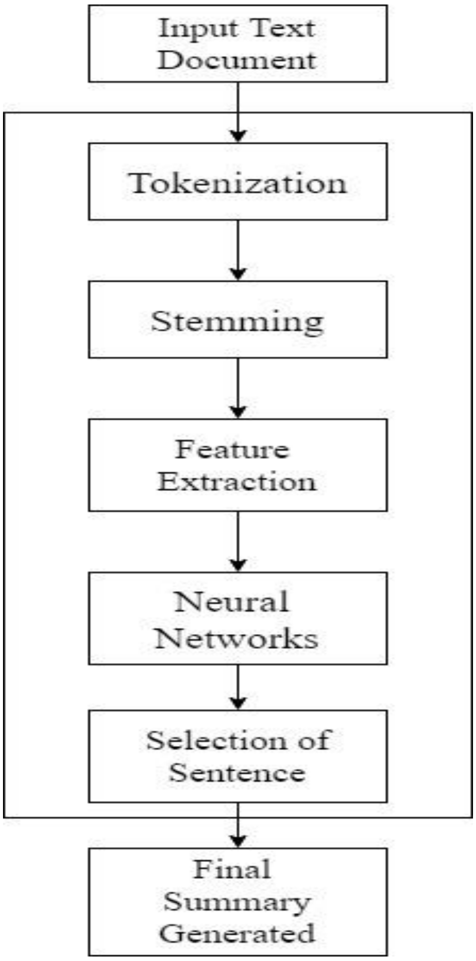


Figure 4:-Flow chart for text summarization using rbm.

V.TEXT SUMMARIZER

An automatic summarization process of creating summaries for a topic relevant to users search and can be divided into three steps:

- Preprocessing step :-It is a structured representation of the original text is obtained.
- Processing step:-An algorithm must transform the text structure into a summary structure.
- Generation step :-the final summary is obtained from the summary structure.

The basic principle is to process the text document given as an input buy the user. This processing will remove all the unwanted and irrelevant words for the document .After which in the processing step relevant words are given[8]. Finally the summary will generated.

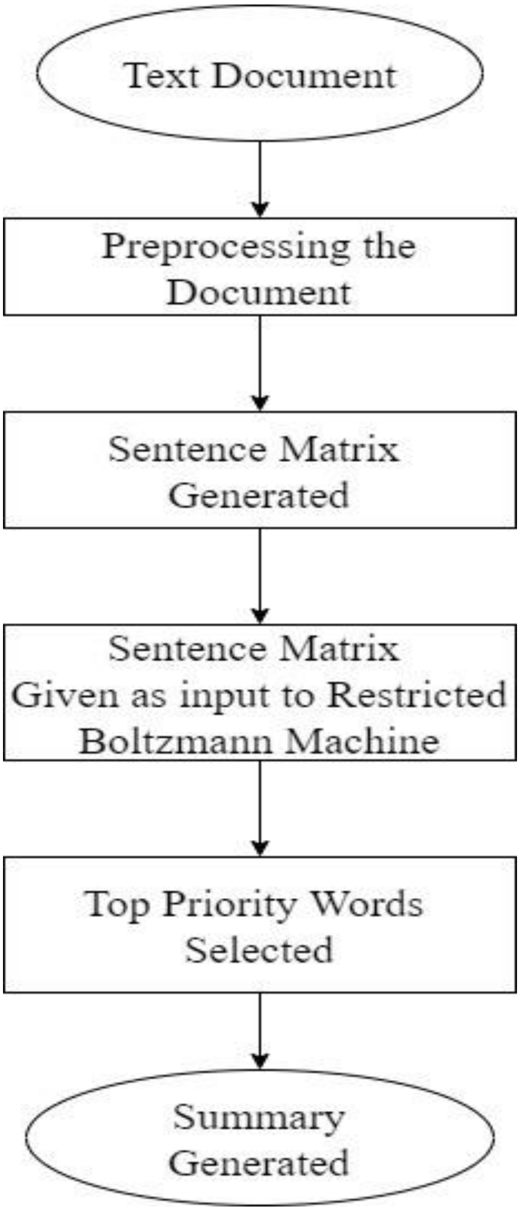

Text Document
Preprocessing the Document
Sentence Matrix Generated
Sentence Matrix Given as input to Restricted Boltzmann Machine
Top Priority Words Selected
Summary Generated

Figure 5:- Text Summarizer implemented.

- First the document is taken as input.
- Then preprocessing is performed for which nltk is used and all the unwanted and irrelevant words from the documents are filtered.
- In second step an algorithm is used which will convert text in summary structure and will give us top priority sentences.

- Then this sentences are rearranged as they appear in the document and finally summary is generated.

5.1 Preprocessing

To make the document light preprocessing of the text document for structuring is done by applying various techniques developed by the linguist[1]. There are myriads of technique by which we can reduce the density of text document. In this study we are using the following techniques.

5.1.1 Stemming

Stemming is process of bringing the word to its base or root form for example using words singular form instead of using the plural (using boys as boy), removing the ing from verb (changing doing to do)[8]. There are number of algorithms, generally referred as stemmers', are there that can be used to perform the stemming.

5.1.2 Part Of Speech Tagging

Part of speech tagging is the process of marking or classifying the words of text on the basis of part of speech category (noun, verbs, adverb, adjectives) they belong[4].

5.1.3 Stop Word Filtering

Stop words are the words which are filtered out prior to or after the preprocessing task generally there is no specific rule on a particular word to be stop word, it is completely subjective depends upon the situation[2]. In our condition we considering words like a, an, in by as stop word and filters this word from the original document

5.2 Feature extraction

After reducing the density of document, the document is structured into a matrix. A sentence matrix S of order n*v is containing the features for every sentence of a matrix[8]. For very informative summarization we are extracting four features of a sentence of text document viz similarity with title, relative position of sentence, term weight of words forming sentences, concept-extraction of sentence

5.2.1 Title Similarity

A sentence is considered important if it's similar to the title of text document. Here similarity is considered on the basis of occurrence of common words in title and sentence[7]. A sentence has good feature score if it has maximum number of words common to the title.

5.2.2 Positional Feature

Positional value of a sentence is also extracted. A sentence is relevant or not can also be judged by its position in the text[6]. To calculate the positional score of sentence we are considering the following conditions:

- Feature score = 1, if sentence is the starting sentence of the text

- Feature score = 0, if sentence comes in the middle paragraphs of text

- Feature score = 1, if sentence comes in the last of the text

5.2.3 Term Weight

This is another very important feature to be consider for summarization of text. Here by term weight we simply mean the term frequency and its importance. This is the most standard feature considered in various natural language processing tasks[3]. The frequency here is the term frequency which reflects the importance of a word in a document, it simply tells number of times a word appears in the text.

5.2.4 Concept Feature

The concept feature from the text document is extracted using the mutual information and windowing process. In windowing process a virtual window of size 'k' is moved over document from left to right[5]. Here we want to find out the co-occurrence of words in same window and it can be calculated by following formula:

5.3 Sentence Matrix

Here sentence matrix S = (s1, s2,……..sn)
where si = (f1, f2,……..f4), i<= n is the feature vector[8].

5.4 Deep Learning Algorithm

The sentence matrix S = (s1, s2,……..sn) which is the feature vector set having element as si which is set contains the all the four features extracted for the sentence si. Here this set of feature vectors S will be given as input to deep architecture of RBM as visible layer. Some random values is selected as bias Hi where i = 1,2 since a RBM can have at least two hidden layer[8]. The whole process can be given by following equation:

S=(s1,s2,……..,sn)

where,
si = (f1,f2,……..f4),
i<= n where n is the number of sentences in the document. Restricted Boltzmann machine contains two hidden layers and for them two set of bias value is selected namely H0H1:

H0={h0,h1,h2,…….,hn}
H1={h0,h1,h2,…….,hn}

These set of bias values are values which are randomly selected. The whole operation of Sentence matrix is performed with these two set of randomly selected value[8]. The whole operation with RBM starts with giving the sentence matrix as input. Here s1,s2,……..sn are given as input to RBM. The RBM generally have two hidden layers as we mentioned above. Two layers are sufficient for our kind of problem. To get the more refined set of sentence features. RBM works in selected value[1]. The whole operation with RBM starts with giving the sentence matrix as input. Here s1,s2,……..sn are given as input to RBM. The

Paper Title

RBM generally have two hidden layers as we mentioned above. Two layers are sufficient for our kind of problem. To get the more refined set of sentence features. RBM works in two step. The input to first step is our set of sentence matrix, S = (s1,s2,……..sn), which is having the four features of sentence as element of each sentence set[1].
for each feature we have calculated. For example we select threshold thrc as a threshold value for the extracted concept- feature. If for any sentence f4<thr then it will be filtered and will become member of new set of feature vector.

Summary Generation

In summary generation phase, the obtained optimal feature vector set is used to generate the extractive summary of the document. For summary generation first task is obtaining the sentence score for each sentence of document. Sentence score is obtained by finding the intersection of user query with the sentence[1]. After this step ranking of the sentence is performed and the final set of sentences for text summary generation defining the summary is obtained.

V. RESULT AND DISCUSSION

The objective of the system is to provide a summary of any given text document. The proposed approach deals with text summarization based on a deep learning method. The method that we proposed incorporates the RBM algorithm for getting better efficiency. The Figures below shows the website and the output ,wherein a sentence matrix is generated and the words will be selected and finally summary will be generated. The summary is generated as such the meaning of the documents does not change and the it makes it easy to determine whether the document is releated to their requirement or not which will help them to save time.

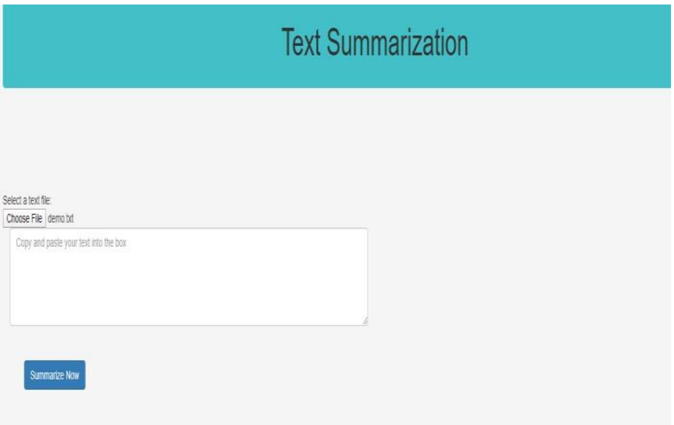
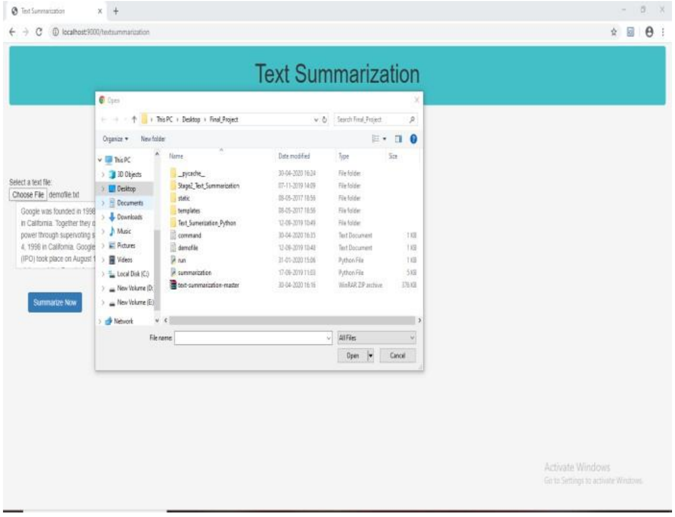


Fig 6 Text Summarization Website.

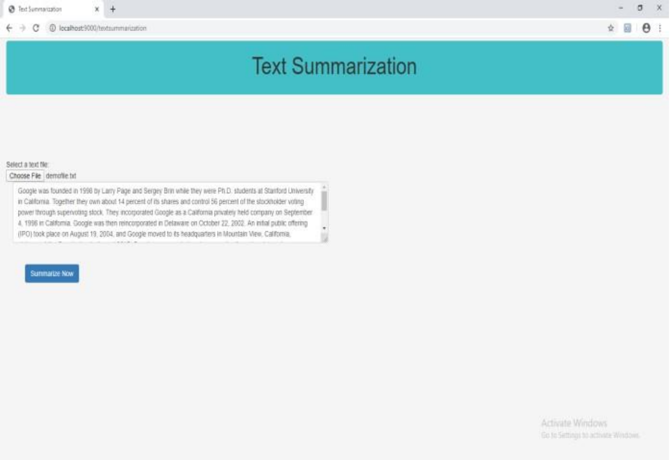
We have compared various methods of summarization based on various features such as title similarity, occurrence of words through out the document ,nouns and date. The nltk libraries are used for summarization. Then the sentence matrix is generated based on the feature score and then based on the score top sentences are selected and arranged as per their occurrence in the documents.



Name	Date modified	Type	Size
__pycache__	30-04-2020 16:24	File folder	
Stage2_Text_Summarization	07-11-2019 14:09	File folder	
static	08-05-2017 18:56	File folder	
templates	08-05-2017 18:56	File folder	
Text_Sumerization_Python	12-09-2019 10:49	File folder	
command	30-04-2020 16:35	Text Document	1 KB
demofile	12-09-2019 10:48	Text Document	1 KB
run	31-01-2020 15:06	Python File	1 KB
summarization	17-09-2019 11:03	Python File	5 KB
text-summarization-master	30-04-2020 16:16	WinRAR ZIP archive	378 KB

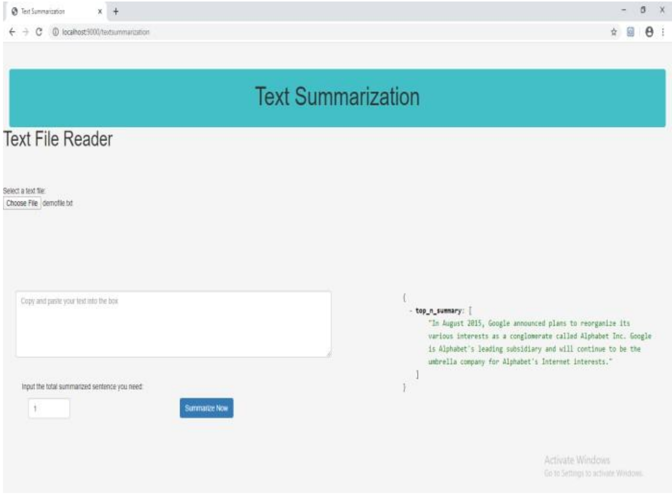
Google was founded in 1998
in California. Together they o
power through supervoting s
4, 1998 in California. Google
(IPO) took place on August 1

Fig 7:-Choose file to be summarized.



Google was founded in 1998 by Larry Page and Sergey Brin while they were Ph.D. students at Stanford University in California. Together they own about 14 percent of its shares and control 56 percent of the stockholder voting power through supervoting stock. They incorporated Google as a California privately held company on September 4, 1998 in California. Google was then reincorporated in Delaware on October 22, 2002. An initial public offering (IPO) took place on August 19, 2004, and Google moved to its headquarters in Mountain View, California,

Fig 8:-Contents of file is displayed in the text area.


Text Summarization
localhost:9000/textsummarization
Text Summarization
Text File Reader
Select a text file:
Choose File demofile.txt
Copy and paste your text into the box
Input the total summarized sentence you need:
1
Summarize Now
Activate Windows
Go to Settings to activate Windows.

```
{
  - top_n_summary: [
        "In August 2015, Google announced plans to reorganize its
        various interests as a conglomerate called Alphabet Inc. Google
        is Alphabet's leading subsidiary and will continue to be the
        umbrella company for Alphabet's Internet interests."
    ]
}
```

Fig 9:-Final Summary Generated

VI .CONCLUSION

Several researches were conducted for summery generation from the multiple documents in recent days. We have developed automatic multi-document summarization system which incorporates the RBM. We have used four different features for feature extraction phase. The feature score of the sentences is applied to the RMB in which the RBM rules are optimized with the help of Deep Learning Algorithm. The features are processed through different levels of the RBM algorithm and the text summary is generated accordingly. The generated result is tested as per the evaluation matrices. The evolution matrices considered in the proposed text summarization algorithm are recall, precision and fmeasure. The experimentation of the proposed text summarization algorithm is carried out by considering three different document sets. The responses of three documents sets to the proposed text summarization algorithm are satisfactory. The futuristic enhancement to the proposed approach can done by considering different features and by adding more hidden layers to the RBM algorithm.

VII .REFERENCES

[1] G.Padmapriya,Dr.K.Duraiswamy,"An Approach for text summarization using Deep Learning Algorithm",JCSSP,2014.

[2] Heena A. Chopade , Dr.Meena Narvekar, "Hybrid Auto Summarization ,"Using Deep Neural Network And Fuzzy Logic System", IEEE, 2017.

[3] Ashwini Anbekar, Kajol Shah , Minakshi Agarwal, Simica Pawar , Asma Shaikh,"Text Summariazation Using Restricted Boltz-mann Machine:Unsupervised Deep Learning Approach",IEEE,2018.

[4] Abdullah Goktug Mert,"Text Summarizer with Deep Learning",METU,2016.

[5] Trun Kumar,"Automatic text summarization",NITR,2014.

[6]S.Santhana Megala, Dr. A.Kavitha, Dr. A.Marimuthu,"Enriching Text Summarization using Fuzzy Logic",IJCSIT,2014.

[7] G.Padmapriya,Dr.K.Duraiswamy,"Association Of Deep Learning Algorithim With Fuzzy Logic For Multi Document Text Summariazation",JATIT,2014.

[8] Mehdi Jafari, Amir Shahab Shahabi, Jing Wang, Yongruri Qin, Xiaohui Tao, Mehdi Gheisari, " Automatic Text Summarization using Fuzzy Inference", IEEE, 2016.