JCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-6, June 2018 E-ISSN: 2347-2693

# Automatic Extractive Text Summarization Using K-Means Clustering

M R Prathima[1*], H R Divakar[2]

[1] Dept. of Master of Computer Applications, PES College of Engineering, Mandya, Karnataka, India
[2] Dept. of Master of Computer Applications, PES College of Engineering, Mandya, Karnataka, India

*Corresponding Author: mrcprathima@gmail.com, Mob.: 7204448143*

**Available online at: www.ijcseonline.org**

Accepted: 22/Jun/2018, Published: 30/Jun/2018

***Abstract***- In recent year, data is emerging rapidly in each and every domain such as social media, news, education, etc. Due to data excessiveness, there is a need for an automatic text summarizer which will be having an ability to summarize the data. Since the research importance focusing on Natural Language Processing (NLP), text summarization can be used in several fields. Text summarization is a process of extracting data from a documents and generating summarized text of that documents. Thus presents an important data to the users in a relatively more concise form. The study of various extractive summarization of text is made and an essential text summarization method is proposed on the basis of Support-Vector-Machine (SVM). The proposed model tries to improve the quality as well as performances of the summary generated by the clustering technique by cascading it with Support-Vector-Machine (SVM). The documents are preprocessed to get the tokens that are obtained after tokenization, stop word removal, case folding and stemming. The various similarity measures are utilized in order to identify the similarity between the sentences of the document and then they are grouped in cluster on the basis of their term frequency and inverse document frequency (tf-idf) values of the words.

***Keywords***-Text Summarization, Extractive Summarization, Natural language Processing (NLP), Clustering, Support-Vector-Machine (SVM), Advanced Encryption Standard (AES), Tokens.

## I. INTRODUCTION

Usually, the existence of the exactly similar information in multiple documents is the main issue facing in an effective information access. Instead of this information redundancy thus accessed or retrieved, users are more concerned in retrieving the information which addresses the one or the other various aspects. In such situations, the text summarization proves to be beneficial not only in information retrieval, but it is found to be highly active research topic in other fields such as Natural Language Processing (NLP) and Machine Learning. Currently the text summarization is gaining much significance and one of the reason for this is, recently due to the extensive development in material, requirement for involuntary text summarization has extended. There is a large quantity of text material available on the internet. However, usually the internet provides more material than is essential. Hence, a twofold problem is encountered i.e., examining for the pertinent documents through a crush amount of existing documents, and absorbing a huge quantity of pertinent information.

Text Summarization has become very important since from many years. To generate a summarized document, a reader and an identifier is required to generate summary. Summary is a data assembled by collecting an equivalent information files and extracting only the important points to be inserted in summary. When the user searches for an information by entering query, the internet will be providing a huge number of files that matches the score of related data in the query posted, and thus user time is wasted in searching for the pertinent data. But it is not possible for the user to conclude on the required file. This issue grows rapidly as information flow into web increases in World Wide Web [1]. Text summarization is also found to be a technique where summary or abstract is automatically generated using computer techniques from one or more text [2]. According to Babar [2, 3], a summary is a text and that can be assembled from one or more texts, that is capable of delivering a significant information in the actual text and it will be of condensed form. The main aim of automatic text summarization is to introduce the source text into a condensed version with the semantics. The most significant advantage of utilizing a summary is due to that reduces the reading time technologies which is able to make a coherent summary to take into the account variables like writing style, length and syntax Abderrafih [2, 4]. Basically there are two methods of summarization: abstraction and extraction. Extractive summarization of text construct summaries by selecting a subset of the sentences from actual documents. The variance identified in abstractive summarization, where the information in the text is expressed in a different way. The extractive summarization provides a better outcome compared to the automatic abstractive summaries. This is due to the reality that the issues seen in an abstractive summarization such as representation of the semantics, inferences and natural language generation, are relatively firm when compared to a data-driven approach such as sentence extraction [5]. Since text summarization is found to be complicated task which involving the Natural Language Processing (NLP), it uses NLP tools such as, Dictionaries, Thesaurus, Tokens, and POS Tagger etc., [6]. The main aim of the summary is to present the key concepts in a document in limited space. Summarization of text usually based on method for the sentence extraction and identifies the set of sentences which are having the greatest meaning for the general understanding of the given documents.

In this paper, Section I contains the introduction of text summarization, Section II contains the literature survey of text summarization, Section III explains the methodology and contains the architecture and essential steps of proposed work, Section IV explains the system implementation with flow chart, Section V describes the results and discussion of proposed work, Section VI concludes research work with future directions.

## II. LITERATURE SURVEY

Text mining is the process of extracting information from the text that have greatest meaning and in recent times which has gained the important attentions. Thus describing the most basic text mining tasks as well as techniques including the pre-processing of the text, classification and clustering [7]. Conditional Random Fields (CRF) model is a state-of-art that has a sequence labelling method, and that can use the characteristics of documents more effectively. At the same time, extraction of keywords can be taken as string labelling. Thus, keywords extraction on the basis of Conditional Random Fields (CRF) is proposed and implemented and the outcome of the experiment shows that the CRF model performs the other machine learning methods such as Multiple linear regression model, Support-Vector-Machine, etc. in the keyword extraction tasks [8]. Manual extraction of relevant keywords are the procedures that are overflowing with errors with loads of time and manual effort. Therefore, a hybrid approach is proposed for an automatic keyword extraction for multi-document text summarization in newspaper articles. Techniques such as term frequency-inverse document frequency (tf-idf), term frequency-adaptive inverse document frequency (tf-aidf), and a number of false alarm (NFA) were proposed for an automatic keyword extraction in e-newspaper articles for preferable analysis [9]. An approach is presented in order to cluster the multiple documents by using the document clustering approach and to produce a summary in a cluster wise on the basis of feature profile oriented sentence extraction strategy. The sentence score is calculated for each sentences on the basis of feature profile. Thus the experimental results shows that the proposed clustering algorithm is more efficient and feature profile can be used to extract the significant sentences from multiple documents [10]. Summarization methods that are based on traditional cluster usually faces the problems with the compressions, the computational speed, selection of peer, and clustering of the sentences to generate the summaries with high quality. The traditional clustering of the documents and summarization methods presumes the adjacency node and information of the neighbourhood in order to build the clusters and summaries. Since p2p networks have faced problems with the adjacency node and identical information, it was complicated to generate an optimal clusters and a summaries in the peers. Thus, the approach is proposed in order to provide better solution to generate an optimal clustering of the document using probabilistic k-representative clustering algorithm and to make summaries more efficient using summarization based on phrase rank. Results obtained by the experiments give the better performance with regarded to the entropy, execution time, and qualities of the clusters are concerned [11].

## III. METHODOLOGY

Several methods can be applied in order to overcome the disadvantages that are exists as said in the introduction. Protecting the information’s against environments is a critical factor in the development of information-based processes in business, industries, and administration as information is the most significant advantage of an organization. Therefore, we use an Advanced Encryption Standard (AES) algorithm. This Advanced Encryption Standard (AES) algorithm is a symmetric block cipher which is used to encrypt and decrypt the information’s [12]. In order to secure the electronic data, an Advanced Encryption Standard (AES) algorithm can be used. Also the term frequency-inverse document frequency (tf-idf) is used which is incredibly powerful and which is used to judge the topic of an article by the words that it contains. Early research worked on the extractive summarization is on simple informative features of the sentences like their position in the text, frequency of words altogether they contain, or key phrases that specifying the significance of the sentences [5, 13, 14, 15]. The statistical method known as the tf-idf is used to make the data more accurate, reflecting the context of the text sample better. One of the usual way of determining the value of the tf is to basically just take the raw frequency of any term in the document and one of the usual way of determining the idf is to get the log of the inverse portion of the documents that having the term. And by multiplying both values, the magic value is obtained. The tf-idf reduces the value of common words that are used across different documents.

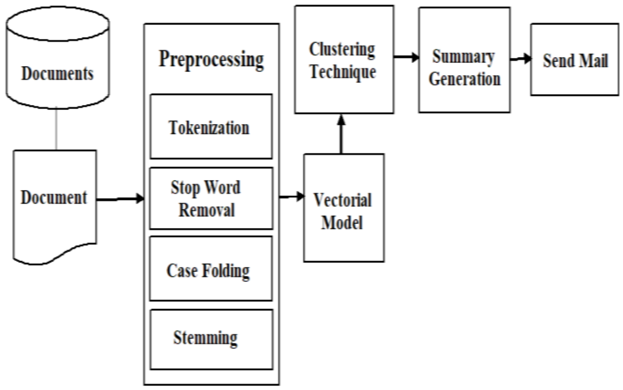

Figure 1. System Architecture

Figure 1 shows the process of the automatic extractive text summarization using k-means clustering in the following steps

### Pre-processing Step

Pre-processing Step is a work before the translation. The document is the input to the summarizer. Document should be transfer into a bag of word or phrases of the document. Here pre-processing step consists of Natural Language Processing (NLP) phases like tokenization, stop word removal, case folding and stemming. Once the pre-processing is done, the term frequency and inverse documents frequency values are calculated for each token.

### Tokenization

Tokenization is the act of dividing sentences into streams of individual tokens that are differentiated by the spaces and that can be used for further processing and understanding. Tokens can be individual words, keywords, phrases, identifiers, etc. In the process of tokenization, tokens or words are segregated by the whitespace, the line breaks or the punctuation marks. The whitespace or the punctuation marks may or may not be involved depending on the requirements.

### Stop Word Removal

Stop Words are the words that occur repeatedly. Stop Word Removal is the process of removing words like “the”, “to”, etc. and stop words are removed in order to support phrase search.

### Case Folding

Case Folding is the process of converting all the letters that the documents contained to the lower case.

### Stemming

Stemming is the process of reducing the derivationally related forms or inflectionally related forms of words to their stem form, common base form or root form – generally a written word form that may help to increase the coverage of Natural Language Processing (NLP) utilities.

### Clustering Technique

Clustering is a process that involves the classification of data points. When the set of data points are given then, the clustering algorithm can be used to classify each data point into a particular class. An algorithm will be generated that containing clustering machine learning technique i.e., Support-Vector-Machine (SVM).

### Summary Generation

Summary of the text document will be generated using two technique, namely:

- The clustering technique and
- The clustering technique cascade with Support-Vector-Machine (SVM).

Summarization of the clustered documents is done based on the ranking and scoring in order to get the brief summaries.

### *A. Advanced Encryption Standard*

This section describes about the algorithm that are used in the automatic extractive text summarization using k-means clustering. The algorithm used in these task is Advanced Encryption Standard (AES) and it is designed as shown in the figure 2 below.

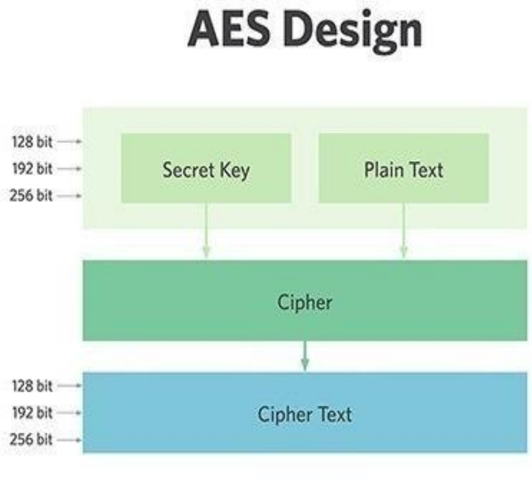

Figure 2. Advanced Encryption Standard Design

AES is a symmetric block cipher which is used to protect the information that are classified and is it implemented in software and hardware throughout the entire globe to encrypt the sensitive data. AES performs its computations on bytes comparatively than bits. Therefore, the 128 bits of a plaintext block is treated as 16 bytes by AES. And these 16 bytes are arranged in an order of four columns and four rows for processing as a matrix. Unlike DES, the number of rounds that are AES is variable and also depends on the key length. AES consider 10 rounds for 128-bit keys, 12 rounds for 192-bit keys and 14 rounds for 256-bit keys. And these each rounds uses a different 128-bit round key and that is calculated from the actual AES key.

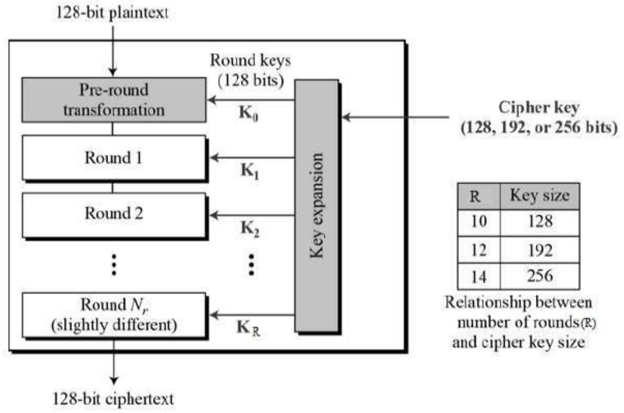

Figure 3. Schematic of AES structure

**Encryption Process:** Here, the description of a typical round of AES encryption is restricted. And each of the round is comprised of four sub-processes. The first round process is represented below

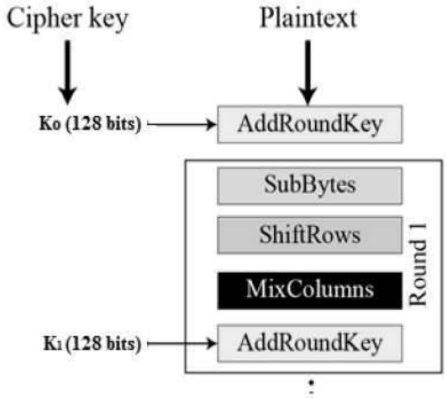

Figure 4. Representation of first round

**Byte Substitution (Sub Bytes):** The 16 input bytes are replaced by taking a look on the fixed table (S-box) given in the design. The result will be in a matrix of four rows and four columns.

**Shift-rows:** Each of these four rows of the matrix are shifted to the left. Any entries which 'fall off' that are re-inserted on the right side of the row. And the shift is carried as follows

- The first row is not shifted.
- The second row is shifted one position to the left side.
- The third row is shifted two positions to the left side.
- The fourth row is shifted three positions to the left.
- Finally the result obtained in a new matrix consisting of the same 16 bytes but shifted with respect to each other.

**Mix-Columns:** Each of these columns of four bytes is now changed using a special mathematical function. Now this function is taken as input the four bytes of one column and outputs four new bytes and that replaces the actual column. The result obtained is another new matrix that consisting of 16 new bytes. And at the last round this step is not performed.

**Add-Roundkey:** Now the 16 bytes of the matrix are considered as 128 bits and are XO-Red to the 128 bits of the round key. The output is the cipher text only when this is the last round. Otherwise, the resulting 128 bits are interpreted as 16 bytes and start another similar round.

*B. The term frequency-inverse document frequency (tf-idf)*

In text summarization, each sentence is given a significance score and this is one of the goodness measure for each sentence. The probability of a sentence which is to be present in a summary is proportional to its score. Each and every sentence is represented by a set of features and the score will be a function of the weighted sum of the separate feature values [5, 16]. The morality of a sentence is usually represented by the significance of the words that are present in it. The term frequency-inverse document frequency (tf-idf) is a simple but it is a powerful informative for ranking the words according to their significance. This feature will be the sum of the tf-idf scores of each individual words of sentence.

The term frequency (tf), which measures how frequently a term that occurs in a document. Since every document will be of different length, there is a possibility of that term would appear several times in a larger documents compared to smaller ones. Therefore, the term frequency is frequently divided by the length of the document as a normalization way

tf (t) = (Number of times term t that occur in the document) / (Total number of the terms present in the document)

The inverse document frequency (idf), which measures how significant a term is. While operating the term frequency, all terms are treated equally significant. However, certain terms such as "as", "that", and "of", may occur several times but with little significance. Therefore, we should be weigh down the frequent terms while presenting the rare ones, by computing as follows

idf (t) = $\mathrm{loge}^{(\text{Total number of the documents / Number of the documents with the term t present in it})}$ [5, 17, 18, 19, 20].

## IV. SYSTEM IMPLEMENTATION

Implementation is one of the most crucial stage in achieving a successful system and giving the user's confidence that the new system is workable and more effective and this modified application overcomes an existing one.

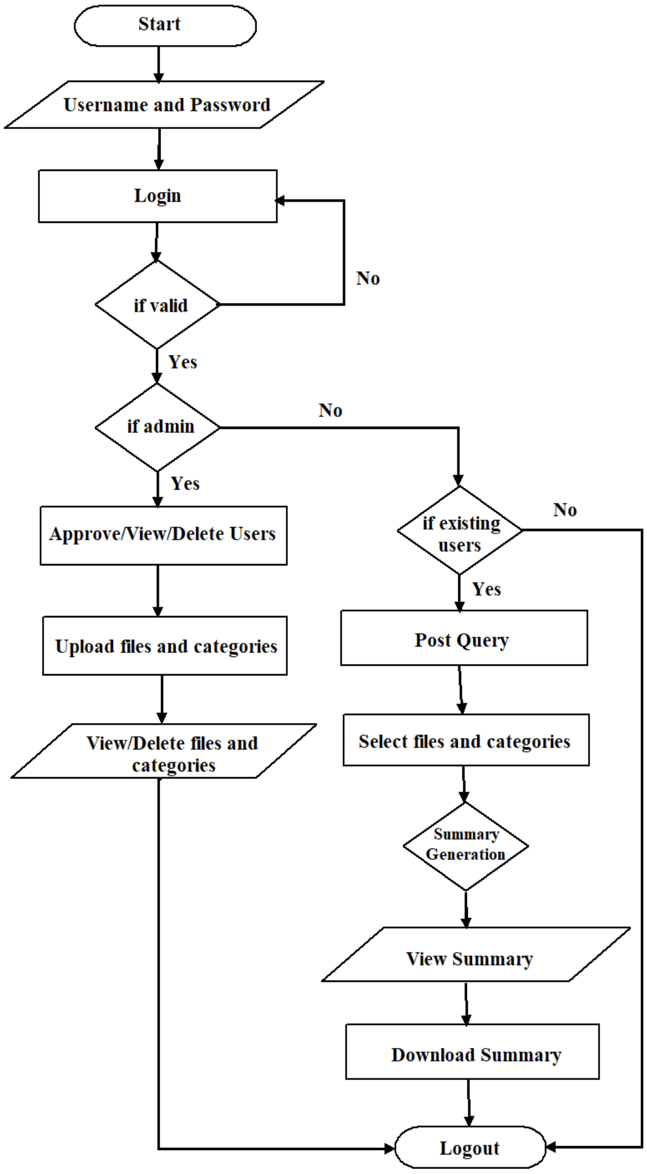

Figure 5. Text Summarization Flowchart

Application has two modules. i.e., admin and user module.

The proposed scheme is as follows

### Admin Module

Admin is the main part of the application he/she has all the authority to access as ADD, EDIT, REMOVE and UPDATE. Admin can view number of users registered and even manages documents. i.e., upload documents, view documents, remove unwanted documents and update the documents and also manages categories. i.e., add different types of categories.

### User Module

User can register into the application. User is given access using a username and password. User will be providing the query by selecting required file and category. The application takes the user query and processes the documents and extracts the data with the help of data mining. Based on the query posted by the user, the summarized text is generated. User can view and download the generated summary.

## V. RESULTS AND DISCUSSION

Result and discussion is to determine the concept and estimate it. The application is written in visual studio 2015 with additional frameworks. Microsoft SQL is used for the storage of data in the databases. CSS are used for the front end design. To estimate, Advanced Encryption Standard (AES) algorithm is used where AES consider 10 rounds for 128-bit keys, 12 rounds for 192-bit keys and 14 rounds for 256-bit keys in order to complete the overall task. The concept of encryption functionality is also done on the basis of AES algorithm itself because AES is more secure and also more efficient both in software as well as hardware. As a result

- Extraction of the text content is done in a secure and in an efficient manner.
- Formation of the text clusters that are relevant to the query is done properly.
- No text redundancy will happen.
- Since there is an increase in the amount of content that are available online, fast and an effective automatic summarization has become more significant. Maximum information can be obtained by spending the less time.

## VI. CONCLUSION AND FUTURE SCOPE

The development of the proposed system is futuristic and secure. This application has helped many users to get the generalized summary based on the query given. It has been thoroughly tested and implemented. After completing this application I am sure that the problems in the existing system would overcome. Here, in this application, the data is captured by reading all the documents and the keywords are identified and based on the keywords segmentation, the data is retrieved and summarized and it is displayed for the users. The future enhancement of this application is to create a procedure to generate an automated way of summary in a better and faster way and process the various types of documents by using multiple data mining techniques and creation of mobile application for the text summarization approach in smart phones.

## REFERENCES

[1] Shiva Kumar K M and Soumya R, "Text Summarization using Clustering Technique and SVM Technique", International Journal of Applied Engineering Research, Vol. 10, No. 12, 2015.

[2] Mgbeafulike IJ nad Christopher, "CONDENZA: A System for Extracting Abstract from a Given Source Document", Journal of Information Technology and Software Engineering, Vol. 8, Issue 1, 2018.

[3] Babar S, "Text Summarization", an overview, 2013.

[4] Lehmam A, "Essential summarizer: Innovative automatic text summarization software in twenty languages", ACM Digital Library, Personalization and Fusion of Heterogeneous Information, 2010.

[5] Ayush Agarwal and Utsav Gupta, "Extraction based approach for text summarization using k-means clustering", International Journal of Scientific and Research Publications, Vol. 4, Issue 11, Nov 2014.

[6] Simran Kaur and wg.cdr Anil Chopra, "CLUSTERING BASED DOCUMENT SUMMARIZATION", International Journal of Emerging Trends and Technology in Computer Science, Volume 5, Issue 1, January-February 2016.

[7] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D.Trippe, Juan B. Gutierrez, and Krys Kochut, "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques", August 2017.

[8] Chengzhi ZHANG, Huilin WANG, Yao LIU, Dan WU, Yi LIAO and Bo WANG, "Automatic Keyword Extraction from Documents Using Conditional Random Fields", Journal of Computational Information Systems 4:3, 2008.

[9] Santosh Kumar Bharti, Korra Sathya Babu, and Anima Pradhan, "Automatic Keyword Extraction for Text Summarization in Multi-document e-Newspapers Articles", European Journal of Advances in Engineering and Technology, 4(6), 2017.

[10] A. Kogilavani and Dr. P. Balasubramani, "CLUSTERING AND FEATURE SPECIFIC SENTENCE EXTRACTION BASED SUMMARIZATION OF MULTIPLE DOCUMENTS", International Journal of Computer Science and Information Technology, Vol. 2, No. 4, August 2010.

[11] A. Srinivasa Roa, Dr. Ch. Divakar, Dr. A. Govardhan, "RANK BASED DOCUMENT CLUSTERING AND SUMMARIZATION APPROACH IN THE DISTRIBUTED P2P NETWORK", Journal of Theoretical and Applied Information Technology, Vol. 78, No. 2, 20th August 2015.

[12] Ayushi Arya, "A Review Paper on Effective AES Implementation", International Journal of Engineering and Computer Science, Vol. 4, Issue 12, Dec 2015.

[13] Baxendale P. B, "Man-made index for technical literature-an experiment", IBM Journal of Research and Development, 2(4), 1958.

[14] Edmundson H. P, "New Methods in Automatic Extracting", Journal of the Association for Computing Machinery, 16(2), April 1969.

[15] Luhn H. P, "The Automatic Creation of Literature Abstracts", IBM Journal of Research and Development, 2(2), April 1958.

[16] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval", Journal of Documentation, 28(1), 1972.

[17] G. Salton, Edward Fox and Wu Harry, "Extended Boolean information retrieval", Communications of the ACM, 26(11), November 1983.

[18] G. Salton and M. J. McGill, "Introduction to modern information retrieval", McGraw-Hill, 1983.

[19] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval", Information Processing and Management, 24(5), 1988.

[20] H. Wu and R. Luk, K. Wong and K. Kwok, "Interpreting TF-IDF term weights as making relevance decisions", ACM Transactions on Information Systems, 26(3), June 2008.