Text Summarization using Artificial Neural Network with Back Propagation in Telugu e-Newspapers

D. Naga Sudha[1], Dr. Y. Madhavee Latha[2]
[1]*JNTUH College of Engineering Jagtial, Telangana, India*
[2]*Malla Reddy Engineering College for Women, India*
(E-mail: nayaniharshi@gmail.com)

***Abstract*—**Text summarization is an automatic system for generating a condensed version of the multiple original documents. Though, the manual text summarization requires a considerable number of qualified unbiased professional, higher budget and consumes more time. So, a new text summarization approach is developed by considering different contributions of training samples; Andhrabhoomi, Andhrajyothy, Vaartha, Eenadu and, Sakshi Telugu newspaper data. Usually, the raw Telugu language contains more noises by means of stop-words, low frequency words, etc., which were effectively eliminated using porter algorithm. After pre-processing the collected data, an effective machine learning based text summarization methodology: Artificial Neural Network with Back Propagation (ANN-BP) was used. The ANN-BP performed Telugu Part-Of-Speech Tagger (POS Tagger) for examining the accurate POS information of any given text based on a sentence or paragraph. In experimental analysis, the effectiveness of proposed approach was measured by means of precision, recall, accuracy, and f-measure. The experimental outcome showed that the proposed method improved the accuracy in text summarization up to 3-4% associated to the existing method; Learning Probability Distribution (LPD).

Keywords*—Artificial neural network with back propagation, Learning probability distribution, Porter algorithm, Text summarization.*

I. INTRODUCTION

Currently, the information contents such as, text, video and audio are easily generated by everyone, due to the development of internet and digital capturing system [1]. So, the search of the require data from the large stored data, increases the difficulty and takes more time. In recent times, automatic text summarization has been an emerging research topic and also showed great interest among the researchers [2-3]. The automatic text summarization used to develop a condensed version of the original document. Manual text summarization needs a considerable number of qualified unbiased professionals, higher budget and considerable time. Though, automatic text summarization is an active research topic in the field of natural language processing for overcoming the above mentioned concerns [4]. A lot of text summarization systems are developed for summarizing the documents in various languages such as, English, Hindi, Telugu, Malayalam, etc. Generally, the text summarization is done in two dissimilar ways; extractive and abstractive. The abstractive summarization simplifies the contents that enhances the coherence among the sentences by reducing the redundancies [5]. Additionally, the extractive summarization identifies more used-words and then score the sentences from dissimilar perspective [6].

Presently, there are several methodologies used for automatic text summarization of multiple documents such as, optimization methods, clustering methods, graph based methods, feature extraction methods, etc. [7-8]. A major issue in existing methods is the summarization time of multiple documents that makes the most of the systems unable to deal with the increase of data sizes [9-10]. To address this issue, an automatic effective text summarization methodology is developed. In this research paper, Telugu language is considered for text summarization, because it is the second most prevalent language in India after Hindi. In addition, the Telugu language ranks nineteenth in the Ethnologue list of most-spoken languages world-wide [11]. Usually, the raw Telugu language datasets consists of more noises by means of stop-words, low frequency words, etc., which were significantly reduced or pre-processed by using porter algorithm. It effectively eliminates affixes; prefixes, circumfixes, infixes, and suffixes from the Telugu articles in order to achieve a word stem. After pre-processing the Telugu data, an effective machine learning based text summarization method (ANN-BP) was used for generating a condensed version of the original multiple document in an effective manner. The ANN-BP has the ability to learn more complex and non-linear relationships that was important in real time applications like text summarization

This research paper is organized as follows. Several recent papers on text summarization are surveyed in the section II. In section III, an effective tag classification method: ANN-BP is presented for automatic keyword extraction. Section IV shows comparative experimental result for existing and proposed strategies. The conclusion is made in section V.

II. LITERATURE SURVEY

Numerous methodologies are developed by the researchers in text summarization. In this sub-section, a brief evaluation of a few essential contributions to the existing literatures are presented.

R. Abbasi-ghalehtaki, H. Khotanlou, and M. Esmaeilpour [12] presented an effective model based on the fuzzy logic system for automatic text summarization. At first, the important features like similarity measure, word features, length and position of a sentence were extracted. In this research paper, artificial bee colony with cellular learning automatic algorithm was employed to determine the similarity measure. The developed method identified less and more essential text features and then allocated fair weights to the features. While implementing optimization approach, the summarization of words was more complex. The developed system requires more repeated words to summarize a particular word.

V. Gupta, and N. Kaur [13] developed a hybrid methodology: conceptual, static, location and linguistic based features for Punjabi text automatic summarization. In this hybrid system, two new statistical entropy measures; Z-score and four new location based features were utilized. The experimental outcome of developed method was compared with different baseline systems by means of precision, rouge-2 score, recall and f-score. The major drawback of hybrid methodology was language dependent and it was only applicable for specific language not for all the languages.

R. Naidu, S.K. Bharti, K. Sathya Babu, and R. K. Mohapatra [14] developed a new learning probability distribution system for extracting the Telugu keywords. In this literature paper, Telugu e-newspaper datasets were utilized for text summarization. The outcomes of qualitative and quantitative evaluation showed that the developed system outperformed the existing approaches in document summarization. The experimental result showed that the developed system effectively enhanced the performance of summarization. In text summarization, the Telugu language has more stop words and low frequency words that reduces the performance of summarization.

R. Kabeer and S.M. Idicula, [15] presented a semantic graph based approach and a statistical sentence scoring method for text summarization. The proposed strategy was evaluated and the results were compared with other well-known Malayalam document summarization algorithms. The experimental result showed that the developed strategy out-performed the existing approaches in Malayalam document summarization. In this research paper, the computational time was a bit high, while summarizing the large dataset contents.

C. Fang, D. Mu, Z. Deng, and Z. Wu [16] illustrated a sentence scoring method for extractive text summarization. In this literature, a word sentence co-ranking approach; CoRank was combined with graph based ranking approach for improving the sentence and word relationship. In experimental analysis, the CoRank methodology serves as a significant building-block of the intelligent summarization systems. Whereas, the CoRank features have limited sets, so it was not possible to apply directly on different languages.

To overcome the above mentioned drawbacks and also to enhance the performance of text summarization, an effective algorithm (ANN-BP) is implemented in this research study.

III. PROPOSED SYSTEM

In this research work, the proposed approach ANN-BP is used for automatic keyword extraction. Here, the Telugu POS tagger [17] determines the suitable POS information in Telugu text and it comprises of totally twenty-one tags. In this scenario, the unwanted stop-words and low frequency words are eliminated before summarizing the news. Then, extract the keywords from pre-processed data using automatic key-word extraction method. The extracted keywords are used for summarizing the article. Finally, the summarization approach chooses the sentence based on the necessitated summary. The proposed architecture consists of five steps: data collection, pre-processing, keyword annotation, tag classification and keyword extraction. The proposed architecture is given in the Fig. 1.

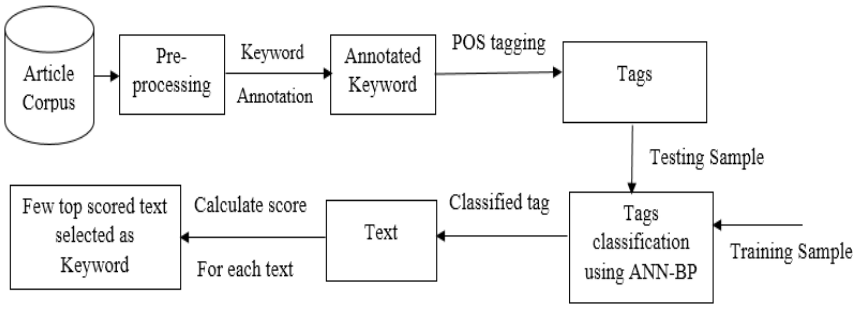

Figure.1 Architecture of proposed system

A. *Collection of Telugu data*

In this experimental analysis, the input data were collected from Telugu e-newspapers. The data were collected from various newspapers such as, Sakshi, Vaartha, Eenadu, Andhrajyothy and Andhrabhoomi. This database consists of 450 articles and every article ranging from 1st October 2016 to 6th December 2016. From Eenadu newspaper, 150 articles were collected with 12233 keywords, and 140 articles were collected from Sakshi

newspaper with 1148 keywords. Similarly, 100 articles were collected from Andhrajyothy with 915 keywords, 70 articles were collected from Vaartha with 785 keywords, and 50 articles were collected from Andhrabhoomi with 568 keywords. The collected data given as the input for the pre-processing procedure.

B. Pre-processing of collected data

After data collection, an important step in text summarization is pre-processing of collected data. The raw Telugu dataset consists of more noises by means of stop-words, low frequency words (words that rarely occur in a language) etc., which are eliminated using stemming process. In this research study, porter algorithm is used for stemming. The developed algorithm eliminates affixes (prefixes, circumfixes, infixes, and suffixes) from an article to achieve a word stem. For instance, the words observer, observe, observes and observations are stemmed to a word "observe". This process exactly matches the stems, which effectively saves memory, space and time.

For instance;

English: My name is kumar

Telugu: నా పేరు ఉంది కుమార్

After removing stop words;

English: name kumar

Telugu: పేరు కుమార్

C. Keyword annotation

After pre-processing the data, human intervention is performed for key annotation in order to train the proposed approach. Though, the human annotators investigate the document and selects the appropriate keywords. In this research, totally four human annotators done this job. The selected keywords used in POS tagger and the extracted POS information are given as the input to the next section of the model.

D. Tag classification using ANN-BP

After keyword annotation, an appropriate tag classification method; ANN-BP is used to improve the accuracy level of the system. ANN is a well-known technique in automated text extraction, especially, it is back-propagated with feed-forward chaining. This effective learning methodology is utilized to train and recognize the patterns. Respectively, BP identifies the error in previous layer output by analyzing the current layer output as a response. This process is consequently iterative and compute the modifications in the weight of the last layer. The error is determined at the output of prior layer and repeat the process.

Again initialize this process by determining the partial derivative of the error, due to a single input data pattern and output of the neurons in the last layer. The error occurred in the single input data pattern is computed as shown in the Eq. (1),

$$E_n^p = \frac{1}{2}\sum\left(x_n^i - T_n^i\right)^2 \tag{1}$$

Where, E_n^p is represented as error, P is denoted as single pattern, T_n^i is stated as target output at the last layer and x_n^i is represented as the actual value of the last output layer. The partial derivation result of the Eq. (1) is represented in the Eq. (2).

$$\frac{\partial E_n^p}{\partial x_n^i} = x_n^i - T_n^i \tag{2}$$

The outcome of Eq. (2) is given as the starting value of BP procedure. These numeric values are utilized as the quantities of Eq. (3) for determining the derivative values. Using the derivative values, the weight of the numeric value calculated that is mathematically denoted in the Eq. (4).

$$\frac{\partial E_n^p}{\partial y_n^i} = G\left(x_n^i\right)\frac{\partial E_n^p}{\partial y_n^i} \tag{3}$$

Where, $G\left(x_n^i\right)$ is represented as the derivative of the activation function.

$$\frac{\partial E_n^p}{\partial w_n^{ij}} = x_{n-1}^i \frac{\partial E_n^p}{\partial y_n^i} \tag{4}$$

Subsequently, use Eq. (3) and (4) for computing the error of the previous layer, with the help of Eq. (5).

$$\frac{\partial E_{n-1}^p}{\partial y_{n-1}^k} = \sum w_n^{ik} \frac{\partial E_n^p}{\partial y_n^i} \tag{5}$$

The resulting values from Eq. (5) is utilized as the starting values for preceding layers, which is the most significant point in understanding BP. The ANN-BP algorithm helps to perform Telugu POS tagger for analyzing the accurate POS information of any given text on the basis of contest like phase, sentence, or paragraph.

E. Keyword extraction

The output of Telugu POS tagger is forwarded to the model for keyword extraction. Initially, the score is determined for each text from the Telugu POS tagger, in that the top scored texts are selected as keywords. Then, calculate the number of words in the document and rank the keywords by using the Eq. (6).

$$Score = P(\text{tag}) \times Count\left(word; tag\right) \tag{6}$$

Where, $P(\text{tag})$ is probability of a tag that is calculated by dividing the tag count by the total number of keywords and and $Count\left(word; tag\right)$ is defined as the number of words in the document. The algorithm for keyword extraction is determined below.

1) Algorithm for keyword extraction

Data: Doc: = Input article

P(Tag):= List of trained probabilities

Num_key-words:=Required number of keywords

Result: key-words []

Train.Pos_ Doc to ANN // Trained from documents, collected from human spectators

Test. Pos_ Doc to ANN

Top:=0

While word in Pos_ Doc do

flag:=0

for i <- 0 to top do

if word.text=wordset[i].text and word.tag=wordset[i].tag then

wordset[i].count:=wordset[i].count+1 flag:=1

end

end

if flag=0 then

wordset[top+1].word:=word.word wordset[top+1].tag:=word.tag

wordset[top+1].count:=1 wordset[top+1].score:=0 top:=top+1

end

end

for i <- 0 to size do

wordset[i].score:=wordset[i].count*P(wordset[i].tag)

end

sort_desc(wordset.score)

for i <- 0 to Num_ key-words do

key-words [i]: = wordset[i]

End

F. Text summarization

After ranking the key-words, text summarization is carried-out to summarize the acquired key-words. For text summarization, the proposed system suggested that the keyword of the article is directly proportional to the score it received. Here, the number of times a word is being repetitive in the article is termed as the keyword's score value. Then, the proposed system derives the sentences by means of crude scoring or clustering means. The working procedure of the proposed system is determined below.

Let us assume a news article about Yuvraj Singh retirement with 17 sentences, the possible key-words are listed in the table 1. Initially, assume the score for each key-words and then extract the sentences for each key-words using the Eq. (7). Finally, the extracted sentences are utilized for summarizing the document.

$$NS = \left[\frac{Keyword\ score \times Number\ of\ sentences\ required}{Toatl\ score\ of\ all\ keywords} \right] \quad (7)$$

Where, *NS* is represented as the number of sentences needed in summary using every keyword.

TABLE I. KEYWORD SCORE AND NUMBER OF SENTENCES ON EVERY KEYWORD

Keyword	యువరాజ్ సింగ్	ఐపియల్ మంచి ప్రదర్శన	ప్రపంచ కప్	విరమణ
Keyword score	2	1.5	2.5	2.5
NS	4	3	5	5

IV. EXPERIMENTAL RESULT AND DISCUSSION

In this section, the proposed system was experimented using Java (NetBeans) with 8 GB RAM, 3.0 GHZ Intel i5 processor, and 1TB hard disc. The performance of proposed system was compared with the existing method (LPD) on the reputed datasets like Eenadu, Sakshi, Andhrajyothy, Vaartha, and Andhrabhoomi Telugu e-newspapers for estimating the efficiency and effectiveness of proposed approach. The proposed system performance was compared by means of accuracy, recall, precision and f-measure.

A. Performance measure

Performance measure is defined as the regular measurement of outcomes and results that develops a reliable information about the effectiveness of proposed system. Also, it is the procedure of reporting, collecting and analyzing information regarding the performance of a group or individual. The general formula to evaluate precision and recall of text summarization are represented in the Eq. (8) and (9).

$$precision = \frac{TP}{TP + FP} \quad (8)$$

$$\text{Re}call = \frac{TP}{TP + FN} \quad (9)$$

In addition, accuracy and f-measure are the appropriate evaluation metrics used to find the effectiveness and efficiency of text summarization. The formula of accuracy and f-measure are represented in the Eq. (10) and (11).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad (10)$$

$$F - measure = \frac{2TP}{2TP + FP + FN} \times 100 \quad (11)$$

Where, FP is represented as false positive, TP is denoted as true positive, FN is indicated as false negative, and TN is specified as true negative.

B. Quantitative analysis

In this experimental analysis, Eenadu, Sakshi, Andhrajyothy, Vaartha, Andhrabhoomi e-newspaper databases were used to relate the performance of existing and proposed method. Here, the performance evaluation is validated with 70% training and 30% testing of data. In table 2, performance evaluation of proposed and existing

methods demonstrated by means of precision and recall. In table 2, the average precision of proposed system (ANN-BP) is 0.97 and the existing methodology (LPD) achieves 0.8298. Similarly, the average recall of proposed system is 0.9 and the existing methodology delivers 0.789 of average recall. The graphical comparison of recall and precision are denoted in the Fig. 2 and 3.

TABLE II. PERFORMANCE ANALYSIS OF PROPOSED AND EXISTING METHOD BY MEANS OF PRECISION AND RECALL

Methods	News paper	Precision	Recall
LPD [14]	Eenadu	0.795	0.761
	Sakshi	0.800	0.753
	Andhrajyothy	0.839	0.806
	Vaartha	0.865	0.816
	Andhrabhoomi	0.85	0.809
ANN-BP [proposed]	Eenadu	0.95	0.93
	Sakshi	1	0.92
	Andhrajyothy	0.9	0.85
	Vaartha	1	0.9
	Andhrabhoomi	1	0.9

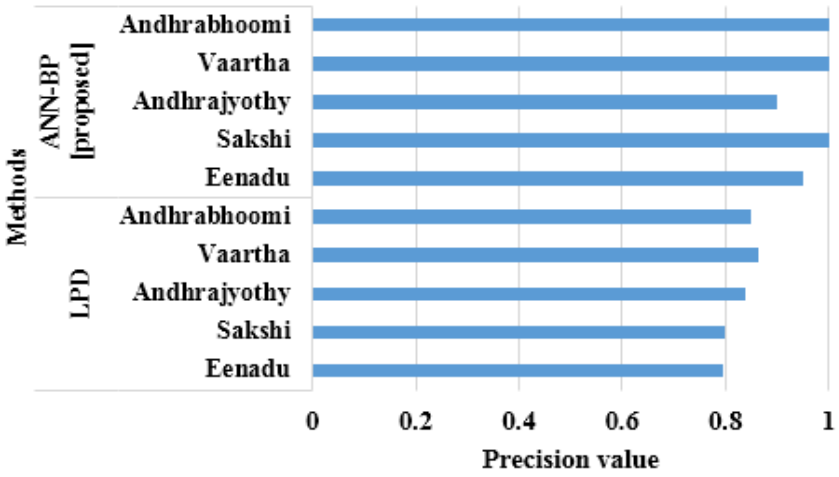

Figure.2 Precision comparison of proposed and existing method

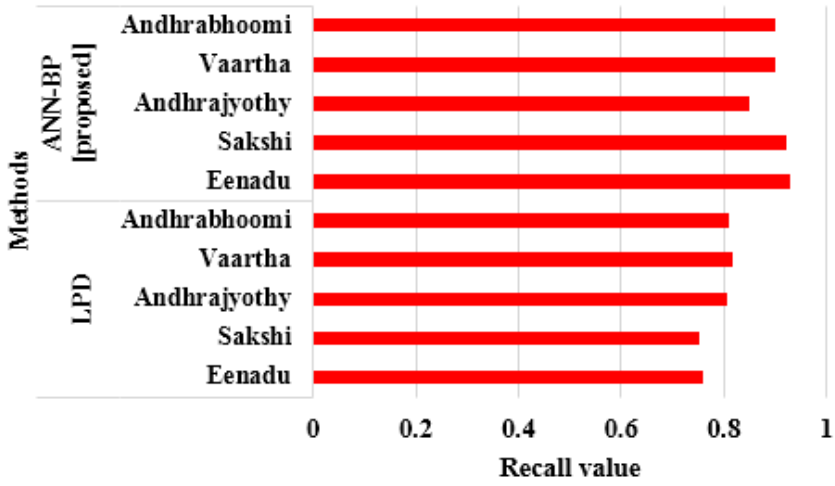

Figure.3 Recall comparison of proposed and existing method

In Table 3, performance evaluation of proposed and existing methods is demonstrated by means of accuracy and f-measure. In Table 3, average accuracy of proposed system (ANN-BP) is 94.62% and the existing methodology (LPD) delivers 90.82% of average accuracy. Similarly, the average f-measure of proposed system (ANN-BP) is 93.41% and the existing methodology (LPD) achieves 80.88% of average f-measure. The graphical comparison of accuracy and f-measure are represented in the Fig. 4 and 5.

Tables 2 and 3 illustrates the performance of existing and proposed system. The experimental result confirmed that the proposed system performs effectively in text summarization related to the previous methods by means of accuracy, f-measure, precision and recall. Here, the proposed system (ANN-BP) has the ability to capture the

non-linear characteristics of the Telugu data and also preserves the quantitative relationships between the trained patterns and recognized patterns.

TABLE III. PERFORMANCE EVALUATION OF PROPOSED AND EXISTING METHOD IN TERMS OF F-MEASURE AND ACCURACY

Methods	News paper	Accuracy (%)	F-measure (%)
LPD [13]	Eenadu	90.70	77.70
	Sakshi	91.70	77.60
	Andhrajyothy	91.30	82.20
	Vaartha	91.50	84
	Andhrabhoomi	88.90	82.90
ANN-BP [proposed]	Eenadu	93	94
	Sakshi	95	96
	Andhrajyothy	94.10	87
	Vaartha	98	95
	Andhrabhoomi	93	95.05

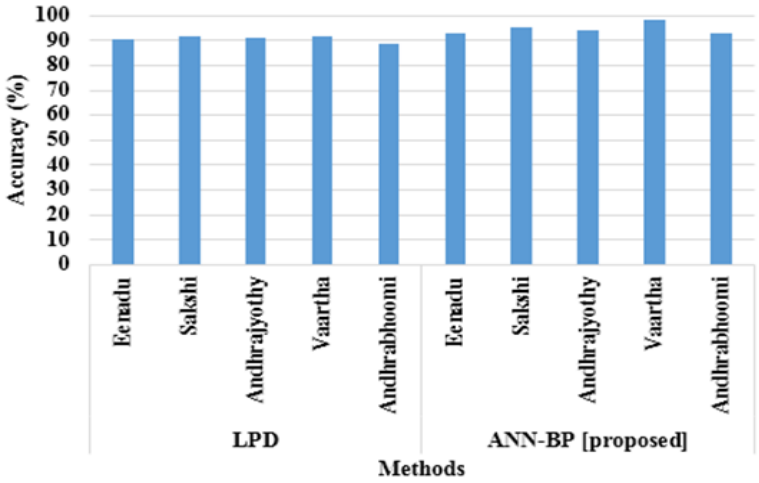

Figure.4 Accuracy comparison of proposed and existing method

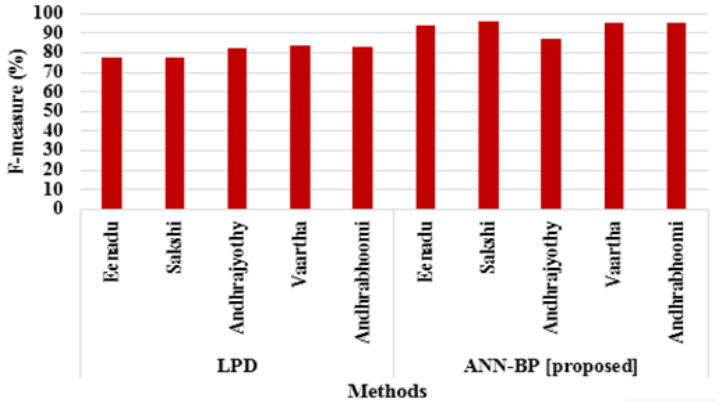

Figure.5 F-measure comparison of proposed and existing method

V. CONCLUSION

Text summarization is the most significant research task in the field of natural language processing system. The objective of the experiment is to develop a proper classification approach for summarizing the text in the multiple documents using Sakshi, Vaartha, Eenadu, Andhrajyothy and Andhrabhoomi Telugu newspapers. In this scenario, porter algorithm was utilized to eliminate the noise in the raw Telugu data. Then, an effective machine learning based text summarization methodology; ANN-BP was employed to inspect the accurate POS information of given Telugu text. Compared to other existing methods in text summarization, the proposed system delivered an effective performance by means of accuracy,

precision, recall and f-measure that shows 3-4% of improvement in text summarization. In future work, an effective automatic POS tagger was combined with an appropriate multi objective classifier for further improving of the summarization rate.

REFERENCES

[1] S. Akter, A.S. Asa, M.P. Uddin, M.D. Hossain, S. K. Roy, and M. I. Afjal, "An extractive text summarization technique for Bengali document (s) using K-means clustering algorithm," In: Imaging, Vision & Pattern Recognition (icIVPR), pp. 1-6, 2017.

[2] 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, 2017.

[3] S.A. Babar, and P.D. Patil, "Improving performance of text summarization," Procedia Computer Science, vol. 46, pp.v354-363, 2015.

[4] M. Yousefi-Azar, and L. Hamey, "Text summarization using effective deep learning," Expert Systems with Applications, Vol.68, pp.93-105, 2017.

[5] T.B. Mirani, and S. Sasi, "Two-level text summarization from online news sources with sentiment analysis", in: Networks & Advances in Computational Technologies (NetACT), pp. 19-24, 2017.

[6] S. Abujar, M. Hasan, M. S. I. Shahin, and S. A. Hossain, "A Heuristic Approach of Text Summarization for Bengali Documentation," In: 8th International Conference on Computing, Communication and Networking (8th ICCCNT), pp.1-8 2017.

[7] A. Kumar, A. Sharma, S. Sharma, and S. Kashyap, "Performance analysis of keyword extraction algorithms assessing extractive text summarization," In: Computer, Communications and Electronics (Comptelix), pp. 408-414, 2017.

[8] R.S. Baraka, and S.N. Al Breem, "Automatic Arabic Text Summarization for Large Scale Multiple Documents Using Genetic Algorithm and MapReduce," In: Information and Communication Technology (PICICT), pp.40-45, 2017.

[9] R.Z. Al-Abdallah, and A.T. Al-Taani, "Arabic Single-Document Text Summarization Using Particle Swarm Optimization Algorithm," Procedia Computer Science, vol. 117, pp. 30-37, 2017.

[10] S. Swamy, T. Shalini, S.P. Nagabhushan, S. Nawaz, and K.V. Ramakrishnan, "Text Dependent Speaker Identification and Speech Recognition Using Artificial Neural Network," In: Global Trends in Computing and Communication Systems, pp. 160-168, 2012.

[11] Summary by language size, Ethnologue. Retrieved, 2016.

[12] R. Abbasi-ghalehtaki, H. Khotanlou, and M. Esmaeilpour, "Fuzzy evolutionary cellular learning automata model for text summarization," Swarm and Evolutionary Computation. vol. 30, pp.11-26, 2016.

[13] V. Gupta, and N. Kaur, "A novel hybrid text summarization system for Punjabi text," Cognitive Computation, vol. 8, pp.261-277, 2016.

[14] R. Naidu, S.K. Bharti, K. Sathya Babu, and R. K. Mohapatra, "Text Summarization with Automatic Keyword Extraction in Telugu e-Newspapers," In: Smart Computing and Informatics. pp. 555-564, 2017.

[15] R. Kabeer, and S.M. Idicula, "Text summarization for Malayalam documents-An experience," In: Data Science & Engineering (ICDSE), pp. 145-150, 2014.

[16] 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.

[17] S. Reddy, and S. Sharoff, "Cross language POS taggers (and other tools) for Indian languages: An experiment with Kannada using Telugu resources," In Proceedings of the Fifth International Workshop on Cross Lingual Information Access, pp. 11-19, 2011.

D. Naga sudha is working as Assitant Professor in Electronics and Communication Engineering department JNTUHCEJ.Her area of specialization is in the field of Digital Systems and Computer Electronics.She has been teaching for more than 15 years and her current research work includes image processing, VLSI design and information retrieval. She also has industry experience of over 3 years at various software organizations.

Y. Madhaveelatha is working as Professor and Principal of Malla Reddy College of Engineering for women. She completed her B.Tech, M.Tech & Ph.D in the area of Electronics and Communication Engineering from JNTU, Hyderabad. Her area of specialization is in the field of Signal Processing and Communication. She has more than 18 years of experience and she took up academic pursuit and served in an array of designations like Professor & Head of Department, ECE. She has attained Patent Right for her Technological Innovation "An Arrangement for Cultivation in a Chamber Using Fuzzy Logic Wind Intelligence Technology". She has delivered more than 50 keynote addresses and expert lectures in Conferences and Workshops. Has established IEEE, IETE, ISTE & HMA Student Professional chapters and has convened more than 300 conferences, workshops & refresher courses for the benefit of Students and Staff on a variety of themes. She is a member of Editorial Boards and Advisory Committees for several Conferences and Journals and has reviewed several Journals in her relevant field.