Journal of Theoretical and Applied Information Technology
15th September 2022. Vol.100. No 17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

TELUGU TEXT SUMMARIZATION USING HISTO FUZZY C-MEANS AND MEDIAN SUPPORT BASED GRASSHOPPER OPTIMIZATION ALGORITHM (MSGOA)

CHINNI BALA VIJAYA DURGA[1], DR. G.RAMA MOHAN BABU[2]

[1]Research Scholar, Dr. YSR ANU College of Engineering & Technology, Department of CSE,

Acharya Nagarjuna University, Nagarjuna nagar-522 510, Guntur (DT.), A.P., India.

[2]Professor, , RVR & JC College of Engineering, Department of CSE-AIML

Chowdavaram, Guntur (DT.), A.P., India.

E-mail: [1]balavijayadurga@gmail.com, [2]rmbgatram@gmail.com

ABSTRACT

In this work, we clearly stated the text summarization. Time is taken to read the big document it's quite complex, at the same time if we summarize the same document it is easy to understand and time-saving. Here we propose the text summarization of the Telugu language. To begin, documents must go through a preprocessing procedure that includes tokenization, stop-word removal, stemming and N-gram analysis. After that, it uses Histo-Fuzzy C-means Clustering to achieve clustering, as well as a technique of sentence ranking based on weights. Finally, the Median Support Based Grasshopper Optimization Algorithm (MSGOA) is utilized to combine the phrases into a clear and succinct summary. The performance of this strategy is evaluated using an online research dataset. When compared to earlier text summarizing methods, the suggested method outperforms them. When compared to existing accuracy, the proposed method performs admirably and obtains an accuracy of 84%.

Keywords: *Telugu language Text Summarization, preprocessing, Median Support Based Grasshopper Optimization (MSGO), Histo-Fuzzy C-Means Clustering, Stemming, Enthalpy.*

1. INTRODUCTION

Now a days every data is available in the internet in reginal languages itself which is very good thing. A common men can read everything easily and can understand what is the thing is conveying in context. There is a lot of difference between understanding in reginal language and in other language. Sometimes it may leads to very big problem when conveying a message because of understanding the other languages. So our government approved a bill NEP2020 (National Education Policy 2020) is a new education system with the main moto of study of a child is in his reginal language which helps to quality of education. With this context the data is available in reginal languages in internet. Telugu language is also one of most popular reginal language which is available in internet. Now the problem is data is available in reginal languages but which is very huge data sometimes may not have that much time to read and understand not only that sometimes the data is not relevant to us that thing can know after seeing the whole data that causes to vesting of the valuable time. To avoid this providing summarization of a text data in reginal languages is useful.

The Internet now allows users to access massive amounts of data. With the rise in popularity, extracting the most significant information from such a big number of data is becoming increasingly difficult. In the case of text documents, gathering and comprehending primary information from a vast quantity of resources insufficient time for human beings is a complicated and time-consuming procedure [1]. For decades, information retrieval algorithms have been performing these tasks automatically. As the amount of data rises, however, various performance difficulties emerge, including insufficient solutions and unmanageable information retrieval task applications. The use of high-tech machines may be able to help alleviate the losses caused by these issues. However, it is possible that the price will be higher. Dimension reduction in raw data can be used to address these challenges and speed up the completion of these jobs as a more appropriate option. Summarization is one of the most important applications in the computer world. Text summarizing is the process of producing a reduced version of a single document or a group of documents. Automatically summarizing text documents is a tough problem to solve since the resulting summaries must encompass as much of the original document(s) as possible [4]. Text summaries save us time and effort by giving us a fast summary of the entire content. Originally, these text summaries were written by hand, but as the amount of data has risen and automation has become more common, automatic summary approaches have become increasingly significant [5].

On the basis of their worth in extractive summarizing, some of the sentences are immediately selected and included in the summary. Abstractive summarization, on the other hand, adjusts or paraphrases terms so that they have the same meaning as the original. On numerous Telugu documents, extractive summarization is conducted. TS is a modern solution to the categorization problem in machine learning. This extractive ts method seeks to know the importance of individual sentences in a document and chose the most important ones. Why because machine not create the sentence and forming the sentences. And why it is not involving this is grammar is differ to every language. Extractive ts has the stages such as sentence rank based on score and sentence selection. Based on the rank the n top most sentences will select to make summary. These process unavoidably link to classification challenges in the ML sector. Determining rank is ML problem that is best handled by supervised learning [10] [11].

The government and the general public have turned to Twitter and Sina microblogs to get information on a number of problems. However, with the rapid expansion of social media platforms, consumers are finding it more difficult to get the information they need right away. For online search users, large amounts of data are likely to cause issues. Query-based summarization is another well-known method for automatic text summarization. This method uses an input query to compute the summary of the text content. This method focuses on the user query parameters and, as a result, on the user queries that are given to the system by users [19]. Text summaries are used to shorten input documents while keeping their overall meaning and information value. As a result, text summarizing is the act of reducing data in order for the user to absorb it more rapidly [20]. This paper's primary contribution is:

- The pre-processing step using the methods of Tokenization, stop word elimination, stemming, and N-gram.
- Evaluation includes the Histo Fuzzy C-means Clustering and Sentence Ranking.
- Sentences selection using MSGOA ranking.
- Finally, the phrases are put together to make an informative and short summary.

The farther paper is as follows: The technique of Telugu language text summary is well detailed in sec 3. Sec 2 includes a review of similar studies, while sec 3 presents the proposed approach of Telugu language text summarization. In sec 4, the results are posted, and in sec 5 concluded the work.

2. LITERATURE SURVEY

Reddy Naidu et al [21] proposed a method for automatically extracting keywords for text summarizing from Telugu e-newspaper datasets. The technique of compressing a text material into a summary that preserves the important ideas is known as summarization. Extractive summarizers use the information they've been given and extract sentences that best reflect the hidden message. The bulk of extractive summarization techniques are founded on the concept of detecting keywords, and extraction is typically achieved by extracting relevant words that occur more frequently than others, with an emphasis on the most important. Manual keyword extraction and annotation is a time-consuming and error-prone process.

Yan Du et al. [22] proposed a new automatic news text summarizing model fuzzy logic, multi-features, and a genetic algorithm. Word features are the most required features. The keywords were selected from the extracted words of higher score than predefined. They apply a fuzzy logic framework to calculate the final score. Because news text is a distinct sort of text with various distinct aspects of place and time these distinct news elements can sometimes be retrieved as keywords. Each feature was weighted using Genetic algorithm. A linear combination reveals the significance of each text.

Angel Hernandez-Castaneda et al [23] suggested a technique to improve keyword recognition used a semantic information for the ATS. The task of automated text summarizing

Journal of Theoretical and Applied Information Technology
15th September 2022. Vol.100. No 17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

(ATS) requires synthesizing a document to create a condensed version of it. Making a summary demands choosing not only on the primary ideas of the sentences, but also on their key relationships. Related works use a ranking system to select which text units (mainly sentences) should be included in the summary. However, because important information may have been omitted, the resulting summaries may not cover all of the topics discussed in the source text. By grouping phrases to locate the primary subjects in the original manuscript, this technique improves coverage and precision.

Rana Alqaisi et al [24] proposed an automatic, generic, and extractive method for summarizing Arabic documents. Because of the rising usage of the Internet and social media, a great volume of textual material is now available online. These online textual data resulted in overabundance and redundancy. When reading online textual data, it's vital to reduce information redundancy and save time. As a result, there is a constant demand for an automatic text summarizing system that pulls the relevant and conspicuous information from a group of texts that share the same or related themes. In the proposed system, clustering-based and evolutionary multi-objective optimization methodologies are applied. The clustering-based method identifies the text's most important subjects, whereas the evolutionary multi-objective optimization method prioritizes three objectives: coverage, diversity/redundancy, and relevancy.

Fucheng You et al [25] Using Fine-tuning BERT introduced a topic information fusion and semantic relevance for text summarization using Fine-tuning BERT (TIF-SR). The focus of a high-quality summarizing system should be on the document's topic substance and the resemblance between the summary and the source document. They extract topic keywords and integrate them with source documents as part of the input because subject information is so vital in summary synthesis [27-32]. Second, calculate the semantic similarity between the produced summary and the original material to improve the quality of the abstract.

Gaps Identified

S.No.	Author	Methodology	Merits	Future work of / Demerits
1	Reddy Naidu	Extractive - Automatic Keyword Extraction	The similar title in five different Telugu e-Newspapers to check the similarity and consistency in summarized results.	key-word extraction technic
2	Yan Du	fuzzy logic rules, multi-feature and Genetic algorithm	used fuzzy logic to calculate the final score	Sometimes leads to over fit
3	Angel Hernandez-Castaneda	Lexical-Semantic Keywords	The most relevant information for generating good-quality summaries	system is language and domain independent
4	Rana Alqaisi	Used evolutionary multi-objective optimization with K-medoid clustering	Optimized redundancy, and relevancy	Can use Genetic Algorithm to find the optimal weights, not encounter coherency or readability of the generated summary
5	Fucheng You	Fusion and Semantic Relevance based on Fine-tuning BERT(TIF-SR)	abstractive summarization	system is language independent

3. PROPOSED METHODOLOGY

It's a time-consuming and difficult process to read massive, lengthy documents. A summary of the same document provides an overview of the content. For inputted documents, the summary can

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

be generated. This work proposes a Telugu text summarization using a new hybrid method that combines Histo-enthalpy Fuzzy C-means Clustering and Median Support Based Grasshopper Optimization Algorithm (MSGOA). Initially, input documents are given to the pre-processing stage for the text summarization purpose. After combining all of the papers, the output document is processed. Tokenization, stop word elimination, stemming, and N-gram are some of the features available. Sentences are ranked by assigning weights and are ranked depending on their weights in Histo Fuzzy C-means Clustering and Sentence Ranking. Highly ranked sentences are taken from the input document, resulting in a high-quality summary of the document. The MSGOA ranking is used to choose the sentences that will be featured in the summary. The sentences with the highest ranking are chosen. The sentences that aren't important are removed here[33-40]. Finally, the phrases are put together to make an informative and short summary. The block diagram of the presented methodology is given in figure 1.

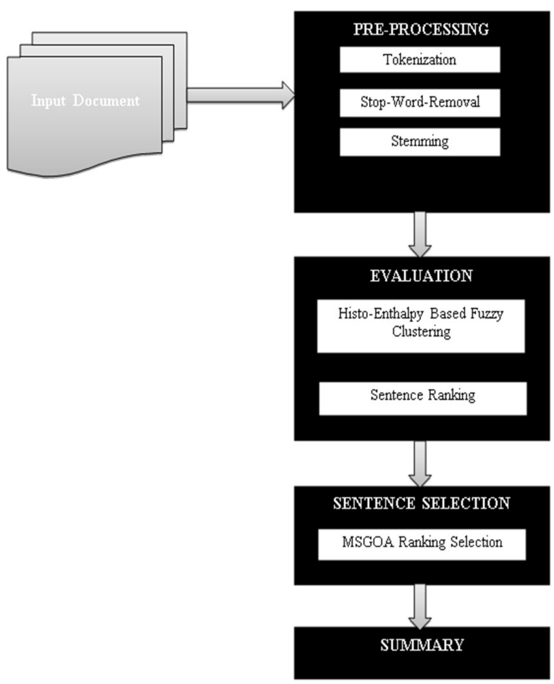

Figure 1: MSGOA sentence selection based block diagram

3.1 Preprocessing

The method of preprocessing for the detection of raw text reports for the input document(D) is first required for the development of text summarization. Tokenization, stop word removal, stemming, and N-gram are all examples of fundamental preprocessing.

3.1.1 Tokenization

Tokens are phrases, symbols, or other meaningful items that are used to break down a stream of text into words. To select a relevant sentence it requires the weight of the sentence. To find the weight of sentence required the verb weight for that splitting of the sentence as token. In generally tokenization is identified by the space in the given sentence. At its most basic level, text data is just a group of characters or a single word. Those words are required for any information retrieval processing. As a result, tokenization of documents is a prerequisite for a parser.

3.1.2 Stop word removal

Most commonly used words like 'the,' 'a,' 'and,' and 'this' are deleted from the textual data, which lacks semantic information, in this step. They are ineffective when it comes to document classification. As a result, they'll have to go. The creation of a stop words list, on the other hand, is difficult and inconsistent between textual sources. This method also decreases the amount of text data and boosts the system's efficiency. These words appear in every text document, although they are not required for text mining applications.

3.1.3 Stemming of Telugu language

Stemming is somewhat difficult Telugu type languages. By removing suffixes and prefixes from the words, it transforms them into their basic form in English language. The phrases: ['రాయటం','రాయడం','వ్రాయటం','రాసిన','వ్రాసిన'] might all be simplified to a single symbol 'వ్రాయడం'. In Telugu not only these phrases so many variants were possible.

3.1.4 N-Gram stemmers

Adamson and Boreham came up with this technique. The shared diagram method is what it's called. A pair of letters is shown in the diagram. The shared unique diagram is used to calculate association measures between pairs of terms in this method.

As an illustration: learning and learner are two words that come to mind.

le ea rn ni in ng learning

learner le ea rn ne er er er er er

In this example, learning contains six distinct diagrams, while learner has five distinct diagrams; these two words share three distinct schematics: le, ea, and rn. After determining the number of distinct diagrams, the Dice coefficient is used to calculate a similarity measure based on the distinct diagrams. [26], which is the definition of the dice coefficient.

$$S=2C/(A+B) \qquad (1)$$

3.2 Histo-Enthalpy Based Fuzzy C-means clustering

To improve the pre-prepared text documents histogram equalization is used. This works by thinking about just little areas and performs contrast upgrade of those regions of documents. Histogram equalization is represented by using the condition (2),

$$H = \sum_{I=1}^{n} H_d \qquad (2)$$

$H_d$is the collective histogram function value in the text document reports. Enthalpy-based fuzzy c-means clustering method clusters raw text document reports.

Input: Pre-processed text document d **Output:** Clustered text document reports
set the random centroid of the cluster $n \leftarrow 1az$ **Repeat** Enthalpy computation using condition (3) Compute the cost function of cluster using condition (6) Upgrade the cluster by condition (8) $n \leftarrow n+1$ **Until** stopping condition reached **Return clustered text document reports** I_c

Algorithm 1: Pseudo-code for Histo-Enthalpy Based fuzzy C-means Clustering

A histogram is a careful depiction of the circulation of mathematical information. The clustering steps are according to the accompanying algorithm1,

Stage 1: The enthalpy count for the histogram text document is given in condition (4).

$$E_y = \frac{1}{Max[H_d]} \qquad (3)$$

$$S(A,B) = 1\left(1+e^{-s_i(t)}\right) * E_y \qquad (4)$$

At this time, equation (3) is multiplied with the enthalpy function of equation (4) for improved text documents.

Stage2: Afterward the enthalpy measure, the enlargement action is one of the bases of morphology handling is utilized. $\overline{A}$ Is expanded by $\overline{B}$, composed as $\overline{A} \oplus \overline{B}$ characterized as:

$$\overline{A} \oplus \overline{B} = \left\{ z \middle| \left(\hat{B}\right)_z \cap \overline{A} \neq \phi \right\} \qquad (5)$$

Amongst them, ϕ is for the unoccupied set, $\overline{A}$ is for the organization component, and $\hat{B}$ is for the imprint of assortment $\overline{B}$. To put it plainly, that $\overline{A}$ is distended by $\overline{B}$ is the set shaped by the preliminary opinion seats of entirely essential fundamentals. Here, the pre-processed Text Documents are grouped into clusters. The clustering is to minimalize the cost function.

$$I_c = \sum_{k=1}^{n} \sum_{c=1}^{C} \frac{u_{kc}^m \left\| x_k - v_c \right\|^2}{S(A,B)} \qquad (6)$$

With constraints of,

$$\sum_{c=1}^{C} u_{kc} = 1, \quad 0 \le u_{kc} \le 1, \qquad (7)$$

By the side of this point, v_c is the centroid of c^{th} cluster, u_{kc} labels the fuzzy association of n^{th} text document in the direction of c^{th} cluster, m depicts the fuzziness of calculating [m ≥ 1], anticipated for example, slighter m recommends crisper grouping as well as $\|.\|$ attitudes for some norm. Here, it is Euclidean distance. Cluster Centre is calculated as,

$$v_c = \frac{\sum_{k=1}^{n} u_{kc} x_k}{\sum_{k=1}^{n} u_{kc}} \qquad (8)$$

Therefore, matrix U = [u_{kc}] as well as vector V = [v_c] necessity be efficient as declared by (8) in addition to (9) up until received the stopover criterion characteristically as

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$\left\| U(t) - U(t-1) \right\| \le \bar{\varepsilon} \qquad (9)$$

At this point, $\bar{\varepsilon}$ signifies the threshold limit. This enthalpy based fuzzy c-means clustering process; actual features are extracted on behalf of Cluster processing.

3.2.1 Sentence ranking

The approach of establishing the rank of each sentence as well as the peak hierarchical semantic sub-graphs is known as sentence ranking. To identify a most mono semantic network and semantic consistency throughout the sentence, the procedure only looks at the first rated rich semantic sub-graph. The average weight of each word and the average weight of the full sentence are calculated using (10) and (11). The frequency with which the phrase concept is used determines its importance (Word net usage popularity). The letter M gives the total number of tokens in a sentence (11). In the line "రాజు తను చేసే పనిలో దిట్ట" in this the word " రాజు " has only one concept (one sense) and hence has a weight of 10. Each of the three notions (senses) in the word "దిట్ట" has a weight of one (10, 7, and 6). In the word " చేసే పనిలో," there is only one notion with a weight of ten. Based on these values, the output rank values of these sentence rich semantic sub-graphs are (10, 9, and 8, 6).

C wht = 10 * (5 * ((n-1)/N)) (10)

S wht = (∑ Mm=0 C wht)/M (11)

3.3 Median Support Value based Grasshopper Optimization Algorithm (GOA)

The median support value estimation is defined on extricated ranging estimates such as P and Q is signified in condition (12).

$$\widetilde{S} = {(P+Q)}/{P*Q} \qquad (12)$$

Where p signifies the received feature extracted Text, Q signifies the time of input Text Document arrival.

The inquiry cycle is intelligently partitioned into two stages: exploration and exploitation, thanks to meta-heuristic calculations. The exploration phase is characterized by grasshoppers' long-range and unexpected developments, whereas the exploitation phase is characterized by neighborhood developments in search of better food supplies. The Support Value, as well as a numerical model for this behavior, are described.

$$x_i = S_i + G + A + \widetilde{S} \qquad (13)$$

Where, x_i - i^th position grasshopper,

S_i - Social connection in a gathering,

G -Power of gravity on the i grasshopper,

A - Wind direction. B

Elaborating S_i, G and A in (13), the equation is:

$$x_i = \sum_{j=1, j\neq 1}^{N} s\left(\left|x_j - x_i\right|\right)\frac{x_j - x_i}{d_{ij}} - g\hat{e}_g + \hat{u}e_w + \widetilde{S} \qquad (14)$$

Where, $s(r) = fe^{-r/l} - e^{-r}$ indicates the number of grasshoppers and denotes a function modeling the effect of social relationships.

Distance is calculated as $d_{ij} = \left|x_j - x_i\right|$.

Because grasshoppers discover acceptable zones quickly and have less convergence, then the equation rewritten as:

$$x_i = c\left(\sum_{j=1, j\neq i}^{N} c\frac{ub - lb}{2}\right) s\left(\left|x_j - x_i\right|\right)\frac{x_j - x_i}{d_{ij}} + \widetilde{T}_d + \widetilde{S} \qquad (15)$$

Where lb and ub - Characterize the upper and lower limits of rank,

$\widetilde{T}$ - Value comparative with the objective (best arrangement found up until now),

c - Diminishing coefficient (sentence ranking function) that the cycles of misuse adjust and investigation, which is characterized as below:

$$c = c_{\max} - I\frac{c_{\max} - c_{\min}}{Max_I} \qquad (16)$$

Where, $c_{\max}$ - Most extreme value (nearly 1),

$c_{\min}$ - Smallest value,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

I - Current iteration,

Max_I - Maximum no. of cycles.

Enhancement was mentioned in algorithm 1.

Initialize the ranking position and the parameters

Initialize $c_{\max}$, $c_{\min}$, and Max_I as the Maximum no. of iterations

Ascertain the ranking of each search operator and let T characterize the best ranking

while ($I < Max_I$) **do**

Compute sentence ranking fitness utilizing condition (16)

  for $i = 1 : N$ **do**

    Update x_i utilizing condition (14)

    Compute the ranking

    if current sentence ranking fitness is worse than target ranking **then**

x_i behaviour utilizing condition (15)

    end if

    Carry the current search agent back if it goes outside the

limits

  end for

  Update T and position

$I = I + 1$

end while

Return the target selecting ranking and position

***Algorithm 2:** Median support value based Grasshopper Optimization Algorithm.*

4. RESULTS

In this, we clearly stated the analysis of results and comparing with other results. The dataset we chosen is online research dataset used for the analysis of results approaches and comparing with other research approaches. This work is implemented using java platform.

4.1 Performances Measures

The statistical metrics of Accuracy, F-measure, Recall and Precision can be expressed in the terms of summary. Using the statistical indicators indicated in this part, we examine the performance of our planned work.

4.1.1 Accuracy

Accuracy will gives the correctness of our work and it is mentioned as the number of correct predictions /the total number of predictions. Number of correct predictions-(TN+TP), Total no. of Predictions-($\mathrm{TN+TP+FN+FP}$)

$$\text{Accuracy} = \frac{(\mathrm{TN+TP})}{(\mathrm{TN+TP+FN+FP})} \quad (17)$$

Where, TN is true negative, TP is the true positive, FP is the false positive, and FN is the false negative.

4.1.2 Recall

Precision is a metric for how relevant the extracted phrases are, whereas recall is a metric for how many truly relevant results are produced when utilizing equations (18),

$$\mathrm{Re}\,call = \frac{\text{Extracted Summary} \cap \text{Provided Summary}}{\text{Provided Summary}} \quad (18)$$

4.1.3 Precision

Precision is the proportion of the predicted positive instances that were correct text size, as calculated using equation (19),

$$\Pr ecision = \frac{\text{ExtractedSummary} \cap \text{ProvidedSummary}}{\text{ExtractedSummary}} \quad (19)$$

4.1.4 F-measure

It is a measure of a test's accuracy. Maintain a balanced state among the Recall and Precision is given in equation (20),

$$Fmeasure = \frac{2\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (20)$$

Figure 2 shows the performance metrics Recall, Precision, F-measure, Time and Accuracy for different number of iterations are taken, and agreeing with the below figure when the iterations are 600 or advanced; then the performance extents it's extreme and become stable.

Table 1: Sentences With And Without Stop Words Of Telugu Text.

S.No	Sentences with and without Stop words	Total Words
1	ఇటీవలి సంవత్సరాలలో, స్మార్ట్ మార్గంలో సమాచారాన్ని ఉత్పత్తి చేయగల పరికరాల సంఖ్య విపరీతంగా పెరిగింది, ఇది ఇంటర్నెట్ ఆఫ్ థింగ్స్ (IoT) కు దారితీసింది ఇటీవలిసంవత్సరాలలోసంఖ్యపరికరాలుసమాచారాన్నిఉత్పత్తిచేయగలస్మార్ట్మార్గంవిపరీతంగాపెరిగింది, ప్రముఖఇంటర్నెట్థింగ్స్ (IoT) In recent years, the number of devices capable of generating information in a smart way has grown exponentially, leading to the Internet of Things (IoT)	18 15
2	IoT పరికరం స్మార్ట్ఫోన్, సెన్సార్, టాబ్లెట్ లేదా ధరించగలిగే ఏదైనా తెలివైన పరికరం కావచ్చు, ఇది క్రొత్త సమాచారాన్ని పంపగలదు లేదా ఉత్పత్తి చేయగలదు IoT పరికరస్మార్ట్ఫోన్, సెన్సార్, టాబ్లెట్రాణిఇతరధరించగలిగేపరికరంపంపడంకొత్తసమాచారాన్నిఉత్పత్తిచేస్తుంది IoT device can be smart phone, sensor, tablet or any other wearable device that can send or generate new information	18 15
3	రెండు సమస్యలను నిజ సమయంలో పరిష్కరించాలి, ప్రభావిత వినియోగదారులకు వెంటనే తెలియజేయబడే పరిస్థితులను గుర్తించడానికి భిన్నమైన డేటాను ప్రాసెస్ చేసే మరియు విశ్లేషించే సామర్థ్యం అవసరం సమస్యలనునిజసమయంలోపరిష్కరించాల్సినఅవసరంఉంది, సామర్థ్యప్రక్రియవేర్వేరుడేటానువిశ్లేషిస్తుంది Both issues need to be addressed in real time, with the ability to process and analyze different data to identify situations where affected customers will be notified immediately	19 11
4	అందువల్ల, వైవిధ్య డేటా వనరులను ప్రాసెస్ చేసే చాలా వ్యవస్థలు ప్రీ-ప్రాసెసింగ్ లేదా సాధారణీకరణను చేయమని బలవంతం చేయబడతాయి, అటువంటి డేటాను విశ్లేషించగలవు. డేటా సాధారణీకరణ దశకు సాధారణంగా చాలా ప్రాసెసింగ్ పనులు అవసరం, వనరులు మరియు సమయాన్ని తీసుకుంటుంది చాలావ్యవస్థలువిభిన్నడేటావనరులనుప్రాసెస్చేస్తాయి, ప్రీ-ప్రాసెసింగ్సాధారణీకరణనుబలవంతంచేస్తాయి, అటువంటిడేటానువిశ్లేషించండి. డేటాసాధారణీకరణదశకుసాధారణంగాచాలాప్రాసెసింగ్పనులు, వనరులసమయంఅవసరం Therefore, most systems that process diverse data sources are forced to do pre-processing or generalization, which can analyze such data. The data normalization phase usually requires a lot of processing tasks, resources and time	29 24
5	కొన్ని విధానాలు మూలాలపై సజాతీయతను నిర్వహించడానికి ప్రయత్నించాయి, పెద్ద ఐయోటి నెట్వర్క్లను భారీ మొత్తంలో పరికరాలతో మూలాలుగా పరిగణించేటప్పుడు ఇది తగినంత సమర్థవంతంగా ఉండదు విధానాలుసజాతీయమూలాలనునిర్వహించడానికిప్రయత్నించాయి, పెద్దIoT నెట్వర్క్లమూలాలనుపెద్దమొత్తంలోపరికరాలనుపరిగణనలోకితీసుకుంటేసరిపోతుంది Some policies have attempted to maintain homogeneity on sources, which may not be effective enough when considering large IoT networks as sources with large amounts of devices	18 15

Journal of Theoretical and Applied Information Technology
15th September 2022. Vol.100. No 17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

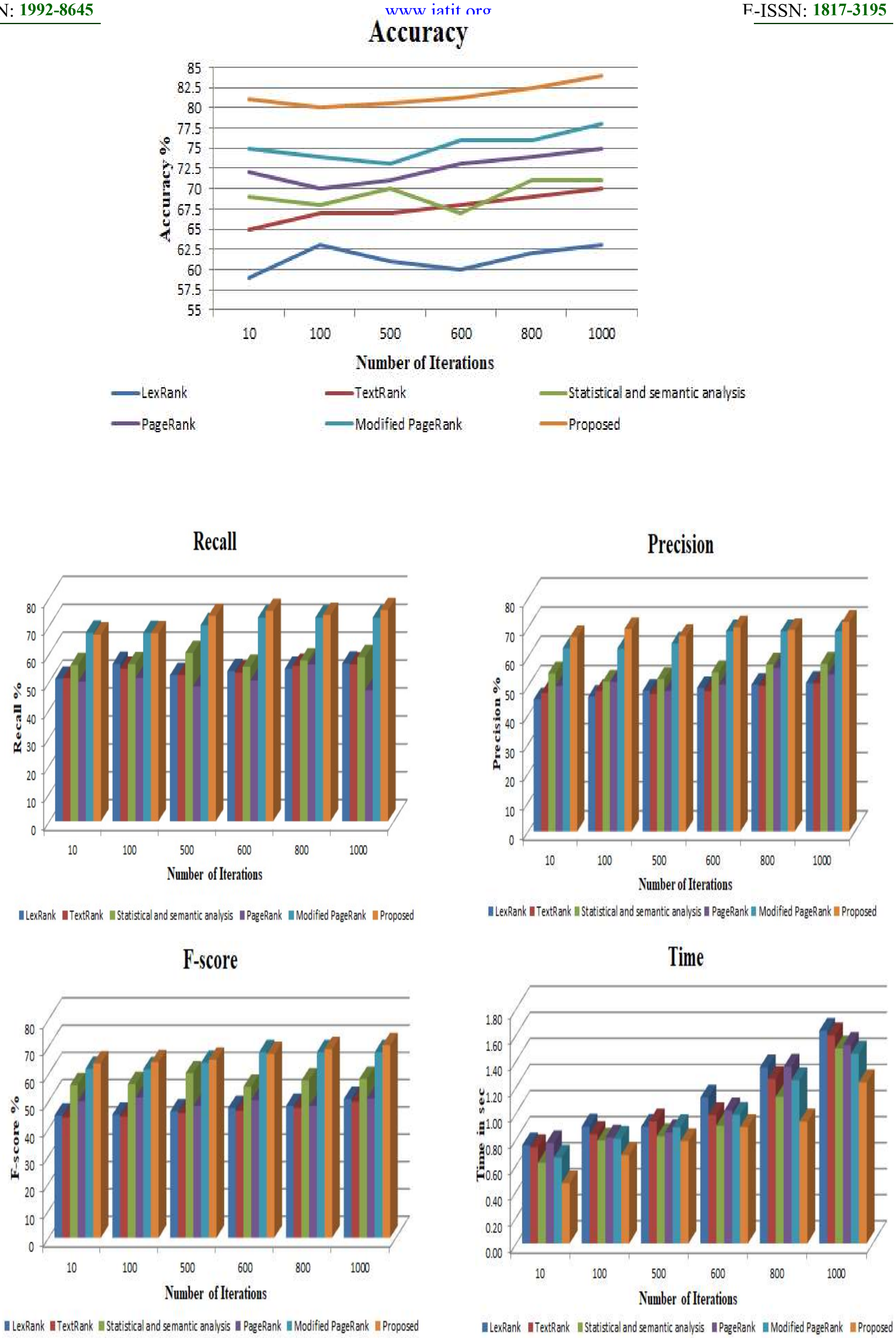

Figure 2: Performance Estimation During Iteration Related With Other Research Approaches.

From table 1 the sentences with and without stop word is clearly stated for the Telugu text document analysis based on our proposed approach techniques. The prediction of Telugu document contains the words the stop words are removed by the Stemming process of the proposed approach of N-Gram stemmers is used to avoid the stop words.

The tenth iteration of performance metrics for existing and proposed techniques was in table 2, The hundredth iteration of performance metrics for existing and proposed techniques in table3, The 500th iteration of performance metrics for existing and proposed techniques in table4, The 600th iteration of performance metrics for existing and proposed techniques in table5, The 800th iteration of performance metrics for existing and proposed techniques in table6, the seventh table according to the data stated below, the improvement of algorithm based on performance as the number of iterations increases until it reaches 600 iterations, at which point it stabilizes. For iteration, the Median Support Value based Grasshopper Optimization Algorithm is employed, and each vertex (sentence) gets a new weight based on the sentence ranking, which is determined by the previous rank of the word and weight linking the nodes. According to no of iterations the performance esteem changes the values iterations of 10, 100, 500, 600, 800, and 1000. When the iterations increase the accuracy value is also increased.

Table 2: 10th Iterations Varying The Performance Metrics Values.

10 Iteration	**Preci sion**	**Rec all**	**F-scor e**	**Ti me**	**Accur acy**
LexRank	45.36	51. 02	44.7 5	0.7 53	59
TextRank	47.55	51. 27	44.1 2	0.7 38	65
Statistical and semantic analysis	54.3	55. 94	55.9 4	0.6 21	69
PageRank	50	50	50	0.7 69	72
Modified PageRank	62.9	67. 51	61.9 1	0.6 59	75
Proposed	66.78	66. 89	63.8 9	0.4 62	81

Table 3: 100th Iterations Varying The Performance Metrics Values Of The Proposed Approach.

100 Iteration	**Preci sion**	**Rec all**	**F-scor e**	**Ti me**	**Accur acy**
LexRank	46.3	56. 35	45.1 2	0.8 96	63
TextRank	48.26	54. 65	44.3 9	0.8 39	67
Statistical and semantic analysis	51.48	56. 32	56.3 2	0.7 89	68
PageRank	51.36	51. 36	51.3 6	0.8 13	70
Modified PageRank	62.76	67. 43	61.7 6	0.8 04	74
Proposed	69.52	67. 32	64.3 2	0.6 78	80

Table 4: 500th Iterations Varying The Performance Metrics Values.

500 Iteration	**Precis ion**	**Rec all**	**F-scor e**	**Ti me**	**Accur acy**
LexRank	48.32	52.3 6	46.3 2	0.8 95	61
TextRank	47.25	52.3 9	45.6 3	0.9 36	67
Statistical and semantic analysis	52.36	60.2 5	60.2 5	0.8 23	70
PageRank	48.3	48.3	48.3	0.8 53	71
Modified PageRank	64.6	70.2 7	64.1 2	0.8 92	73
Proposed	67.12	73.5 6	65.2 5	0.7 85	80.56

Table 5: 600th iterations Varying The Performance Metrics Values.

600 Iteration	**Preci sion**	**Rec all**	**F-scor e**	**Ti me**	**Accur acy**
LexRank	49.26	53. 79	47.5 6	1.1 2	60
TextRank	48.26	53. 25	46.5 3	0.9 85	68
Statistical and semantic analysis	54.68	55. 34	55.3 4	0.9 03	67
PageRank	50.45	50. 45	50.4 5	1.0 21	73
Modified PageRank	68.75	72. 93	67.9 8	0.9 87	76
Proposed	70.13	75. 36	67.2 9	0.8 93	81.25

Table 6: 800thiterations Varying The Performance Metrics

800 Iteration	Precision	Recall	F-score	Time	Accuracy
LexRank	50.36	54.68	48.32	1.352	62
TextRank	50	55.65	47.56	1.263	69
Statistical and semantic analysis	57.32	57.62	57.62	1.124	71
PageRank	56.23	56.23	48.28	1.356	74
Modified PageRank	68.75	72.93	67.98	1.254	76
Proposed	69.2	73.89	69.12	0.935	82.48

Table 7: 1000thIterations varying the performance metrics.

1000 Iteration	Precision	Recall	F-score	Time	Accuracy
LexRank	51.03	56.5	50.86	1.632	63
TextRank	50.88	56.22	49.81	1.596	70
Statistical and semantic analysis	57.62	58.8	58.2	1.496	71
PageRank	54	47	51	1.523	75
Modified PageRank	68.75	72.94	67.98	1.456	78
Proposed	72	75.69	70.56	1.236	84

Table 8: Iterations varying the performance metrics values of the proposed approach

*Iterations	Precision	Recall	F-score	Time	Accuracy
10	66.78	66.89	63.89	0.462	81
100	69.52	67.32	64.32	0.678	80
500	67.12	73.56	65.25	0.785	80.56
600	70.13	75.36	67.29	0.893	81.25
800	69.2	73.89	69.12	0.935	82.48
1000	72	75.69	70.56	1.236	84

Table 7 and Figure2 shows the difference between the current research performances metrics other results LexRank, TextRank, Statistical and semantic analysis, PageRank, Modified Page Rank results. The dataset used in the proposed work is online research article dataset. The value prediction of Precision, recall, f-measure, time and accuracy are derived according to the proposed MSGOA. Figure 3 estimates the number of data that is sentences given increases the performance measures in Accuracy, Recall, Precision, F-score and Time.

.

Journal of Theoretical and Applied Information Technology
15th September 2022. Vol.100. No 17

ISSN: 1992-8645 E-ISSN: 1817-3195

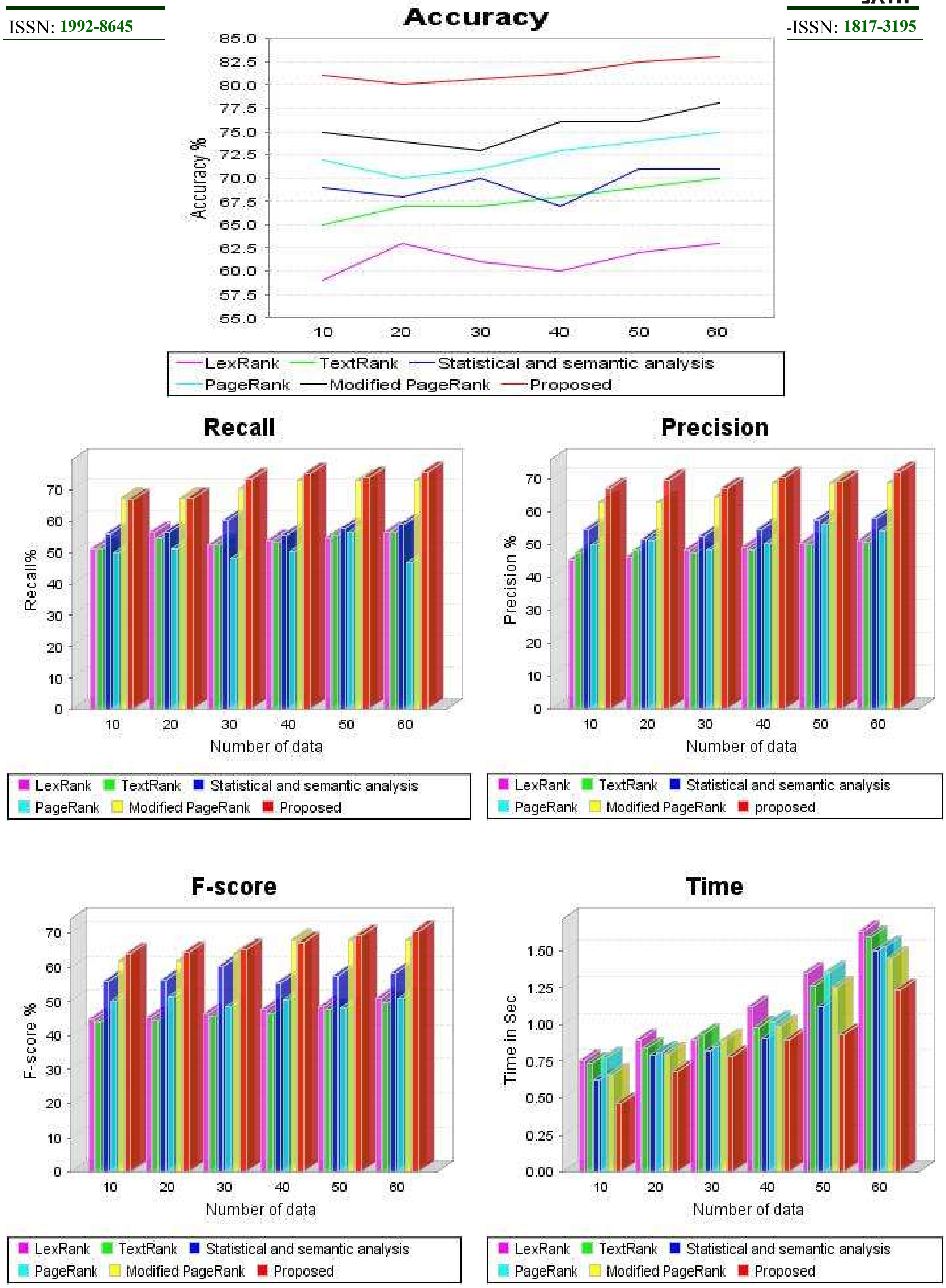

Figure 3: Performance Estimation Of Sentences Related With Other Research Approaches.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5. CONCLUSION

The research article concludes that the Telugu language Text summarization attains the basic stages of preprocessing, Tokenization, Stop-word removal and stemming using the clustering method of fuzzy C-means combined with Histo-Enthalpy based. After that ranking selection of text Optimization technique used is Median Support Based Grasshopper Optimization Algorithm (MSGOA) successfully attained the summary. The results are derived finally with proposed approaches to predict the text summary accurately. The results show that as the number of iterations grows, the median support value based Grasshopper Optimization Algorithm (GOA) produces better outcomes. This investigation, according to past studies, has superior outcomes than the others. This study's final accuracy rating is 84%.

REFERENCES

[1] Nawaz A, Bakhtyar M, Baber J, Ullah I, Noor W, Basit A. Extractive Text Summarization Models for Urdu Language. Information Processing & Management. 2020 Nov 1; 57(6):102383.

[2] Mutlu B, Sezer EA, Akcayol MA. Candidate sentence selection for extractive text summarization. Information Processing & Management. 2020 Nov 1; 57(6):102359.

[3] Tanwi, Satanik Ghosh, Viplav Kumar, Yashika S Jain, Mr. Avinash B, " Aut omat ic T ext Summarizat ion using T ext Rank", (IRJET), Volume: 06 Issue: 04, April 2019, e-ISSN: 2395-0056.

[4] K. U. Manjari, S. Rousha, D. Sumanth and J. Sirisha Devi, "Ext ract ive Text Summarizat ion from Web pages using Selenium and TF-IDF algorithm," 2020 4th Internat ional Conference on Trends in Elect ronics and Informat ics (ICOEI)(48184), Tirunelveli, India, 2020, pp. 648-652, doi: 10.1109/ICOEI48184.2020.9142938.

[5] Sana Shashikanth, Sriram Sanghavi, "Text Summarization Techniques Survey on T elugu and Foreign Languages", International Journal of Research in Engineering, Science and Management, Volume-2, Issue-1, January-2019. ISSN: 2581- 5792.

[6] Rumagit, R. Y., Setiyawati, N., & Bangkalang, D. H. (2019). Comparison of graph-based and term weighting method for automatic summarization of online news. Procedia Computer Science, 157, 663–672.

[7] Qaroush, A., Farha, I. A., Ghanem, W., Washaha, M., & Maali, E. (2019). An efficient single document arabic text summarization using a combination of statistical and semantic features. Journal of King Saud University-Computer and Information Sciences.

[8] Noor, F., Bakhtyar, M., & Baber, J. (2019). Sentiment analysis in e-commerce using svm on roman urdu text. International conference for emerging technologies in computing. Springer213–222.

[9] Miao, L., Cao, D., Li, J., & Guan, W. (2020). Multi-modal product title compression. Information Processing & Management, 57(1), 102123. https://doi.org/10.1016/j.ipm.2019. 102123.

[10] Mutlu, B., Sezer, E. A., & Akcayol, M. A. (2019). Multi-document extractive text summarization: A comparative assessment on features. Knowledge-Based Systems, 183, 104848.

[11] Mutlu, B., Sezer, E. A., & Akcayol, M. A. (2020). Investigation of text summarization features by fuzzy systems and convolutional neural networks. Submitted: Semantics conference 2020.

[12] Joshi, A., Fidalgo, E., Alegre, E., & Fernández-Robles, L. (2019). Summcoder: An unsupervised framework for extractive text summarization based on deep auto-encoders. Expert Systems with Applications, 129, 200–215.

[13] Nasar, Z., Jaffry, S. W., & Malik, M. K. (2019). Textual keyword extraction and summarization: State-of-the-art. Information Processing & Management, 56(6), 102088. https://doi. org/10.1016/j.ipm.2019.102088.

[14] Yasunaga, M., Kasai, J., Zhang, R., Fabbri, A., Li, I., Friedman, D., & Radev, D. (2019). ScisummNet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. Proceedings of AAAI 2019.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[15] R. Lan, Y. Zhou, Z. Liu, X. Luo, Prior knowledge-based probabilistic collaborative representation for visual recognition, IEEE Trans. Cybern. (2018) 1–11, doi: 10. 1109/TCYB.2018.2880290 .

[16] Liang Z, Du J, Li C. Abstractive Social Media Text Summarization using Selective Reinforced Seq2Seq Attention Model. Neurocomputing. 2020 May 12.

[17] Qian, X., Li, M., Ren, Y., Jiang, S., 2019. Social media based event summarization by user text image co-clustering. Knowl. Based Syst. 164, 107–121. https://doi.org/10.1016/j.knosys.2018.10.028.

[18] Curiel A, Gutiérrez-Soto C, Rojano-Cáceres JR. An online multi-source summarization algorithm for text readability in topic-based search. Computer Speech & Language.;66:101143.

[19] Curiel A, Gutiérrez-Soto C, Rojano-Cáceres JR. An online multi-source summarization algorithm for text readability in topic-based search. Computer Speech & Language.;66:101143.

[20] Mao, X., Yang, H., Huang, S., Liu, Y., and Li, R. (2019). Extractive summarization using supervised and unsupervised learning. Expert Systems with Applications, 133:173–181.

[21] Naidu, Reddy, Santosh Kumar Bharti, Korra Sathya Babu, and Ramesh Kumar Mohapatra. "Text summarization with automatic keyword extraction in telugu e-newspapers." In Smart Computing and Informatics, pp. 555-564. Springer, Singapore, 2018.

[22] Du, Yan, and Hua Huo. "News Text Summarization Based on Multi-Feature and Fuzzy Logic." IEEE Access 8 (2020): 140261-140272.

[23] Hernández-Castañeda, Ángel, René Arnulfo García-Hernández, Yulia Ledeneva, and Christian Eduardo Millán-Hernández. "Extractive Automatic Text Summarization Based on Lexical-Semantic Keywords." IEEE Access 8 (2020): 49896-49907.

[24] Alqaisi, Rana, Wasel Ghanem, and Aziz Qaroush. "Extractive Multi-Document Arabic Text Summarization Using Evolutionary Multi-Objective Optimization With K-Medoid Clustering." IEEE Access 8 (2020): 228206-228224.

[25] Fucheng You, Shuai Zhao, and Jingjing Chen. "A Topic Information Fusion and Semantic Relevance for Text Summarization." IEEE Access 8 (2020): 178946-178953.

[26] Kannan, Subbu, Vairaprakash Gurusamy, S. Vijayarani, J. Ilamathi, and M. Nithya. "Preprocessing techniques for text mining." International Journal of Computer Science & Communication Networks 5, no. 1 (2014): 7-16

[27] Moawad, Ibrahim F., and Mostafa Aref. "Semantic graph reduction approach for abstractive Text Summarization." In 2012 Seventh International Conference on Computer Engineering & Systems (ICCES), pp. 132-138. IEEE, 2012.

[28]Mallikarjuna Reddy, A., Rupa Kinnera, G., Chandrasekhara Reddy, T., Vishnu Murthy, G., et al., (2019), "Generating cancelable fingerprint template using triangular structures", Journal of Computational and Theoretical Nanoscience, Volume 16, Numbers 5-6, pp. 1951-1955(5), doi: https://doi.org/10.1166/jctn.2019.7830.

[29]Mallikarjuna Reddy, A.,Venkata Krishna, V. and Sumalatha, L." Face recognition approaches: A survey" International Journal of Engineering and Technology (UAE), 4.6 Special Issue 6, volume number 7 , 117-121,2018.

[30]C. R. T, G. Sirisha and A. M. Reddy, "Smart Healthcare Analysis and Therapy for Voice Disorder using Cloud and Edge Computing," *2018 4th International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, Mangalore, India, 2018, pp. 103-106, doi: 10.1109/iCATccT44854.2018.9001280.

[31]Ilaiah Kavati, A. Mallikarjuna Reddy, E. Suresh Babu, K. Sudheer Reddy, Ramalinga Swamy Cheruku,Design of a fingerprint template protection scheme using elliptical structures,ICT Express,Volume 7, Issue 4,2021,Pages 497-500,ISSN 2405-9595,https://doi.org/10.1016/j.icte.2021.04.001

[32]M. Samel, A. Mallikarjuna Reddy. (2022). An Empirical Study on Copy-Move Forgery Detection Techniques in Images. Mathematical Statistician and Engineering Applications, 71(3), 183 –. Retrieved from

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

https://www.philstat.org.ph/index.php/MSEA/article/view/136

[33]Salma Samreen Shaik, A. M. R. . (2022). An Automatic Plant Disease Detection System Using Deep Learning Technique. Mathematical Statistician and Engineering Applications, 71(3), 152–158. Retrieved from https://www.philstat.org.ph/index.php/MSEA/article/view/121

[34]A Mallikarjuna Reddy, Vakulabharanam Venkata Krishna, Lingamgunta Sumalatha and Avuku Obulesh, “Age Classification Using Motif and Statistical Features Derived On Gradient Facial Images”, Recent Advances in Computer Science and Communications (2020) 13: 965. https://doi.org/10.2174/2213275912666190417151247.

[35]Swarajya lakshmi v papineni, A.Mallikarjuna Reddy, Sudeepti yarlagadda , Snigdha Yarlagadda, Haritha Akkineni *"An Extensive Analytical Approach on Human Resources using Random Forest Algorithm"* International Journal of Engineering Trends and Technology 69.5(2021):119-127.

[36]Dayaker, P., Honey Diana, Chandrasekhara Reddy, T., Mallikarjuna Reddyreddy, A.” Advancements of security and privacy of sensitive data cloud computing” Jour of Adv Research in Dynamical & Control Systems, Vol. 10, 11-Special Issue, 2018.

[37]A. Mallikarjuna Reddy, V. Venkata Krishna, L. Sumalatha, “Efficient Face Recognition by Compact Symmetric Elliptical Texture Matrix (CSETM)”, Jour of Adv Research in Dynamical & Control Systems, Vol. 10, 4-Regular Issue, 2018.

[38]Swarajya Lakshmi V Papineni, Snigdha Yarlagadda, Harita Akkineni, A. Mallikarjuna Reddy. Big Data Analytics Applying the Fusion Approach of Multicriteria Decision Making with Deep Learning Algorithms *International Journal of Engineering Trends and Technology*, 69(1), 24-28, doi: 10.14445/22315381/IJETT-V69I1P204.

[38]Chandrasekhara Reddy, T., Pranathi, P., Mallikarjun Reddy, A., Vishnu Murthy, G., Kavati,I.”Biometric template security using convex hulls features” Journal of Computational and Theoretical Nanoscience, Volume 16, Numbers 5-6, May 2019, pp. 1947-1950(4),doi: 10.1166/jctn.2019.7829.

[40] B. Pruthvi Raj Goud, G. L. Anand Babu, G. Sekhar Reddy, A. Mallikarjuna Reddy “Multiple object detection interface using HSV, hough and haar-classifier” International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-9, July 2019.