# Design and Development of an Automatic Text Summarization Using Pragmatic-Enabled Features with LMS based Neural network

**Ch. Sita Kameswari, J.A. Chandulal**

*Professor, Departmentof CSE, BABA institute of technology and sciences, Andhra Pradesh, India*
*Principal, Swarnandhra College of Engineering Technology, Andhra Pradesh, India*

---

**ABSTRACT**

The main intend of the text summarization is to generate a condensed version of one or more texts using computer techniques. This will help reader to decide if a document contains needed information with minimum effort and time loss.In past decades, a number of literatures have been presented different text summarization techniques. In this paper, we have proposed a text summarization approach combining pragmatic-enabled features and LMS based neural network. At first, the preprocessing steps are applied through pragmatic analysis. In this step, the text contents are filtered using WorldNet dictionary. After that, four set of features like asTitle Similarity, Positional Feature, Term Weight and Concept Feature are extracted and feature matrix is generated. Once feature matrix is created, text summarization is done via LMS based neural network. The summary of the document is created based upon the score level using LMSNN.The performance of the proposed approach is evaluated through precision, recall and f-measure. Experiment result shows that the pragmatic analysis based text summarization method provides better performance than existing method.

**Keywords:** *pragmatic analysis, LMS based neural network, WorldNet, feature*

## 1. INTRODUCTION

With the rocking advancement of the internet cutting across all barriers of language, there are a zooming number of people browsing through the cyberspace [1]. Laden with current data burden, the inquisitive investigators find it a Herculean Task to keep abreast of the hi-tech advancement appearing like a blitz every second in every domain, it has become all the more necessary to identify a piece of data only when essential. By extracting the quintessence of data, the automatic summarization lends a helping hand to the humans to tackle the enormity of the data. [2]. Summarization is compiled as per the equation: "summarization = topic identification + interpretation + generation". For the purpose of recognition, the target is to sprain the input to reschedule only the most noteworthy, and vital subjects. In order to clarify, the objective is to perform the compaction by means of re-interpretation and fusion the extracted topics into an abridged one [3].

In this regard, the Automatic Text summarization may be broadly segmented into two categories such as the extraction and abstraction [4]. By 'extraction' what is meant is the choice of the phrases or sentences possessing the maximum score from the original text and their integration to generate the novel abridged text without any alteration in the source text. On the other hand, the abstraction involves the probe and interpretation of the text by means of linguistic techniques. In a large majority of the cases, extraction

approach is employed to generate the summary in automated text summarization mechanism. Many an investigator has identified the automated part of speech (POS) tagging, sentence feature computation and score generation for every sentence as the most common challenges [6] in the text summarization. The extraction, in turn, includes concatenating extracts collected from the corpus into a synopsis, whereas the abstraction constitutes the generation of innovative sentences from the data extracted from the corpus. It is found in the backdrop of the multi-document summarization of news articles, that the extraction is likely to be misfit as it may generate summaries which are excessively wordy or prejudiced in favor of certain sources [5, 7].

In particular, the model of text summarization has an integer linear programming problem. In [16], a method is proposed an automatic summarization approach based on the analysis of review articles' internal topic structure to assemble customer concerns. A trainable summarizer, which takes into account several features, including sentence position, positive keyword, negative keyword, sentence centrality, sentence resemblance to the title, sentence inclusion of name entity, sentence inclusion of numerical data, sentence relative length, Bushy path of the sentence and aggregated similarity for each sentence to generate summaries is discussed in [17]. A document summary is useful since it can give an overview of the original document in a shorter period of time [18]. The main goal of a summary is to present the main ideas in a document/set of documents in a short and readable paragraph. Summaries can be produced either from a single document or many documents [18, 19].

Most of the previous studies on the sentence extraction-based text summarization task use a graph-based algorithm to calculate the saliency of each sentence in a document and the most salient sentences are extracted to build the document summary[9]. The sentence extraction techniques give an indexing weight to the document terms and use these weights to compute the sentence similarity [9, 20] and/or document centroid [9, 19] and so on. The sentence similarity calculation remains central to the existing approaches. The indexing weights of the document terms are utilized to compute the sentence similarity values. In [10], the authors have presented an anatomy-based summarization method called Topic Summarization and Content ANatomy (TSCAN), which organizes and summarizes the content of a temporal topic described by a set of documents. As internet is growing, the ratio of people using it, is increasing without having any language barrier.

In this paper, we have proposed a document summarization system using LMS based neural network. At first, a preprocessing process is applied via pragmetic anlaysis. Then, a feature extraction scheme is carried out via four set of features like title similarity, positional feature, term weight, concept feature. Once feature extarction is completed, the feature matrix is constructed. Afterthat, text summuarization is done through LMS based neural network.The basic organization of the paper is as follows: Section 2 presents the review of literature survey and the proposed system model is explained section 3. The result and discussion part is presented in the section 4 and the conclusion part is given in section 5.

## 2. LITERATURE REVIEW

A handful of research related to text summarization and different methods used for text summarization are plotted in the following section.It was Hien and Eugene [8] who shelled out the effect of the cognitive fashions of the user, while analyzing the multi-document summaries. Especially, they shortlisted the two

vital dimensions forming part of the cognitive fashion of the user such as, analytic/wholist and verbal/imagery dimensions. Accordingly, they investigated their effects on the way the user evaluated a summary which was produced from a set of documents. Particularly, the category of a document set indicated whether the content of set was slackly or strongly connected.

Moreover, Pawan *et al* [9] proficiently propounded a context sensitive document indexing model based on the Bernoulli model of randomness. The Bernoulli model of randomness, in turn, was employed to locate the prospect of the co-occurrences of two terms in a huge corpus. A novel technique employing the lexical linkages between the terms to furnish a context sensitive weight to the document terms was envisioned. The innovative sentence similarity measure was widely employed with the baseline graph-based ranking models for sentence extraction.

Chin and Meng [10] characteristically defined a topic as a decisive event or activity together with the entire openly connected events and activities. It was characterized by a chronological series of documents published by several authors on the web. In their document, a task known as the topic anatomy was defined which summarized and linked the central segments of a topic for a temporary period so that readers were able to comprehend the content without any difficulty. The innovative topic anatomy model was afforded the pet name the TSCAN, which gathered the vital themes of a topic from the eigenvectors of a temporal block association matrix. Subsequently, the important events of the themes and their summaries were extorted by assessing the configuration of the Eigen vectors. In the long run, the extorted events were linked in terms of their temporal proximity and context resemblance to generate an evolution graph of the topic.

Additionally, Chin Liu *et al* [11] achieved world-wide acclaim for launching a movie-rating and review-summarization technique in a mobile scenario. The movie-rating information was mainly dependent on the sentiment-classification outcomes. The abridged portrayal of movie reviews was produced from the feature-based summarization. An innovative technique in accordance with the latent semantic analysis (LSA) approach was green-signaled to locate the product characteristics, based on which, a novel method was envisaged to cutback the dimension of the summary. The sentiment-classification precision and the system feedback duration were taken well-care of while configuring the novel mechanism.

Similarly, Feng Yang [12] fantastically gave vent to the method for satisfying the user requisites which played a leading role in the investigation of the query-oriented automatic summarization. By deftly blending the impact of text granularities and query data, they were able to configure a novel model to elucidate the linkages between the text granularities. They also endeavored to point out the way in accordance with the semantic association to assess the text similarity and followed an innovative technique to dynamically allocate the sentences for the generation of the summarization.

Nowshath *et al* [13] were instrumental in proposing the feature extraction as the most significant issue to be tackled in algebraic based Automatic Text Summarization (ATS) techniques. In this regard, the most critical role of any ATS included the recognition of most significant sentences from the specified text, which could be a reality only when the accurate characteristics of the sentences were appropriately detected. Accordingly, in their document, they elegantly launched an innovative Conditional Random Field (CRF) based ATS which were competent to detect and extort the appropriate characteristics which

was vital challenge which plagued the Non-negative Matrix Factorization (NMF) based ATS. Their investigation led to the launch of a trainable supervised technique.

Moreover, Hitesh and Durga [14] were instrumental in deftly designing an innovative Aspect Based Sentiment Analysis and Summarization (ASAS) System, which successfully tackled the context dependent opinion words which were found to trigger several challenges. At the outset, with an eye on locating the opinion polarity, an online dictionary was employed for organizing the context independent opinion word. Thereafter, a natural linguistic rule for allocating the polarity was made use to a large majority of context dependent words, thereby leading to the generation of the training data set. Subsequently, for the organization of the residual opinion words, they resorted to the employment of the opinion words and feature jointly instead of deploying the opinion words singly, as the identical opinion word could have divergent polarity within the identical realm. Thereafter, the Interaction Information technique was utilized to organize the feature-opinion couples. Further, sensing the critical role played by the negation words, they were utilized to spin the polarity of the matching opinion word. In the long run, after orchestrating each and every opinion word, the system went on to create a concise summary for the specific product based on each and every characteristic.

The credit goes to Han Zhang for the efficient launch of the Automatic summarization [15] intended to offer assistance in organizing the outcomes of the biomedical data retrieval mechanisms. In this regard, the Semantic MEDLINE summarized the semantic predications characterizing the assertions in MEDLINE citations. The outcomes were pictorially represented by means of a graph preserving associations to the original citations. However, it was very difficult to go through the graphs which summarized a whopping 500 plus citations. The innovative method was in accordance with the degree centrality, which estimated the associations in a graph. In the end, four types of clinical theories linked to the cure of disease were located and offered as a summary of input text.

## 3. PROPOSED ALGORITHM FOR TEXT SUMMARIZATION

Due to increasing the text data avaliable over the internet it becomes difficult for users to find the desired information quickly. Automatic text summarization solves the problem by genrating summarizes that could be used as a condensed replica of a documeits. For that reason, automatic text summarization can be defined as the process condensing the source text document or set of documents while retaining main information contents using a automatic machine.In past years, a number of literatures have been presented for this process. In this research, we have proposed a document summarization system using LMS based neural network.The scematic diagram of proposed system is presented in figure 1.

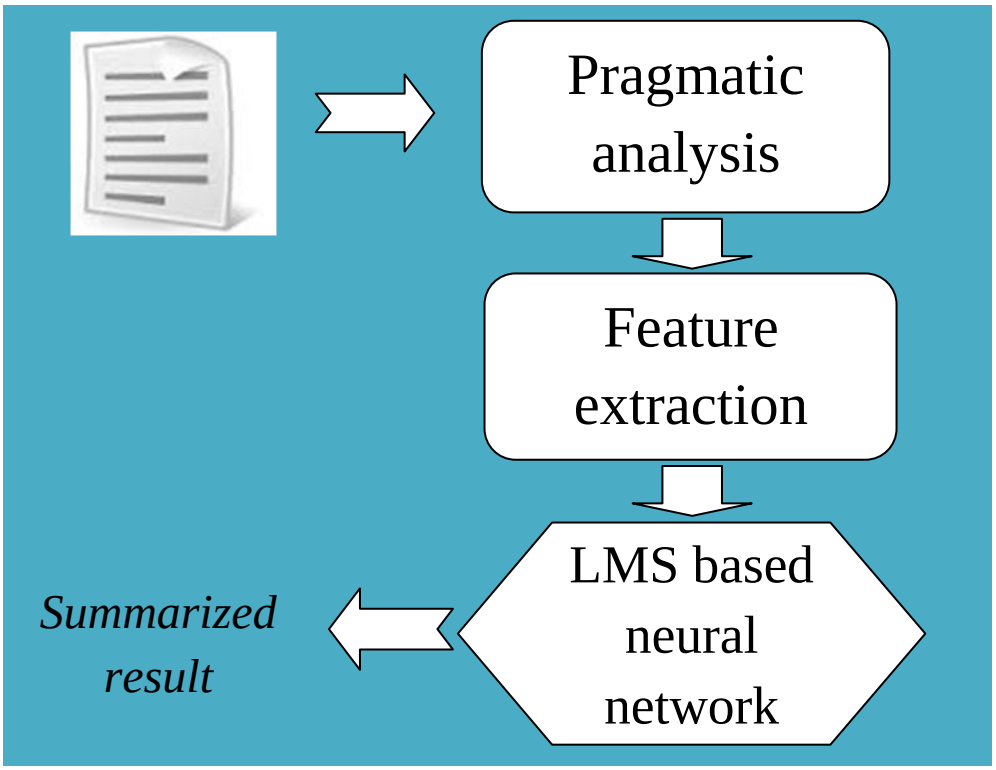


**Figure 1:** Illustration of the proposed approach

**A. FEATURE COMPUTATION**

1. **Filtering via pragmatic analysis**

Before the feature compution, pragmatic analysis is carried out here to find the important topical words along with context-aware.Pragmatics can be defined as the study of the meaning in context. It is an effort to get the intended meaning of text. In this step, a content filtering analysis is carried out to improve the summraization process. In order to do this, Wordnet dictionary based distance computation is proposed to filtering the documents. Wordnet is the one of the most widely used and largest lexical databases of english.In general as a dictionary, WordNet covers some specific terms from every subject related to their terms. It maps all the stemmed words from the standard documents into their specifies lexical categories. The pragmatic anlaysis consists of two stages: (1) Word positioning in Wordnet, (2) Word distance determeination and Content filtering.

- ***Word position***

Words position is hierarchial structure of Wordnet structure of Wordnet dictionary is found in this step. At first, each sentence is taken from the document. Subsequently, each important word searching process is done on the wordnet dictionary.In order to this, Hypernyms option is used in each word search. By using this option, parent of each word is determined. For instance,a sentence "*WSN is vulnerable to*

*various problems related to security*" is taken. Then, parent of each important word is determined. For example, searching "security" will result three parents like "protection", "cretirficate" and "surety".

- ***Word distance determination and filtering:***

From previous step, each important word consists of some parents words. Consequently, each meaning word finds deapth with sentecne other words. The most deapth or related words are replaced in original sentence. For example, as desribed above, searching "security" will result three parents like "protection", "cretirficate" and "surety". Then, each word like "protection" or "cretirficate" or "surety" finds deapth with sentence words. From the deapth results, we get more deapth for word "protection". Therfore, the original sentence is changed as "*WSN is vulnerable to various problems related to protection*". The schematic diagram of thepragmatic analysis is illustrated in figure2.

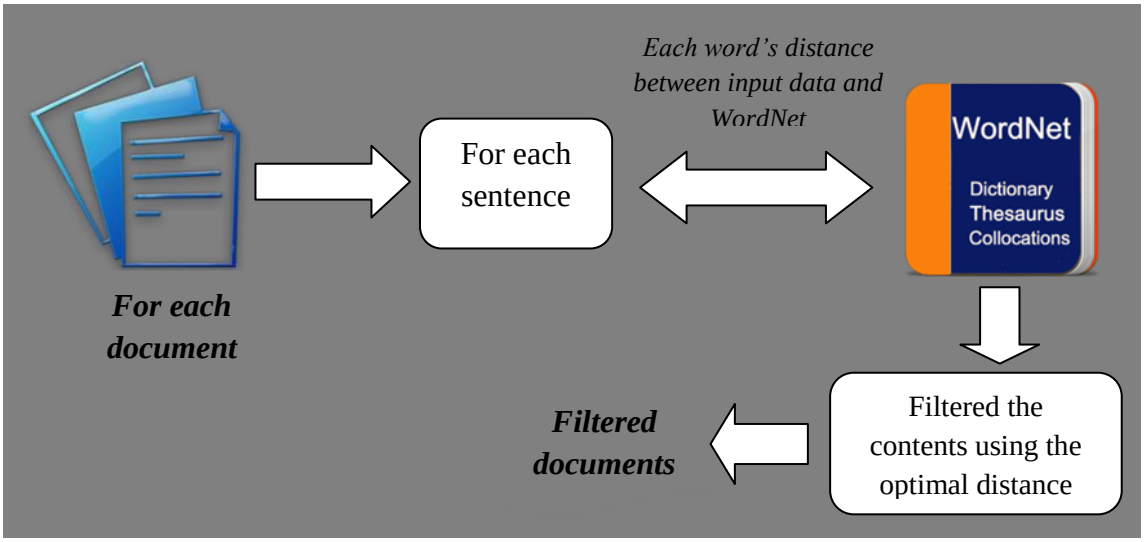


**Figure 2:**Content filtering through pragmatic analysis

## 2. Feature extraction

The text document is represented by set, $TD = \{S_1, S_2, ..., S_k\}$, where, $S_i$ signifies a sentence contained in the text document $TD$. The filtered document is given to the feature extraction. The significant word and sentence features to be used are decided. This work uses features like as Title similarity, Positional feature, term weight and concept feature.

- ***Title Similarity***

A sentence is deemed to be significant, if it is identical to the title of the text document. For our purposes similarity is construed based on the incidence of common words in the title and the sentence. A sentence is equipped with excellent feature score if it contains the maximum number of words common to the title. The ratio of the number of words in the sentence which crop up in title to the total number of words in the

title enables us to compute the score of a sentence for this feature. It is estimated by means of the following expression:

$$F_1 = \frac{S \cap T}{T} \quad (1)$$

Where,

$S$ →Set of words of sentence
$T$ →Set of words of title
$S \cap T$ →Common words in sentence and title of document

- ***Positional Feature***

Further, the positional value of a sentence is also extrated. Whether a sentence is relevant or not is decided by its position in the text. To estimate the positional score of sentence, the following parameters are taken into account:

$F_2 = 1$, if sentence is the starting sentence of the text
$F_2 = 0$, if sentence comes in the middle paragraphs of text
$F_2 = 1$, if sentence comes in the last of the text

- ***Term Weight***

The term weight constitutes a very significant feature to be taken into account for the summarization of the text. Incidentally, by the term 'weight' what is construed is the term frequency and its significance. Further it is deemed as the most typical feature in severa; natural language processing functions. The frequency, in this case, represents the term frequency throwing light on the relevance of of a particular word in a document, and basically reveals the number of occasions a word occurs in the text. The term frequency of a word is expressed by the expression $if(f,d)$ where $f$ refers to the frequency of the word and $d$ indicates the text document. The total term weight is estimated by calculating the $tf(f,d)$ and the $idf$ for a document. Now, the idf indicates the inverse document frequency which gives us a hint of whether the term is frequent or uncommon across all the documents. It is estimated by dividing the total number of documents by the number of documents containing the term and thereafter, calculating the log of that quotient. The idf is expressed the by the equation shown below:

$$idf(t,D) = \log\left(\frac{D}{d \in D : t \in d}\right) \quad (2)$$

Where, $D$ is the total number of documents, $\in D : t \in d$, it is the number of documents where term $t$ appears. The total term weight is given by $tf * idf$ which can be calculated by:

$$tf * idf(t,d,D) = tf(f,d) + idf(t,D) \quad (3)$$

$$f_3 = tf * idf \quad (4)$$

- ***Concept Feature***

The concept feature from the text document is derived by means of the mutual data and the windowing procedure, where a virtual window of size 'k' is shifted in the document from left to right. Now, the co-occurrence of words in same window is found out which can be estimated with the help of the following equation:

$$MI(w_i, w_j) = \log_2\left(\frac{p(w_i, w_j)}{p(w_i) * p(w_j)}\right) \quad (5)$$

Where, $p(w_i, w_j)$ -joint probability that both keyword appeared together in a text window.

$p(w_i)$ -probability that a keyword $w_i$ appears in a text window and can be computed by:

$$p(w_i) = \frac{|sw_t|}{|sw|} \quad (6)$$

Where,

$sw_t$ →the number of windows containing the keyword $w_i$

$|sw|$ →total number of windows constructed from a text document

The sentence matrix produced by means of the above-mentioned steps is given by:

$$\begin{array}{c} \\ S1 \\ S2 \\ . \\ . \\ Sn \end{array} \begin{pmatrix} T & P & Tw & C \\ f1 & f2 & f3 & f4 \\ .. & ... & .. & .. \\ .. & ... & .. & .. \\ ... & .. & .. & .. \end{pmatrix} \quad (7)$$

### 3. Sentence Matrix

Once features are extrated in feature extraction phase, a feature matrix is created.Here, sentence matrix $S_M = (s_1, s_2, ..., s_n)$ where $s_i = (f_1, f_2, f_{3,} f_4), i \le n$ is the feature vector.

### B. SUMMARIZATION VIA LMS BASED NEURAL NETWORK

After that feature matrix is formed, the Least Mean Square (LMS) based neural network is used for the summarization process. The Least Means square (LMS) algorithm was introduced by widrow and Hoff in 1959. LMS is an example of supervised learning algorithm in neural network (NN). In LMS, the algorithm trains the perceptron using the termination criterion until it correctly classifies the output of training set while the mean-square-error (MSE) is greater than a certain value. LMS is the faster algorithm that minimizes the MSE. The MSE is the average of the weighted sum of the error for $N$ training sample which defined as:

$$MSE = \frac{\sum_{j=1}^{N} (T - P_i)^2}{N} \quad (8)$$

Where, $T$ is the target, $P$ is the predicted result.

In order to train the perceptron by using LMS, we can iterate the test set, taking a set of inputs computing the output and then using the error to adjust the weight. The learning rule of LMS is given as:

$$w_{t+1} = w + \beta(T - P)E \quad (9)$$

The learning rule adjusts the weight based on the error. Once error is computed, the weight is adjusted for small amount, $\beta$ in the direction of the input, $E$.

The implementation of LMS is very simple. Initially, the weights vector is initialized with small random weights. The main repetition then randomly selects a test, calculates the output of the neuron, and then calculates the error. Using the error, the formula of learning rule is applied to each weight in the vector. Then continues the repetition to check the MSE to see if it has reached an acceptable value, and if so, exit and emit the computed truth table for the neuron.The simplest description of LMS training algorithm is explained in figure 3.

1. **Initialization. Set**

   $\hat{w}_k = 0$ for $k = 1,2,\ldots, p$

2. **Filtering. For time** $n = 1,2,\ldots,$

   Compute $y(n) = \sum_{j=1}^{p} \hat{w}_j(n) x_j(n)$

   $e(n) = d(n) - y(n)$

   $\hat{w}_k(n+1) = w_k(n) + \eta\, e(n)\, x_k(n)$ for $k = 1,2,\ldots, p$

**Figure 3:** Summarization of Least Means Square (LMS) algorithm

- **Summary generation**

The summary generation is done by two important stages, (1) Training, and (2) Testing. In training stage, the feature matrix is given to LMS based neural network structure. The proposed neural network will train the system based on target given. The target would be based on the documents taken for training belongs to the topic, as we make out the topic (domain) for the documents taken for training. The testing is made by giving the testing document after training the system based on the neural network. When the testing document is specified as input to the system, the frequency matrix is produced for the input document by means of the sentences that formed the frequency matrix in the training process. The testing document is given to the LMS based NN classifier, where the trained weight is used as hidden

layer weight for the testing stage. Finally, the LMS based NN will offer a score for the specified input document and based on the score the document will be summarized to which topic it belongs.

## 4. RESULT AND DISCUSSION:

We have offered the results of our suggested methodology and examined their presentation in this part. The proposed automatic text summarization based on multiple documents systemis implemented in the JAVA program and the text summarization process is experimented with the documents are collected from specific area like data mining, software engineering. We have implemented our proposed automatic text summarization system using Java (jdk 1.6) and a series of experiments were performed on a PC with Windows XP Operating system at 2 GHz dual core PC machine with 4 GB main memory running a 64-bit version of Windows 2007.

### 4.1 Dataset Description

The experimental evaluation of the proposed text summarization algorithm is executed on different documents. The documents are collected from specific area like data mining, networking and software engineering. Multiple documents from each of the different domains are collected and processed, since the proposed approach is based on multiple documents. The data mining keyword is given in the Google search and the top ten results are selected. The top ten results are stored as ten documents and given to the feature extraction phase to extract the feature vectors. Similarly, the document set for software engineering and networking are created and features are extracted.

### 4.2Evaluation metrics

The evaluation of proposed automatic text summarization systemis carried out using the following metrics as suggested by below equations,

**Precision (P):**Precision is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved which is given in equation (10).

$$P = \frac{\{relevant\ sentence\} \cap \{retrived\ sentences\}}{\{retrived\ sentences\}} \qquad (10)$$

**Recall (R):**Recall is the ratio of the number of relevant sentences retrieved to the total number of relevant records in the summary which is given in equation (11).

$$R = \frac{\{relevant\ sentence\} \cap \{retrived\ sentences\}}{\{relevant\ sentence\}} \qquad (11)$$

**F-measure(F):**F-measure is defined as the harmonic mean of precision and recalls metrics which is given in equation (12).

$$F = \frac{2PR}{(P+R)} \qquad (12)$$

Where;

P →precision, R →Recall, F →F-measure

### 4.3 Performance analysis:

In the following section, we plot the analysis and discussion of the experiments carried out on the proposed text summarization approach. The experiments are carried out with extreme precaution as most relevant sentences are selected for generating the summary. The summary is evaluated based on the above plotted parameter, precision, recall and f-measure. The figures 4 to 6 shows the performance of the proposed work and table 1 shows the features of proposed work utilization.

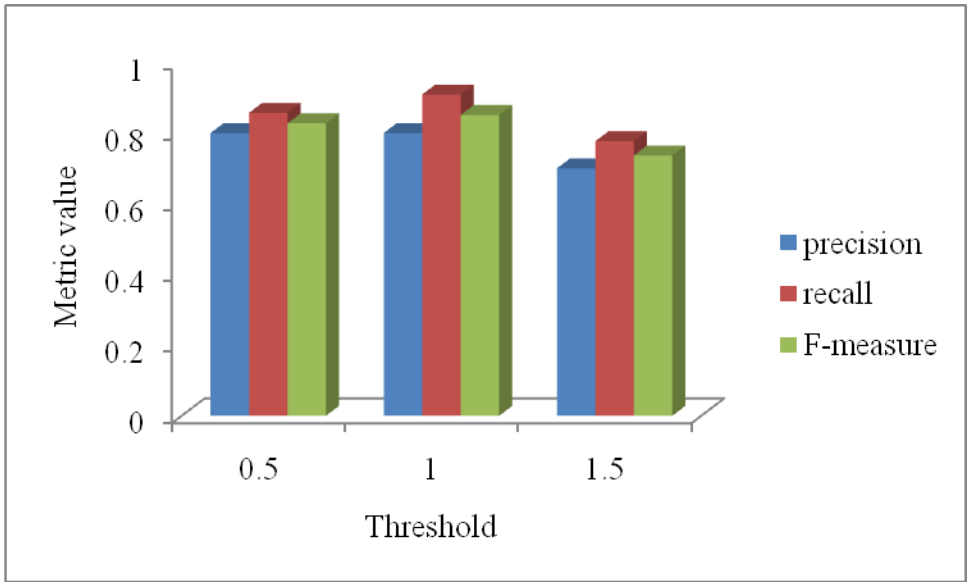


Figure 4: Performance of proposed approach using networking domain

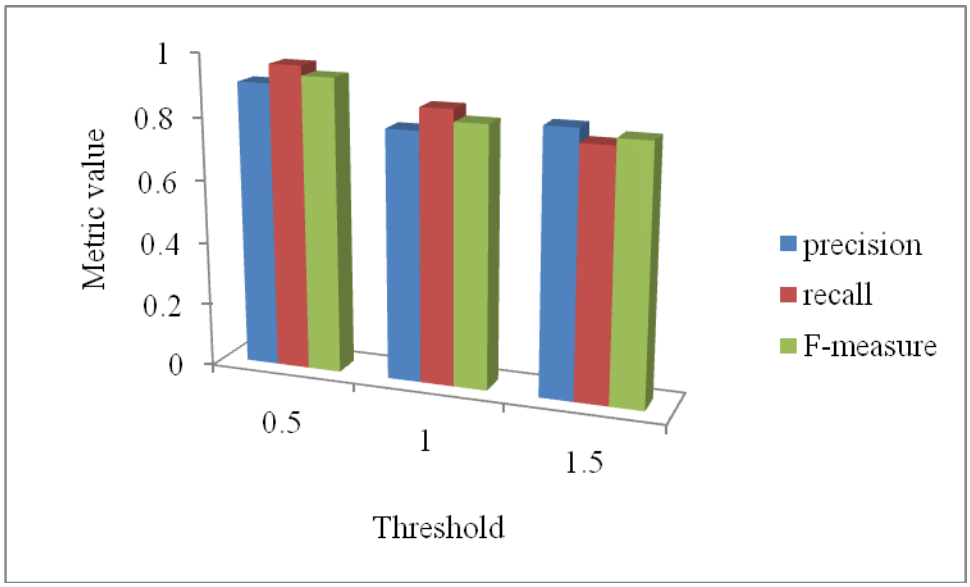


Figure 5: Performance of proposed approach using software engineering domain

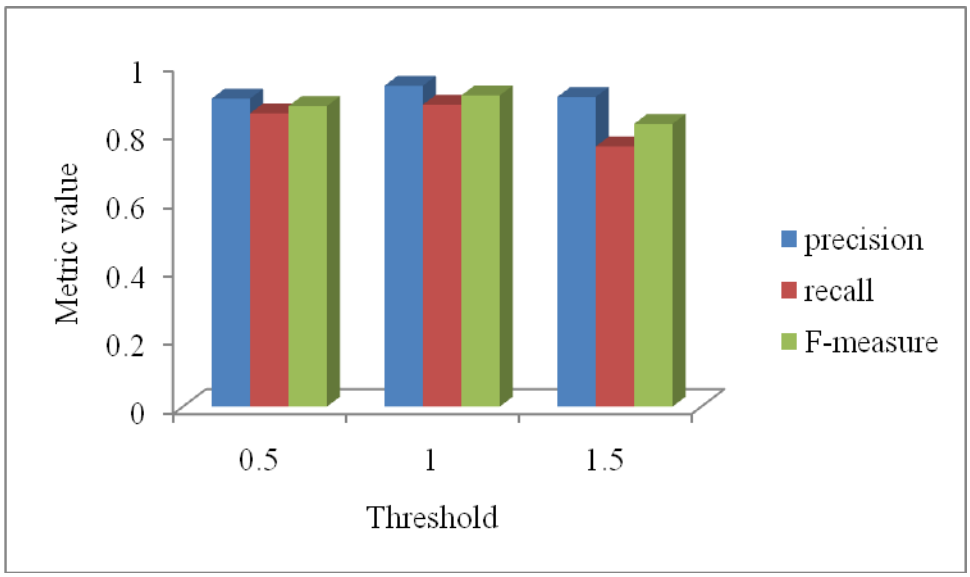


Figure 6: Performance of proposed approach using data mining domain

The basic idea of our research is to automatic text summarization using pragmatic-enabled features with modified mutual information. At first, the input documents are given to the preprocessing stage, here the sentence and words are extracted and indexed. Then, stop word removal and stemming process is applied to remove the unwanted and repeating words.An automatic text summarization is carried out using pragmatic-enabled features with modified mutual information. The final score values decides whether that sentence is summary sentence or not. The above figure 4 shows the performance of

proposed automatic text summarization using networking domain. When the threshold value is 0.5 and 1 we obtain the maximum precision value of 80% which is 70% for using threshold value is 2. When we using the threshold value are 1, we obtain the maximum recall of 90.9% and maximum F-measure of 85.1%. In figure 5, shows the performance of proposed approach using software domain. When analyzing figure 5, we obtain the maximum precision of 90.6%, recall of 96.6% and F-measure of 93.5% using the threshold value is 0.5. Similarly, in figure 6 shows the performance of automatic text summarization of proposed approach. Here, also we obtain the maximum precision of 93.75% using the threshold value is 1. When analyzing above three figures our proposed automatic text summarization achieves the very good performance.

| Document no | Parameter no | Line no | Title value | Pos value | Tf idf | Concept |
|---|---|---|---|---|---|---|
| 2 | 0 | 1 | 1 | 3 | 0.703669 | 0.236971 |
| 2 | 2 | 1 | 0.777778 | 2.8 | 0.683092 | 0.721901 |
| 2 | 2 | 2 | 0.777778 | 2.6 | 0.75243 | 0.635871 |
| 2 | 2 | 3 | 0 | 2.4 | 0.245638 | 0.311167 |
| 2 | 2 | 4 | 0 | 2.2 | 0.547057 | 0.509735 |
| 2 | 2 | 5 | 0.666667 | 2 | 0.831807 | 0.610243 |
| 2 | 2 | 6 | 0.111111 | 1.8 | 0.345836 | 0.242785 |
| 2 | 2 | 7 | 0.555556 | 1.6 | 0.590491 | 0.545455 |
| 2 | 2 | 8 | 0.111111 | 1.4 | 0.516636 | 0.605461 |
| 2 | 3 | 1 | 0.777778 | 1.2 | 0.804196 | 0.272783 |
| 2 | 5 | 1 | 0 | 1 | 0.004763 | 0 |
| 2 | 7 | 1 | 0.333333 | 0 | 0.308124 | 0.129836 |
| 2 | 8 | 1 | 0.333333 | 0 | 0.400183 | 0.500228 |
| 2 | 8 | 2 | 0.222222 | 0 | 0.295333 | 0.311654 |
| 2 | 8 | 3 | 0.555556 | 0 | 0.557147 | 0.484436 |
| 2 | 8 | 4 | 0.222222 | 0 | 0.449435 | 0.545455 |
| 2 | 8 | 5 | 0.1111111 | 0 | 0.239608 | 0.363636 |

Table 1: Features using in the proposed work

**5.4 Comparative analysis:**

In this section we compare our proposed automatic text summarization with existing approach. In existing we use only without pragmatic analysis approach to check whether that sentence is summary sentence or not.

| Name of the sub-domain in text corpus | Value of $\gamma$ during testing | Input query document name | Computation time (in seconds) | | Precision | | Recall | | F-measure | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Proposed (with pragmatic analysis) | existing (without pragmatic analysis) | proposed (with pragmatic analysis) | Existing (without pragmatic analysis) | proposed (with pragmatic analysis) | Existing (without pragmatic analysis) | proposed (with pragmatic analysis) | Existing (without pragmatic analysis) |
| data mining | 5 | Dm1.doc | 14658 | 15741 | 80 | 70 | 90 | 80 | 84.70 | 74.66 |
| | 5 | Dm2.doc | 15687 | 16927 | 80 | 80 | 85 | 70 | 82.42 | 74.66 |
| | 5 | Dm3.doc | 14699 | 15337 | 90 | 80 | 70 | 60 | 78.75 | 68.57 |
| | 5 | Dm4.doc | 16587 | 17884 | 90 | 80 | 80 | 70 | 84.70 | 74.6 |
| | 5 | Dm5.doc | 13698 | 14699 | 90 | 90 | 85 | 75 | 87.42 | 81.81 |
| Average | 5 | | 15065.8 | 16117.6 | 86 | 80 | 82 | 71 | 83.59 | 74.86 |

Table 2: Comparison between Proposed and existing approach using data mining domain

| Name of the sub-domain in text corpus | Value of $\gamma$ during testing | Input query document name | Computation time (in seconds) | | Precision | | Recall | | F-measure | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Proposed (with pragmatic analysis) | Existing (without pragmatic analysis) | Proposed (with pragmatic analysis) | Existing (without pragmatic analysis) | Proposed (with pragmatic analysis) | Existing (without pragmatic analysis) | Proposed (with pragmatic analysis) | Existing (without pragmatic analysis) |
| Networking | 5 | N1.doc | 13687 | 16985 | 100 | 80 | 90 | 80 | 94.73 | 80 |
| | 5 | N2.doc | 14771 | 15982 | 90 | 80 | 85 | 70 | 87.42 | 74.66 |
| | 5 | N3.doc | 13698 | 15412 | 80 | 70 | 70 | 60 | 74.66 | 64.61 |
| | 5 | N4.doc | 15981 | 16748 | 90 | 80 | 80 | 70 | 84.70 | 74.66 |
| | 5 | N5.doc | 17841 | 18741 | 90 | 70 | 85 | 75 | 87.42 | 75 |
| Average | 5 | | 15195.6 | 16773.6 | 90 | 76 | 82 | 71 | 85.78 | 73.78 |

Table 3: Comparison between Proposed and existing approach using Networking domain

| Name of the sub-domain in text corpus | Value of $\gamma$ during testing | Input query document name | Computation time (in seconds) | | Precision | | Recall | | F-measure | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Proposed (with pragmatic analysis) | Existing (without pragmatic analysis) | Proposed(with pragmatic analysis) | Existing (without pragmatic analysis) | Proposed (with pragmatic analysis) | Existing (without pragmatic analysis) | Proposed (with pragmatic analysis) | Existing (without pragmatic analysis) |
| Software engineering | 5 | Se1.doc | 17896 | 18748 | 80 | 70 | 90 | 70 | 84.70 | 70 |
| | 5 | Se2.doc | 17412 | 19734 | 80 | 80 | 90 | 70 | 84.7 | 74.7 |
| | 5 | Se3.doc | 16387 | 17698 | 90 | 80 | 80 | 60 | 84.7 | 68.57 |
| | 5 | Se4.doc | 13287 | 14687 | 90 | 80 | 80 | 70 | 84.7 | 74.7 |
| | 5 | Se5.doc | 15699 | 16982 | 90 | 90 | 90 | 80 | 90 | 84.7 |
| Average | 5 | | 16136.2 | 17569.8 | 86 | 80 | 86 | 70 | 85.76 | 74.53 |

Table 4: Comparison between Proposed and existing approach using software engineering domain

In table 2, shows the Comparison between Proposed and existing approach using data mining domain. In existing work we cannot use pragmatic analysis. When analyzing the table 2, we obtain the minimum average computation time of 15065.8 sec for using proposed work with pragmatic analysis and 16117.6 sec for using without pragmatic analysis. When comparing the precision, recall and f-measure value to the existing work, our proposed work achieves the maximum output values. In table 3 shows the comparison between proposed and existing approach using networking domain. In this table 3 clearly shows our proposed approach achieves the minimum computation time of 15195.6 sec and maximum precision, recall and f-measure values. Similarly, in table 4 shows the comparison between Proposed and existing approach using software engineering domain. Here also in all the metrics we obtain the maximum output. Overall, we clearly understand that our proposed approach achieves the maximum precision, recall and f-measure compare to existing approach.

## 5. **CONCLUSION**

Automatic text summarization aims to generate summaries for one or more texts using machine techniques. A variety of techniques have been developed in recent years. In this article, we proposed a text summarization approach using pragmatic-enabled features and LMS based neural network. Initially, the preprocessing steps were applied through pragmatic analysis. In this step, the text contents were filtered using WorldNet dictionary. After that, four set of features like as Title Similarity, Positional Feature, Term Weight and Concept Feature were extracted and feature matrix was generated. Once feature matrix was created, text summarization was done via LMS based neural network. The summary of the document is created based upon the score level using LMSNN. The performance of the proposed approach is evaluated through precision, recall and f-measure. Simulation results explicitly indicate that the proposed system offers a competitive performance with respect to the existing approach in terms of precision, recall and f-measure.

## References:

[1] Eduard Hovy, Chin-Yew, “Automated text summarization system and the summarist”, TIPSTER '98 Proceedings of a workshop on held at Baltimore, Maryland, PP. 197-214,1999.

[2] Luhn, Hans Peter, "The automatic creation of literature abstracts," IBM Journal of research and development, PP 159-165, 1958.

[3] Jen-Yuan Yeha, Hao-Ren Keb, Wei-Pang Yanga and Heng Menga, "Text summarization using a trainable summarizer and latent semantic analysis ", Information Processing & Management, Volume 41, Issue 1, Pages 75–95, January 2005.

[4] Jimmy Lin., “Summarization.”, Encyclopedia of Database Systems. Heidelberg, Germany: Springer-Verlag, 2009.

[5] Regina Barzilay, Kathleen R. McKeown, and Michael Elhadad, “Information fusion in the context of multi-document summarization”, In Proc. 37th ACL, 550–557, 1999.

[6] Ratnaparkhi, “A maximum entropy part-of-speech tagger”, Proceedings 1st Conference on Empirical Methods in Natural Language Processing, EMNLP, 1996

[7] Vishal Gupta, G.Sl Lehal, “A Survey of Text Mining Techniques and Applications”, Journal of Emerging Technologies in Web Intelligence, vol. 1, no. 1, 60-76, 2009

[8] Hien Nguyen and Eugene Santos, "Evaluation of the Impact of User-Cognitive Styles on the Assessment of Text Summarization", IEEE transactions on systems, man, and cybernetics,vol. 41, no. 6, 2011.

[9] Pawan Goyal, Laxmidhar Behera and Thomas Martin McGinnity, "A Context-Based Word Indexing Model for Document Summarization", IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 8, August 2013.

[10] Chien Chin Chen and Meng Chang Chen, "TSCAN: A Content Anatomy Approach to Temporal Topic Summarization", IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 1, January 2012.

[11] Chien-Liang Liu, Wen-Hoar Hsaio, Chia-Hoang Lee, Gen-Chi Lu, and Emery Jou, "Movie Rating and Review Summarization in Mobile Environment", IEEE Transactions On Systems, Man, And Cybernetics,Vol. 42, No. 3, May 2012.

[12] Feng Yang, "Study on core Technologies of Query-oriented Automatic Summarization", Journal on Advanced in Control Engineering and Information Science,PP 3600 – 3603,2011.

[13] Nowshath K. Batchaa, Normaziah A. Azizb and Sharil I. Shafiea, "CRF Based Feature Extraction Applied for Supervised AutomaticText Summarization", Journal on Electrical Engineering and Informatics,2013.

[14] Hitesh Kansal and Durga Toshniwal, "Aspect based summarization of context dependent opinion words", Journal on Knowledge-Based and Intelligent Information & Engineering Systems, pp 166 – 175, 2014

[15] Han Zhang , Marcelo Fiszman b, Dongwook Shin b, Christopher M. Miller b, Graciela Rosemblat b and Thomas C. Rindflesch, "Degree centrality for semantic abstraction summarization of therapeutic studies", Journal of Biomedical Informatics, pp 830–838,2011.

[16] Jiaming Zhan, Han Tong Loh, Ying Liu, Gather customer concerns from online product reviews – A text summarization approach, Expert Systems with Applications, Volume 36, Issue 2, Part 1, March 2009, Pages 2107-2115.

[17] Mohamed Abdel Fattah, Fuji Ren, GA, MR, FFNN, PNN and GMM based models for automatic text summarization, Computer Speech & Language, Volume 23, Issue 1, January 2009, Pages 126-144.

[18] X. Wan, "Towards a Unified Approach to Simultaneous Single-Document and Multi-Document Summarizations," Proc. 23rd Int'l Conf. Computational Linguistics, pp. 1137-1145, 2010.

[19] D.R. Radev, H. Jing, M. Sty_s, and D. Tam, "Centroid-Based Summarization of Multiple Documents," Information Processing and Management, vol. 40, pp. 919-938, 2004

[20] X. Wan and J. Xiao, "Exploiting Neighborhood Knowledge for Single Document Summarization and Key phrase Extraction," ACM Trans. Information Systems, vol. 28, pp. 8:1-8:34, June 2010.