Intelligent Automation & Soft Computing
DOI: 10.32604/iasc.2022.025235
Article

Tech Science Press

Latent Semantic Based Fuzzy Kernel Support Vector Machine for Automatic Content Summarization

T. Vetriselvi[1,*], J. Albert Mayan[2], K. V. Priyadharshini[3], K. Sathyamoorthy[4], S. Venkata Lakshmi[5] and P. Vishnu Raja[6]

[1]Department of Computer Science and Engineering, K. Ramakrishnan College of Technology, Tiruchirappalli, 621112, India
[2]Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, 600119, India
[3]Department of Information Technology, M. Kumarasamy College of Engineering, Karur, 639113, India
[4]Department of Computer Science and Engineering, Panimalar Institute of Technology, Chennai, 600123, India
[5]Department of Artificial Intelligence and Data Science, Sri Krishna College of Engineering and Technology, Coimbatore, 641008, India
[6]Department of Computer Science and Engineering, Kongu Engineering College, Perundurai, 638060, India
*Corresponding Author: T. Vetriselvi. Email: tvetriselvi21@outlook.com
Received: 17 November 2021; Accepted: 10 January 2022

Abstract: Recently, the bounteous amount of data/information has been available on the Internet which makes it very complicated to the customers to calculate the preferred data. Because the huge amount of data in a system is mandated to discover the most proper data from the corpus. Content summarization selects and extracts the related sentence depends upon the calculation of the score and rank of the corpus. Automatic content summarization technique translates from the higher corpus into smaller concise description. This chooses the very important level of the texts and implements the complete statistics summary. This paper proposes the novel technique that employs the latent semantic analysis (LSA) method where the LSA is derived from natural language processing. Also, it depends upon the particular threshold provided with the device. Statistical feature based model used to compact with inaccurate and ambiguity of the feature weights. Redundancy is removed with cosine similarity and it was presented an enhancement to the proposed method. Finally, fuzzy kernel support vector machine approach of machine learning technique is applied, so this novel model trains the classifier and predicts the statistics summary. This paper focuses to compare together with the another summarization dataset DUC (Document Understanding Conference) like ItemSum, Baseline, Summarizer, Recall Oriented Understudy for Gisting Evaluation (ROUGE) S and ROUGE L on DUC2007. The experiments and result section displays that our proposed model obtains an important performance improvement over the other classifier text summarizes.

Keywords: Automatic content summarization; LSA; redundancy removal; fuzzy kernel; support vector machine

1 Introduction

The recognition of an Internet is expanding radically. An abundant data is accessible on the Internet, it turns into a profoundly tedious and monotonous job to examine the whole content and corpus then acquire

the pertinent data on explicit themes. Content Summarization is recognized as an answer for this problem as it makes programming instructions of the data [1]. Peoples were generally good for this task to understand the correct meaning in the test and can able to mine the features or attributes to get a summarization of the document. Since the automatic text summarization were critical due to the volume of data and less manpower, time to interpret the data. So many reasons are there for telling text summarization is useful for the people:

i) Reading the full content takes more time. In order to overcome this summarization is introduced. To reduce the time
ii) Text statistics summary creates the text projections is easy when analyzing the data in documents.
iii) Efficiency of indexing is improved by automatic summarization method.
iv) The algorithms are likely to become biased than human summarizes in automatic Summarization.
v) In order to process the text documents, abstract data services are introduced in text summarization.

Content Summarization be able to characterized when an abbreviated modification of word identified from at least 1 corpus without losing primary substance or thought of new corpus(s) hence, this is no longer than the incoming text [2] Content Summarization automation have turn into promising way toward deal with data and automatically produces the summary from all the textual data of the theme [3]. The other fields that will be associated to the content summarization like automatic content classification [4] and clustering [5], data recovery and projection, answer querying [6] and words ordering [7].

The main objective of the content summarization framework was able to create compact and easy summary of the provided content with wrapping the significant portion of the content and deal with least redundancy from various incoming sources. There exist various types of content summarization depends upon the input type, depends upon the purpose and depends upon the output type are depicted in Fig. 1. The output type is having two kind of algorithms in which different job has been projected with content summarization. Two kinds of output type algorithms are extractive and abstractive. In this paper, focuses on extractive type summarization where the text is extracted from the corpus without content loss in any of the formats.

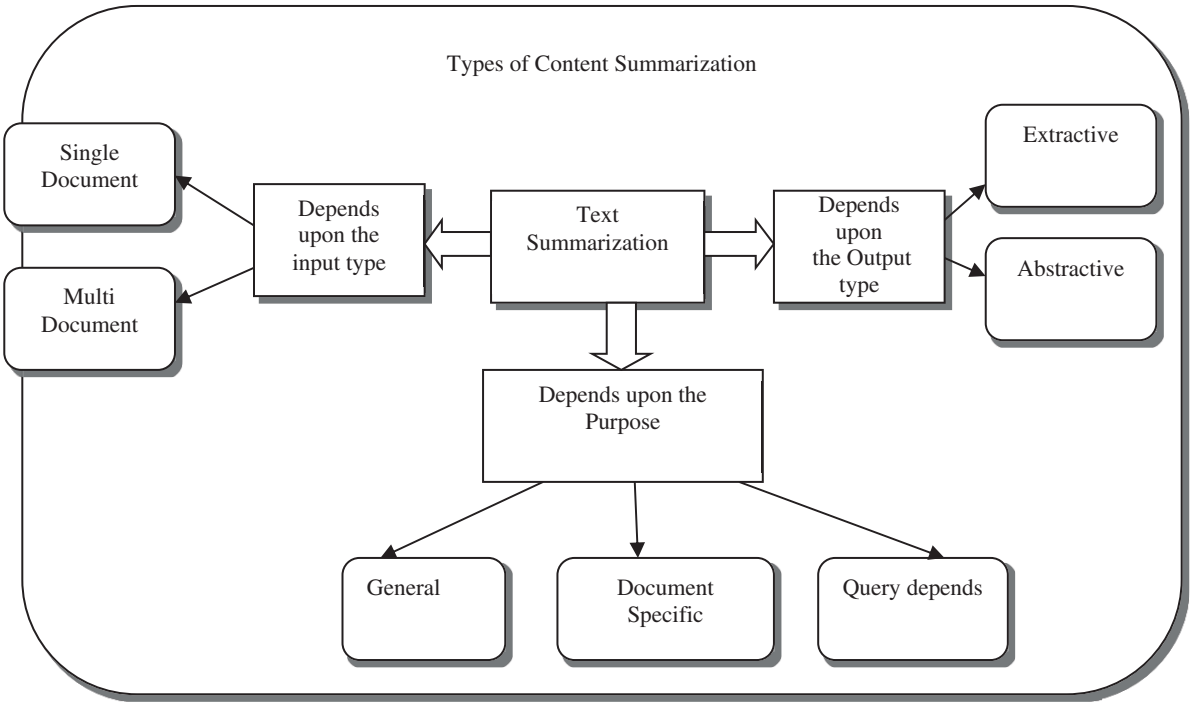

Figure 1: Type of content summarization approaches

Types of Content Summarization Techniques

Content summarization method was classified into different categories.

Depends upon the input type: The existing systems of the text summarization implement a single document summary. Input can be arbitrarily long in multi document,

Depends upon the purpose:

General: No assumption about the field and take care of content to be summarized & takes the incoming text as homogeneous.

Field-specific: It uses field specific pattern to create a more accurate summary.

Query-based: Summary contains data to responds of the input in the natural language.

Depends upon the outcome:

Extractive: Significant words were chosen for getting the summary from the input text.

Most summarization approaches today are extractive in nature.

Abstractive: This method provides its own sentences and phrases to implement the summary in a

Consistent way.

This paper proposes fuzzy kernel support vector machine technique with a multi corpus summarization machine for extracting the significant words to provide a non redundant summary. This extractive summarization system depends on the general summary framework which is provided from more than one relevant document Fig. 1.

The remaining part of the paper is arranged as below. Part two projects previous studies on the content summarization. Part three gives proposed methodology with preprocessing and extracting the significant features. Part four and five explains the fuzzy kernel support vector machine, followed with cosine distance similarity measure utilized to remove the redundancy. Experiments and result outcome analysis is described in Part 6. At end, conclusion and future work projected part 7.

2 Literature Survey

A small number attempt has been made for the movement of content summarization frameworks and utilization of various approaches, methods and devices. The 3 regular methodologies utilized within texts such as statistical methodology, semantics method and mix of both the methodologies are called as hybrid methods [8]. The most important words were extracted with statistical method. It is based on superficial features of content such as topic of the corpus, words position, keywords, prompt words and thematic words etc. it computes the importance of the words and derives the word weight and creates the summary. Validates the results that summary quality will be increased with various combination of superficial content features based on the framework. Statistical approach looks into the syntactic and semantic terms and recognizes the relation between the terms with part of speech (POS), synonymous, vocabulary use, syntax analysis etc. the most effective summary is provided with statistical techniques where as linguistic methods see the semantic usage so it creates average summary by using the hybrid approach.

Traditional techniques depend on statistics to make summary. Anyway a few frameworks are created which take favorable circumstances of the both statistics and linguistics methods. This utilizes semantic similarity for making statistics summary. Based on the hybrid technique, a single document is proposed where the emotion plays a vital part in communication to pass the information very effectively. It uses statistical features tor analyzing the semantic structure.

Many of the researchers concentrate the optimization model for single document classification [9] presented artificial bee colony method for optimization. It increases functions of the whole text and removal of redundancy is taking place while creating the summary. On the other hand, another optimization model is used for multi document summarization [10] presented cat swarm optimization for multi document summarization [11]. Extractive Summary (ES) of single document is proposed for binary classification problem. It utilizes unique statistical features for sentence score. ES methods are very easy to utilize, and this have better focus in the area of content summarization. Normally graph method does not utilize for sentences scoring.

In [12] proposes machine learning algorithms for word ranking which collects sentences relations through graph rank algorithm. It achieves more accurate text summary. Another graph based rank algorithm is also proposed, which utilizes the semantic responsible data for increasing the accuracy of the multi content summarization. Unsupervised technique like clustering strongly investigated under the area of content summarization. The aim of this clustering is to group the document that is considered to be a similar in text [13]. Presented the clustering method which uses WordNet and lexical for acquiring idea of the document. Many of the existing research focus on the relevancy and boundary of the document for solving the redundancy removal and it is taken as a post processing method.

Currently, medical domain contains diversity of summarization models [14,15]. Utilizes 6 various feature selection methods to recognize the significant idea and categorizes words as statistical summary. Also describes the correctness of 2 content summarization methods of automatic creation of abstract starting from the particular medical field. Since, both extractive and abstractive techniques were very helpful to give statistics summary in medical investigation where as an abstractive approach is slightly better than the human perspective [16–18].

Over the past 5 decades, marvelous research has been taking place in the domain of content summarization. Different new techniques and methods like probability &statistical depends, linguistic depends, acyclic graph depends, topic& name depends, discussion depends techniques are incorporated. Different data mining approach, optimization approach& abstractive method then mathematical techniques are implemented for increasing the better quality of the text summary. With studying existing approach, identify the current techniques are over performed then this will be a open problems like at most text coverage, redundancy, word arranging and coupling in the text contents.

This leads to give some encouragement for enhancing the input task of text content summarization; the objective of this proposed work to provide the issues and challenging input task of an Automatic content Summarization of the news feed. In sentence scoring, the following methods such as statistical, data mining and graph based, rank of sentences is applied for extracting the features. The feature weight could be dynamic and not useful. Only few of the attributes are more significant and less have very less significant. So an attributes are not calculated correctly and legitimately.

In order to estimate the calculation of the attribute/feature vector, an LSA with fuzzy kernel SVM is proposed in this proposed work. The main point of LSA was depends upon the concept of natural language processing (NLP) & be able to manage fairly accurate & dynamic data which provides extreme concept of extractive method of text summarization. In this proposed work, redundancy is the major factor as a negative one to affect the text summary quality. In preprocessing, sentences which are considered as the redundant are eliminated from the input for getting better text summary.

3 Proposed Methodology for Multi-Corpus (MC) Summarization

Multi-Corpus summarization framework have additional task when evaluated with single corpus summarization, for example, sentence extraction from various archives, subject recognition, and a

sentence requesting and redundancy reduction. Proposed multi corpus summarization was improved edition of single corpus summarization. In this paper, the following methods are proposed for content summarization.

i) An important content coverage.
ii) Similarity within the content.
iii) The expected or targeted outcome length compression ratio.

3.1 Preprocessing

Few preprocessing steps were mandatory for the raw corpus prior to the multi corpus summarization.

3.1.1 Sentence Segmentation

This is the way toward splitting each and every sentence separately from the documents. Every one of the sentences is extracted from the archives. It is considered to be the task where each text was partitioned into word, unit and topic.

3.1.2 Tokenization

Tokenization is the crucial step that takes place for all the sentences after each of the sentence segmentation. This is the task where each and every word are partitioned from sentence and also utilized to identify the string structure like number, punctuation mark and date time.

3.1.3 Stop Word Removal

The most frequently used words like ‘a’, ‘an’, ‘as’, ‘the’ etc., does not have any semantic data with corresponding documents are rejected. Independent file contains every predefined stop word.

3.1.4 Stemming

Stemming is considered to be the process of translating every word into the corresponding base form by removing the related prefix and suffix.

3.1.5 Tagged Documents

Every corpus special characters such as examination, exclamation and punctuation are removed with a space character efficiency of the various scoring techniques based on the different kind of input text, structure and language of text input [19]. The basic view is having various topics can incline toward various viewpoints; various perspectives will be represented with a different collection of features all of these features classified as word stage and sentence stage features. The important Fig. 2 features that are utilized in the proposed techniques are derived below.

i) Feature Extraction

Efficiency of the various words scoring techniques are computed based on the type of content, the variety of the content semantic structure of the input content. Every text features were categorized as word/sentence types of features. Various statistical features were experimented with data set, then selects best collection of features which will provide the better outcome or results with respect to related domain.

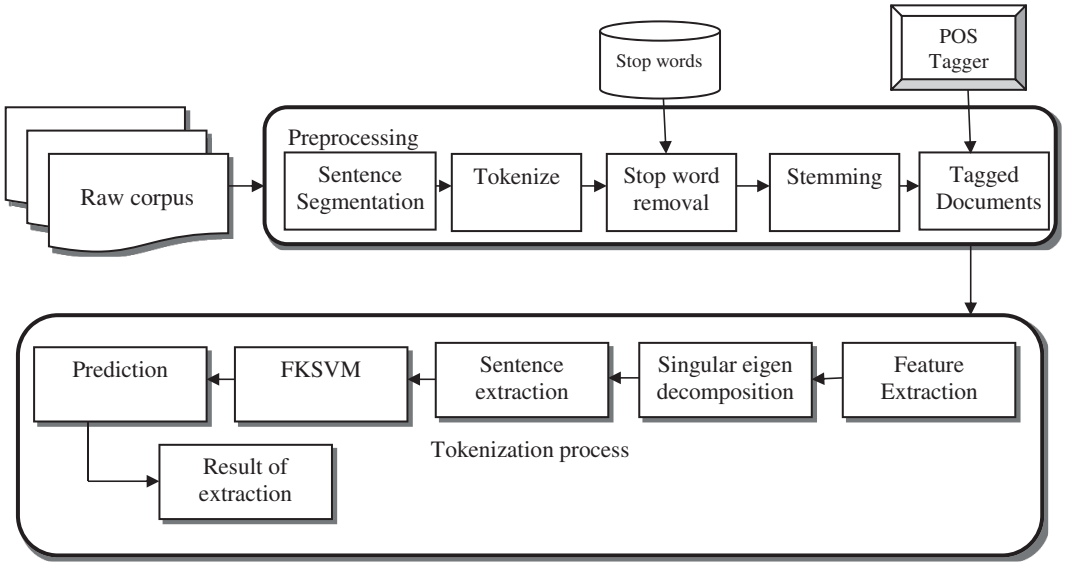

Figure 2: Proposed feature extraction techniques

Word level features

The sentence is having the collection of words. So unique score will be given significant responsibility to provide a sentence score. Each and every various word features are computed as follows

Name-Word: The sentence composed of a word which is identified in the topic of the corpus; these sentences were taken as related text or topic of the text. Number of words in the sentences is computed as word name score. The final score of the name, word (w) in sentence (s) computed by Eq. (1)

$$\text{Name word score } (S_{ws}) = 6/10 \tag{1}$$

Topic-Word: Topic words are the collection of field specific and the most frequent words in a sentence. Here, highest numbers of words are selected in a sentence. The topic-word final score is computed as in Eq. (2)

$$\textit{Topic word score} = 5/10 \tag{2}$$

Named-Entities: Sentence having name entities like a person's name, location is taken as a crucial part in a sentence and it is considered in a summary. The cumulative score of the named entities is calculated as in Eq. (3)

$$\textit{name entities score} = 5/10 \tag{3}$$

Keywords: The high probability of the word is incorporated in the content summary. Normally keywords are considered to be the nouns. Term frequency/inverse document frequency is statistical data by measuring particular word relevant or appears in the document.

Terms 'T' in the document 'D' in a given document set is computed as follows in Eqs. (4) and (5):

$$TF\backslash IDF(t,\ d) = tf(t,\ d) * idf(t) \tag{4}$$

$$IDF(t) = \log\left(\frac{n}{df(t)}\right) + 1 \tag{5}$$

where 'n' is number of documents present in document set.

It is computed with the term of the frequency (TEFR) and inverse document of the frequency (INDOFR) is given as in Eqs. (6) and (7)

$$T_eF_r - I_n\, D_oF_r(C_i) \sum_{te_i}^{C} \mathrm{T_e}F_r \times I_n\, D_oF_r \tag{6}$$

$$T_eF_r - I_n\, D_oF_r(\mathrm{t_{ei}}) = T_eF_r(\mathrm{t_{ei}}) \times I_n\, D_oF_r(\mathrm{t_{ei}}) \tag{7}$$

where t_{ei} denotes ith term T = total number of terms, C = total number of corpus/documents

Statistical data: Numerical information is considered to be the crucial part and it can be incorporated in summary. The final score of statistical data is calculated as in Eq. (8)

$$\mathrm{Statistical - data - score} = 1/10 \tag{8}$$

Sentence Level Features

Sentence features have recognized as two ways like position and length.

Position: The most importance of sentence is location and this can be incorporated in text summary. The final rank in the potions are computed using Eq. (9)

$$\mathrm{Position - score} = 1 - \frac{i-1}{S} \tag{9}$$

where S is the number of total sentence.

Length: Short sentences are neglected. The score is computed using Eq. (10)

$$\mathrm{Length - score} = \frac{number\ of\ word\ occuring\ in\ S}{number\ of\ occuring\ in\ longest\ sentence} \tag{10}$$

3.2 Latent Semantic Analysis

Latent semantic analysis (LSA) is a descriptive statistical method to mine the procedure of sentences with statistical method. It computes high document's text and it equalizes the semantic similarity among the text and words. This is utilized to increase the performance of the data recovery. Various words and terms will be utilized for projects the same semantic conception in LSA. The pair is corpus and the corresponding terms also analyzed by using LSA.

The LSA will be having 3 phases i) generation of matrix ii) singular value decomposition iii) extraction of the sentence. The input is represented by matrix where rows are denoted as words and columns denoted as paragraphs. Each and every cell indicated the importance of sentence in the matrix. The sizes of the input matrix get increases when the complexities get increased. The reduction of matrix is carried out by removing the stop words and tokenization.

The relationship between the word and sentences are shown in the singular value decomposition (SVD). SVD was used to decompose the input into another matrix as given as in Eq. (11)

$$A = U\Sigma\, V^T \tag{11}$$

where A denotes input, which is represented as matrix (x × y), U = words × extracted input (y × y) Σ = diagonal matrix (y × y)

After Fig. 3 implementing the SVD, the matrix which is extracted is utilized for selecting the important sentences. Finally, the higher value denoted the content is much more relevant to the conception. Here the recursive feature rejection is adopted.

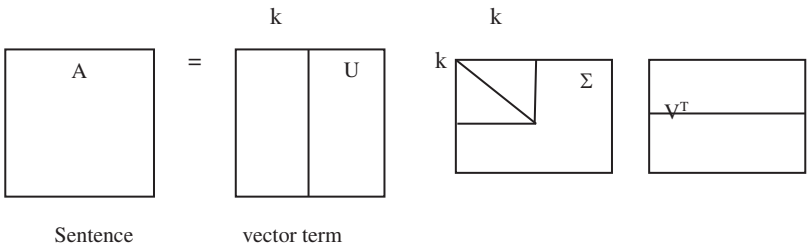

Figure 3: SVD skeleton diagram

3.3 Redundancy Removal

Redundancy removal is the one of the crucial steps in content summarization and this is used to minimize the redundancy from last summary. Here similar text is identified and removed by using correct similarity measures. In this technique, each and every vector word and the similarity between words are computed. In content summarization, cosine similarity measures are one of the most commonly used similarity measures. Two sentences are denoted in Eq. (12)

$$S_x = \{w_{x1},\ w_{x2},\ \ldots w_{xn}\}\ and\ S_y = \{w_{y1},\ w_{y2},\ \ldots w_{yn}\} \tag{12}$$

Distance similarity measures of cosine among the 2 sentences are denoted as in Eq. (13)

$$S(S_x,\ S_y) = \frac{\sum_{i=1}^{n} w_{xi} w_{yi}}{\sqrt{\sum_{i=1}^{n} w_{xi}^2 \cdot \sum_{i=1}^{n} w_{yi}^2}} \tag{13}$$

where, $w_{xi}w_{yi}$ denotes the weight of the terms in sentences S_x and S_y. The weight is related with t_{ei} in each sentence. In the experiment and result section, applied distance similarity cosine function to calculate the highest scoring part of the sentences. Cosine similarity is calculated by python tool.

3.4 Fuzzy Kernel Support Vector Machine

Machine learning algorithm one such type of supervised learning to do a class label classification in terms of binary classification using Support Vector Machine (SVM) implements the binary classification. The primary aim is to create a hyperplane with training dataset and that will be used as model for incoming dataset is displayed in Fig. 2. Then the test data set applied for validation of the input test set. Vapnik chervonenkis implements the method of SVM where it uses the technique of supervised learning algorithm like classification and regression. The very best hyperplane was chosen for calculating the margin which is considering as a maximum with nearest dataset is denoted as support vectors. With increasing the common characteristics, hyperplane margin which is maximal was utilized for building capability.

The building of SVM hyperplane is denoted with Eq. (14)

$$\mathrm{W}^T X + b = 0 \tag{14}$$

Here, W denotes weight input vector and b denotes offset parameter. The margin within the hyperplane described as maximum.

Term margin in the SVM is process of computing the perpendicular distance between hyperplane in observing training data set. shortest distance is referred as minimal or margin in hyperplane. linear support vectors and its kernel trick used to diminishes the overall complexity of classification. The non linear margin function was denoted as in Eqs. (15) and (16)

$$f(x) = sign\ (d(x)) \tag{15}$$

$$d(x) = \sum_{n=1}^{n} \alpha_i y_i(x,\ x_i^n) + \quad b \tag{16}$$

Here, d(x) denotes the Euclidean distance function, α_i represents the lagrange multiplier, number of support vector as n and finally b is the parameter

The kernel trick function $K(x,\ x_i^n)$ calculate the nonlinear argument function Eq. (17)

$$x \rightarrow \varphi(x) \tag{17}$$

Here the kernel trick $K(x,\ x_i^n)$ function generally utilized for content summarization and kernel represented with Radial Basis Kernel Function (RBKF), the equation is as per Eq. (18)

$$K(x,\ \ x_i) = exp\left(\frac{||x - x_i||^2}{2\sigma^2}\right) \tag{18}$$

Two important functions are taking when radial basis kernel was taken as the parameter 1. Kernel parameter 2. Trade off. These 2 parameters are used for better performance with n fold cross validation. Each and every these type of data will be taken for training the input data.

3.5 Fussy Kernel Support Vector Machine (FKSVM)

Based on the amount of data, the best hyperplane is constructed with SVM then it provides the outliers in the training data. To The optimal hyperplaneis calculated by the SVM classifier depend upon the low amount of data. It leads to errors or outliers in training data. To overcome this problem, FKSVM is proposed with fuzzy SVM membership of input data. FKSVM was projected for concentrating the maximal margin of SVM but need to take care of outliers with less membership can prevents the artifacts to create data in terms of high probability.

For example, the aim of binary classification by training data denoted as [X1, y1, m1],…,[Xt, yt, mt]. Then the training set will be as per Eq. (19)

$$X_i \in \Re^N \tag{19}$$

Output label yi ϵ +1, −1 & fuzzy membership value mi $\epsilon[\sigma, 1]$.

I denoted as i = 1,…m when $\sigma > 0$

The training data is equal to 0 provides empty

The minimum error term consider as the optimal hyperplane which issued to measure the outlier in SVM. In order to minimize the outlier or error function as Eq. (20)

$$\frac{1}{2}\boldsymbol{W}.\boldsymbol{W} + C\ \sum_{i=1}^{t} m_i \xi_i \tag{20}$$

With respect to Eq. (21)

$$\boldsymbol{y_i}(\boldsymbol{W}.\boldsymbol{x_i} + b) \geq 1 - \ \xi_i \tag{21}$$

Here, $i = 1 \ldots t$

Kuhn-Tucjer conditions and Lagrange multiplier was utilized for calculating the best hyperplane. The Lagrange multipliers are a method used for finding the best local maxima and minima with equality

constraints on the functions. Here more than one equations are need to be satisfied for the given input variables.

The selection of best hyperplane is very critical task so the aim is to execute the membership value which is very useful and appropriate for the input data. This is mainly depends upon their classes. The analysis of the data and result are projected to calculate the accuracy of the system.

4 Experimental Setup and Result Analysis

The proposed work is validated with DUC 2007 Dataset it uses text for multi-document summarization. Here DUC 2007 has huge types of magazines of various people. All of the DUC 2007 articles are grouped related with corresponding subjects and each and every group was taken as input to this proposed automatic content summarization framework. DUC 2007 gives 4 people created statistics summary as a reference or normal statistics summary. Participants are evaluated the summary with normal summary for validation. The size of the summary may be not more than 100 word content. It contains 10 headings each is categorized with 3 groups: X, Y and Z. traditional summarization task is performed with group X. the generation of updated summary with read X is performed in group Y simultaneously group C is summarized with group A and B. Various summary is written with 4 people the overall rank of group is estimated with just mean rank of the rank of the group.

ROUGE denotes recall-oriented understudy for gisting evaluation. This is a collection of measures and metrics of validating automatic content summarization as well as the system transformation. This will work by evaluating the automatic content summarization and transformation against the collection of base references (typically human-produced).

Performance Metrics

Simplistically take, Recall of ROUGE describes the amount of base reference point summary & statistics summary retrieving? it can be computed as per Eq. (22)

$$recall\ of\ \ ROUG = \frac{number\_of\_overlapping\_words}{total\_word\_in\_reference\_summary} \tag{22}$$

This discusses that every words in reference summary will be acquired with system statistics summary. However, it will not give the next side of the experiment. A system that generated a statistics summary (system summary) will be very huge, acquiring every sentence in the base reference point text summary. However, many sentences available in system might consider as useless, creating statistics summary unnecessarily is more demonstrative. So next metric considered. In precision, the amount of data which is measured and relevant is taken. Precision can be measured with Eq. (23)

$$precision = \frac{number\_of\_overlapping\_words}{total\_words\_in\_system\_summary} \tag{23}$$

ROUGE-L–used to measure the highest matching sequence. This is not taking care of built in sentence order because it takes general highest n grams automatically.

ROUGE-S–This is also can be denoted as skip gram correlation. It used to measure the gaps between the sentences.

In Tab. 1 proposed automatic content summarization with other methods is described with various peer value systems. The parameters such as precision and recall are considered for evaluation. In order to find out the various performance improvements against other DUC data set, the correlated p-value and t-test is depicted in Tab. 2 where the proposed method produces the most important outcome compared with

other methods. Tab. 3 shows the results of DUC2007 dataset where group such as X, Y and Z is validated Fig. 5.

Table 1: Comparison between automatic content summarization with other type of summary

Summary	ROUGE S		ROUGE L		ROUGE L group
	Precision	Recall	Precision	Recall	Recall
Peer 16	0.082	0.084	0.081	0.084	0.4
Peer 19	0.083	0.089	0.084	0.085	0.35
Peer 2	0.081	0.082	0.082	0.081	0.36
Peer 24	0.092	0.091	0.090	0.094	0.32
Peer 25	0.092	0.097	0.091	0.092	0.347
Peer 26	0.085	0.084	0.082	0.081	0.38
Peer 28	0.091	0.090	0.089	0.088	0.342
Peer 29	0.090	0.089	0.088	0.089	0.30
Peer 3	0.078	0.089	0.079	0.078	0.36
ItemSum	0.085	0.084	0.085	0.086	0.38
Baseline	**0.093**	0.092	0.091	0.092	0.37
Summarizer	**0.095**	**0.096**	**0.094**	**0.095**	**0.389**
Proposed multi-doc-summary	0.078	**0.166**	**0.096**	**0.095**	**0.39**

Table 2: Related value of p-values and t-test

Methods	ROUGE S	ROUGE L
Proposed *vs.* Peer 16	1.8e−2	2.3e−3
Proposed *vs.* Peer 19	5.2e−3	5.3e−2
Proposed *vs.* Peer 2	4.2e−2	3.4e−3
Proposed *vs.* Peer 28	5e−2	7.1e−3
Proposed *vs.* Peer 29	6e−3	5.2e−2

Table 3: Results on DUC2007 dataset

Group	ROUGE L
Group X	0.389
	0.368
	0.376

(Continued)

Table 3 (continued)

Group	ROUGE L
Group Y	0.378
	0.376
	0.3754
Group Z	0.398
	0.365
	0.375
Average	0.38
	0.398
	0.387

Figs. 4 to 7 gives chart comparison of ROUGE S and ROUGE L calculated rank with various metrics like precision and recall of text summarization is computed.

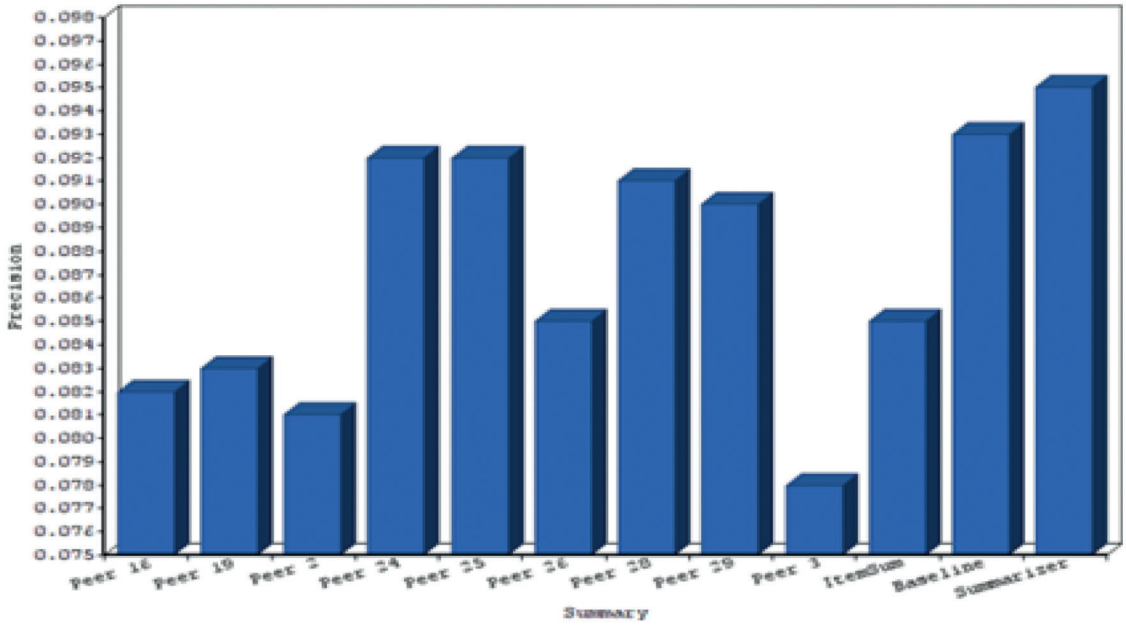

Figure 4: Chart for rouge s-precision

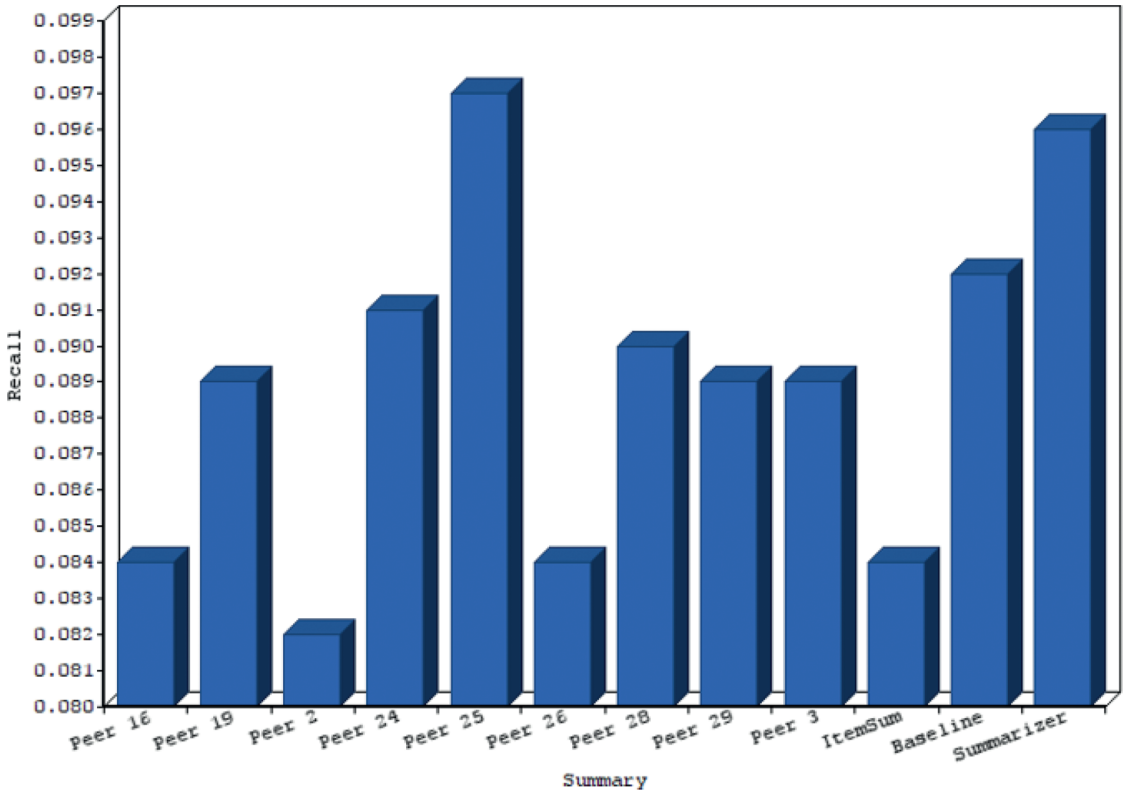

Figure 5: Chart for rouge s-recall

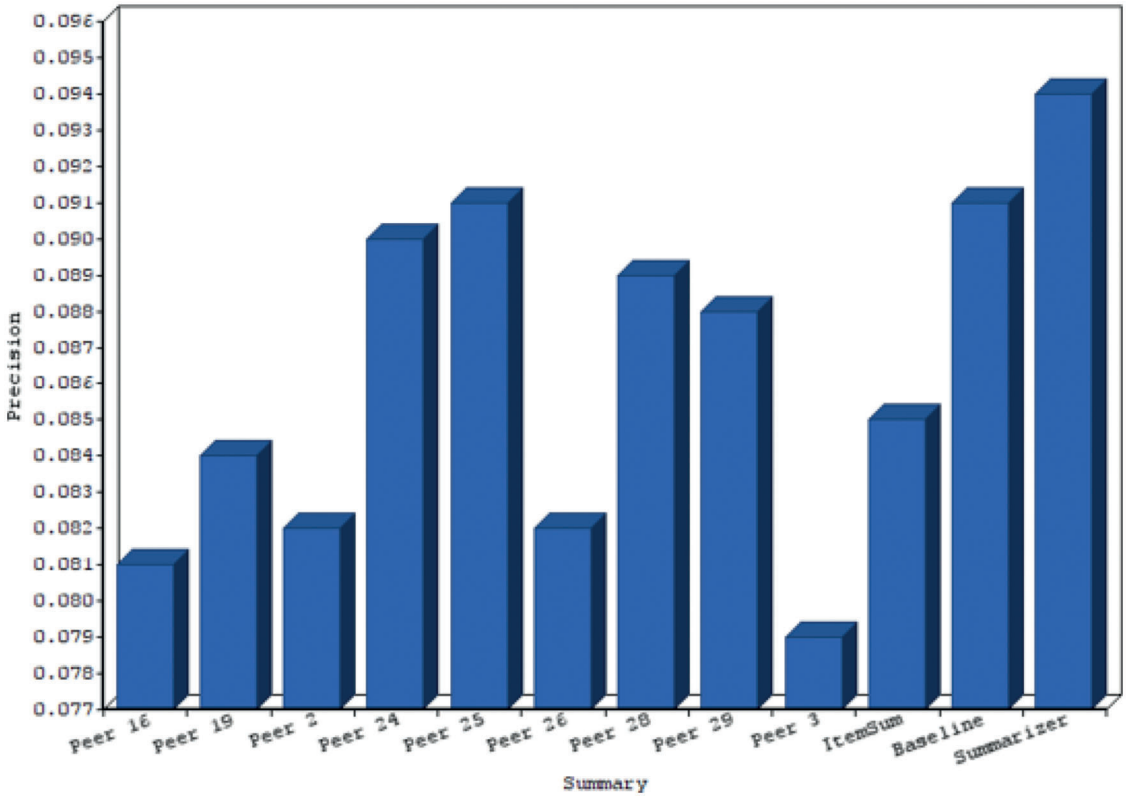

Figure 6: Chart for rouge precision

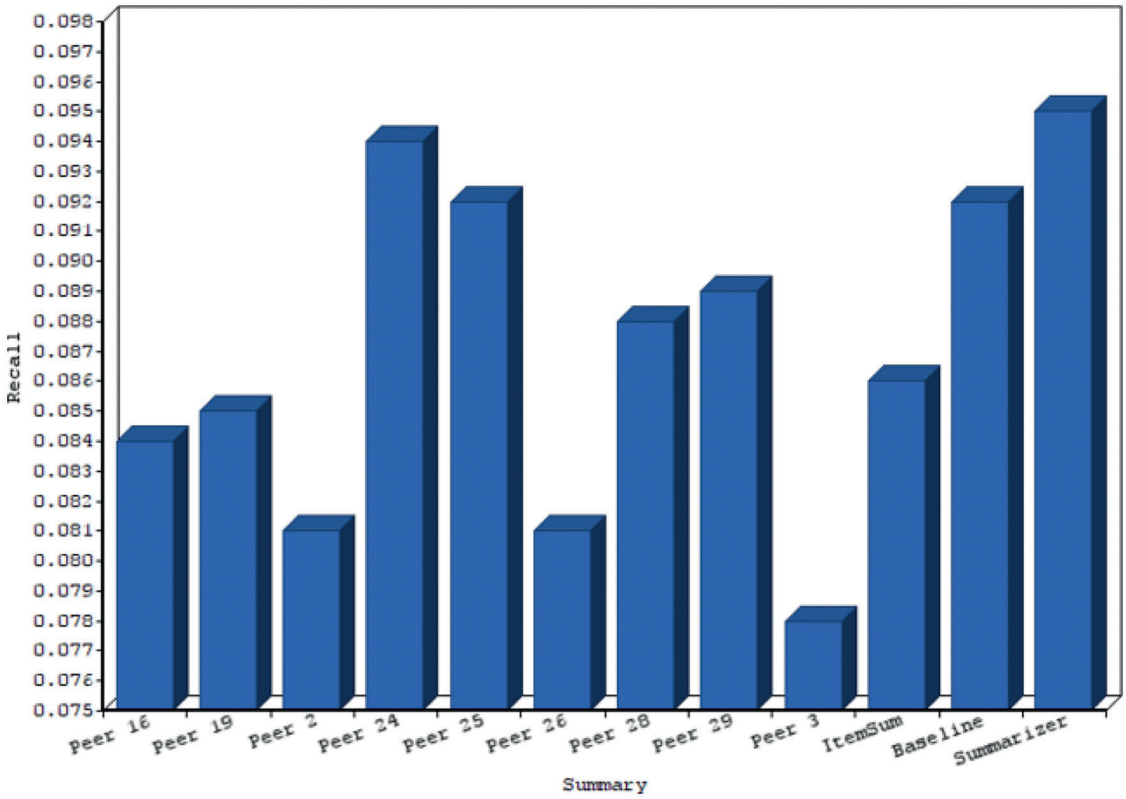

Figure 7: Chart for rouge recall

5 Conclusions and Future Work

The proposed work implements the automatic content summarization where general extractive summary technique is applied. Fuzzy kernel SVM is utilized for improving the performance of the content summarization. After completing the preprocessing, the entire words/sentences scoring is validated, then these words were sorted with decreasing order depends on the rank of the system. Latent semantic analysis (LSA) is a descriptive statistical method for mining the relevant sentences/words with statistical values. Redundancy is only main problem of multi content summarization. Cosine distance measure similarity is applied for rejecting the words that is having same content from the extracted words/sentence for creating text end summary. All of the experimentation was validated with DUC 2007 data set. The result gives performance comparison.

Additionally, many ways are considered for enhancing the present work. This technique was validated with the new dataset, these processing is better when it is used in text corpus. The proposed automatic content summarization device may be improved for creating the appropriate statistics summary through adding some statistics and linguistics features. Multi-document and content summarization has become more likelihood of uncertainty than the single document summary. Lexical chain method will mainly use to rectify the issue. Increasing the system with morphological analyzer, a lexical database and semantic tools with statistical methods is projected in future. Mainly the significant implementation of Sentence Ordering will also be very difficult and it is useful for summarization and interviewing system.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] J. M. Sanchez Gomez, M. A. Vega Rodríguez and C. J. Pérez, "Extractive multi-document text summarization using a multi-objective artificial bee colony optimization approach," *Knowledge-Based Systems*, vol. 159, pp. 1–8, 2018.

[2] R. M. Aliguliyev, "A new sentence similarity measure and sentence based extractive technique for automatic text summarization," *Expert Systems with Applications*, vol. 36, no. 4, pp. 7764–7772, 2009.

[3] M. Moradi and N. Ghadiri, "Different approaches for identifying important concepts in probabilistic biomedical text summarization," *Artificial Intelligence in Medicine*, vol. 84, pp. 101–116, 2018.

[4] T. Wei, Y. Lu, H. Chang, Q. Zhou and X. Bao, "A semantic approach for text clustering using WordNet and lexical chains," *Expert Systems with Applications*, vol. 42, no. 4, pp. 2264–2275, 2015.

[5] E. Yulianti, R. C. Chen, F. Scholer, W. B. Croft and M. Sanderson, "Document summarization for answering non-factoid queries," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 1, pp. 15–28, 2017.

[6] D. Bollegala, N. Okazaki and M. Ishizuka, "A preference learning approach to sentence ordering for multi-document summarization," *Information Sciences*, vol. 217, pp. 78–95, 2012.

[7] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.

[8] C. S. Yadav and A. Sharan, "Hybrid approach for single text document summarization using statistical and sentiment features," *International Journal of Information Retrieval Research (IJIRR)*, vol. 5, no. 4, pp. 46–70, 2015.

[9] R. Rautray and R. C. Balabantaray, "Cat swarm optimization based evolutionary framework for multi document summarization," *Physica A: Statistical Mechanics and its Applications*, vol. 477, pp. 174–186, 2017.

[10] M. Mendoza, S. Bonilla, C. Noguera, C. Cobos and E. León, "Extractive single-document summarization based on genetic operators and guided local search," *Expert Systems with Applications*, vol. 41, no. 9, pp. 4158–4169, 2014.

[11] C. Fang, D. Mu, Z. Deng and Z. Wu, "Word-sentence co-ranking for automatic extractive text summarization," *Expert Systems with Applications*, vol. 72, pp. 189–195, 2017.

[12] S. Yan and X. Wan, "SRRank: Leveraging semantic roles for extractive multi-document summarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2048–2058, 2014.

[13] K. Venkatachalam, A. Devipriya, J. Maniraj, M. Sivaram, A. Ambikapathy *et al.,* "A novel method of motor imagery classification using eeg signal," *Artificial Intelligence in Medicine*, vol. 103, pp. 101787, 2020.

[14] A. Louis and A. Nenkova, "Automatically assessing machine summary content without a gold standard," *Computational Linguistics*, vol. 39, no. 2, pp. 267–300, 2013.

[15] K. Yasoda, R. Ponmagal, K. Bhuvaneshwari and K. Venkatachalam, "Automatic detection and classification of EEG artifacts using fuzzy kernel SVM and wavelet ICA (WICA)," *Soft Computing*, vol. 24, no. 21, pp. 16011–16019, 2020.

[16] P. Prabu, A. N. Ahmed, K. Venkatachalam, S. Nalini and R. Manikandan, "Energy efficient data collection in sparse sensor networks using multiple mobile data patrons," *Computers & Electrical Engineering*, vol. 87, pp. 106778, 2020.

[17] R. Ferreira, L. de Souza Cabral, F. Freitas, R. D. Lins, G. De França Silva *et al.,* "A Multi-document summarization system based on statistics and linguistic treatment," *Expert Systems with Applications*, vol. 41, no. 13, pp. 5780–5787, 2014.

[18] R. Ferreira, F. Freitas, L. de Souza Cabral, R. D. Lins, R. Lima *et al.,* "A context based text summarization system," in *Proc. 2014 11th IAPR Int. Workshop on Document Analysis Systems*, Tours, France, IEEE, pp. 66–70, 2014.

[19] A. Belousov, S. Verzakov and J. Von Frese, "A flexible classification approach with optimal generalisation performance: Support vector machines," *Chemometrics and Intelligent Laboratory Systems*, vol. 64, no. 1, pp. 15–25, 2002.