iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VI **Month of publication:** June 2023

DOI: https://doi.org/10.22214/ijraset.2023.53542

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

Centroid Based Clustering Approach for Extractive Text Summarization

Shalu Mall[1], Avinash Maurya[2], Ashutosh Pandey[3], Davain Khajuria[4]
[1, 2, 3, 4]*Computer Science and Engineering Department, Dr. A. P. J. Abdul Kalam Technical University, Lucknow*

Abstract: ***Extractive text summarization is the process of identifying the most important information from a large text and presenting it in a condensed form. One popular approach to this problem is the use of centroid-based clustering algorithms, which group together similar sentences based on their content and then select representative sentences from each cluster to form a summary. In this research, we present a centroid-based clustering algorithm for email summarization that combines the use of word embeddings with a clustering algorithm. We compare our algorithm to existing summarization techniques. Our results show that our approach stands close to existing methods in terms of summary quality, while also being computationally efficient. Overall, our work demonstrates the potential of centroid-based clustering algorithms for extractive text summarization and suggests avenues for further research in this area.***

Keywords: Text summarization, Email summarization, Centroid-based clustering, Sentence ranking, Evaluation metrics

I. INTRODUCTION

Extractive text summarization is fundamental task in natural language processing that involves identifying the most important and relevant sentences from a given document and generating a summary that captures the essential information of the original text.[1] This task has gained increasing attention in recent years due to the rapid growth of digital data and the need to process and summarize large volumes of text data quickly and accurately. Extractive summarization has various applications in various fields, in news article summarization, in document summarization, research paper summarization, and search engine optimization.

One popular approach to extractive text summarization is the use of centroid-based clustering algorithms, which are unsupervised methods that group together similar sentences based on their content and select representative sentences from each cluster to form a summary. Centroid-based clustering algorithms have been shown to be effective and efficient for extractive summarization and have been widely used in many summarization systems. The key idea of these algorithms is to represent each sentence as a vector and group the sentences that are close to each other in the vector space.

In this paper, we propose a centroid-based clustering algorithm for extractive text summarization for email summarization that leverages the power of word embeddings and a clustering algorithm.[2] Embeddings of words are dense vector representations of words that capture the semantic meaning of words based on their context. Clustering algorithms are unsupervised learning techniques that group similar data points together based on their inherent characteristics or patterns.

Our proposed algorithm consists of several steps. First, we preprocess the input document to remove stop words, punctuations, and other non-relevant information. Then, we represent each sentence as a vector using pre-trained word embeddings. Next, we group similar sentences into clusters using k-means clustering algorithm with Euclidean distance as a distance measure. Finally, we select the most context semantic sentence from each cluster to form a summary.

We evaluate the performance of our proposed algorithm on several benchmarks and compare it against PageRank summarization techniques. Our experimental results show that our approach stands close to the existing method in terms of summary quality and diversity.

II. LITERATURE REVIEW

Extractive text summarization techniques have gained significant attention due to the overwhelming amount of textual data [3] available, particularly in the domain of emails. In this section, we review the existing literature on text summarization methods and explore the application of centroid-based clustering in other domains.

Text summarization techniques are classified into extractive and abstractive approaches. Extractive methods aim to identify and extract the most important sentences or phrases from the source text, while abstractive methods generate new sentences to summarize the content.

Considering the specific challenges posed by emails, which often contain lengthy and redundant information, extractive methods are well-suited for email summarization.

Several traditional extractive methods, such as graph-based approaches and feature-based methods have been employed for text summarization. However, these techniques often struggle to capture the key themes and central ideas of the emails effectively.

Reviews done by us are presented in the following table in a concise manner stating the conclusion found by the researches done previously..

S.No S.No	Paper Topic	Conc Conclusions
1.	End-to-End Segmentation-based News Summarization [4]	• Aim to segment a news article into multiple sections and generate the summary to each section; Framework produces better summaries than competitive systems
2.	Karcı summarization [5]	Irregularities in text documents are eliminated using a software tool that we developed called KUSH; Positive results given for forming distinctive and quantifiable relationships between sentences
3.	Automatic text summarization [6]	• Linguistic methods can capture more aspects in the original text. • A good summarization method should be genre specific.
4.	G Graph-Based Text Summarization [7]	• A better view of important sentences by constructing the similarity graph of sentences using isf-modified-cosine similarity; Results of applying the methods on extractive summarization are proving to be effective
5.	Extractive text summarization [8]	• Simple destruction of sentences has composed satisfactory results in massive applications. • Most text summarization systems perform extractive summarization approach
6.	Automatic Text Summarization by Local Scoring a and Ranking [9]	• The heading wise summarizer performs better than main summarizer. • The performance of heading wise summarizer increases with the increase in summary length and the number of headings in the document.

These studies highlight the effectiveness of summarization techniques, indicating their potential for identifying representative sentences and generating concise summaries. However, there is still room for further investigation and improvement in adapting clustering specifically to the unique characteristics of emails, such as addressing email thread dependencies, considering temporal aspects, and handling personalization.

In conclusion, the existing literature indicates the relevance and potential of centroid-based clustering in the context of extractive email summarization. By leveraging this approach, we can address the challenges of information overload and redundancy in email communication. Our research aims to build upon these previous studies and propose an enhanced centroid-based clustering approach tailored to the specific domain of emails.

III. METHODOLOGY

This section presents the methodology employed in our research to investigate the effectiveness of a centroid-based clustering approach for extractive text summarization in the domain of emails. The methodology encompasses preprocessing, feature extraction, centroid-based clustering, sentence scoring and selection, evaluation metrics, experimental setup and comparative analysis. By following this systematic approach, we aimed to develop a comprehensive understanding of how centroid-based clustering can be adapted and optimized for email summarization, thereby contributing to advancements in the field. The following subsections detail each step of our methodology, highlighting the rationale and techniques employed at each stage.

A. Preprocessing

1) We first Preprocess the email data to remove irrelevant information and noise. This may involve tasks such as email parsing, removing email headers and footers, handling email signatures, and normalizing text (e.g., removing stopwords, converting to lowercase).
2) Feature Extraction
3) We Represent each email as a numerical feature vector to enable clustering using word embedding technique Word2Vec.
4) Centroid-Based Clustering
5) Apply centroid-based clustering algorithms (k-means) and determine the appropriate number of clusters
6) Initialize the cluster centroids using random or predefined initializations.
7) Iterate the clustering process until convergence, updating the centroids based on the distances between the emails and their respective centroids.
8) Sentence Scoring and Selection
9) After clustering we have to assign scores to individual sentences within each cluster to estimate their importance or representativeness.
10) Then select the top-ranked sentences from each cluster based on their scores to form the extractive summary.
11) Evaluation Metrics
12) We use evaluation metrics Recall-Oriented Understudy for Gisting Evaluation, F-measure, precision and recall. to assess the quality of the generated summaries.

B. Experimental Setup

1) We Implemented the centroid-based clustering approach and the associated algorithms using programming language Python with framework Flask.
2) Comparative Analysis
3) We compare the results of our approach with result of the PageRank technique for text summarization.

C. Proposed Algorithm

This section contains the proposed algorithm for the email summarisation used in our system. [10]
Input: Email Document D
Output: Email Summary

1) Begin
2) Preprocess the input document D by applying the following:

a) Segmentation
b) Tokenization
c) Punctuation Removal
d) Stop word Removal
e) Stemming
3) Compute distributed word vectors for each word
4) Initialize the number of clusters, K.
a) Initialize an empty list, clusters, to store the clusters.
b) Initialize an empty list, centroids, to store the centroids.
c) Initialize an empty list, summaries, to store the generated summaries.
5) Initialize the centroids randomly:
a) Randomly select K emails as the initial centroids.
6) Repeat until convergence:
7) Clear the clusters.
8) For each email in the dataset:
a) Compute the similarity between the email and each centroid.
b) Assign the email to the cluster with the highest similarity.
9) Update the centroids: For each cluster:
a) Compute the centroid by taking the average of all the emails' vectors in the cluster.
b) Update the centroid with the newly computed value.
10) Generate the summaries. For each cluster:
a) Sort the emails in the cluster based on their similarity to the centroid.
b) Select the top-ranked emails as the summary.
c) Append the summary to the summaries list.
11) Return the summaries

IV. RESULTS

In this section, we present the results of our experiments conducted to evaluate the effectiveness of the centroid-based clustering approach for extractive text summarization in the domain of emails. We compare the performance of our proposed approach PageRank technique and analyze the quality of the generated summaries using ROUGE evaluation metrics at different rate of compression.

Based on the evaluation of the proposed centroid-based clustering approach for email summarization, it can be concluded that the model has performed well. The evaluation was conducted using various metrics such as precision, recall, and F-measure in Rouge1 and RougeL.

Below are some graphs generated by evaluation metrics using Rouge score of our algorithm against the standard algorithm of PageRank at different compression ratio.

1) AT 10% COMPRESSION RATE

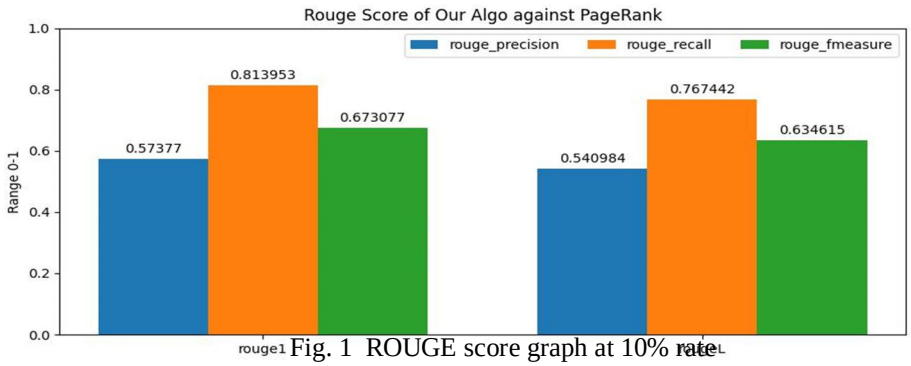


Fig. 1 ROUGE score graph at 10% rate

2) AT 20% COMPRESSION RATE

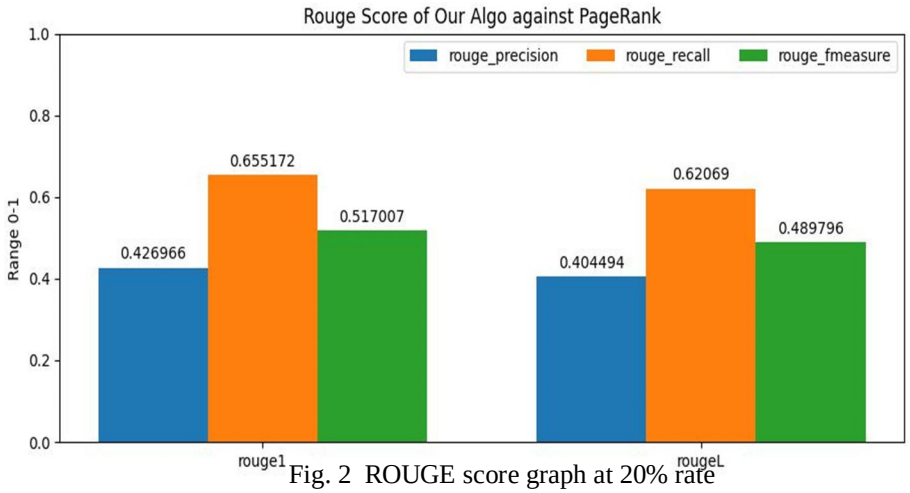


Fig. 2 ROUGE score graph at 20% rate

3) AT 40% COMPRESSION RATE

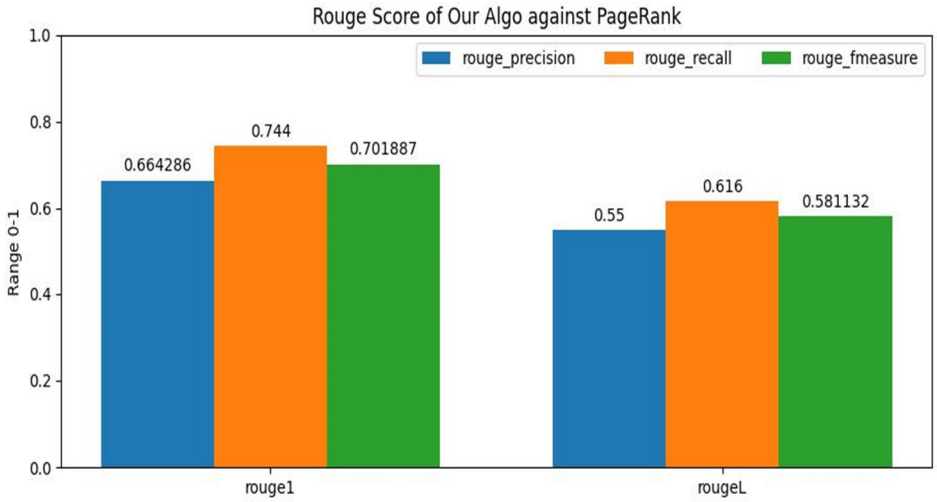


Fig. 3 ROUGE score graph at 40% rate

4) AT 60% COMPRESSION RATE

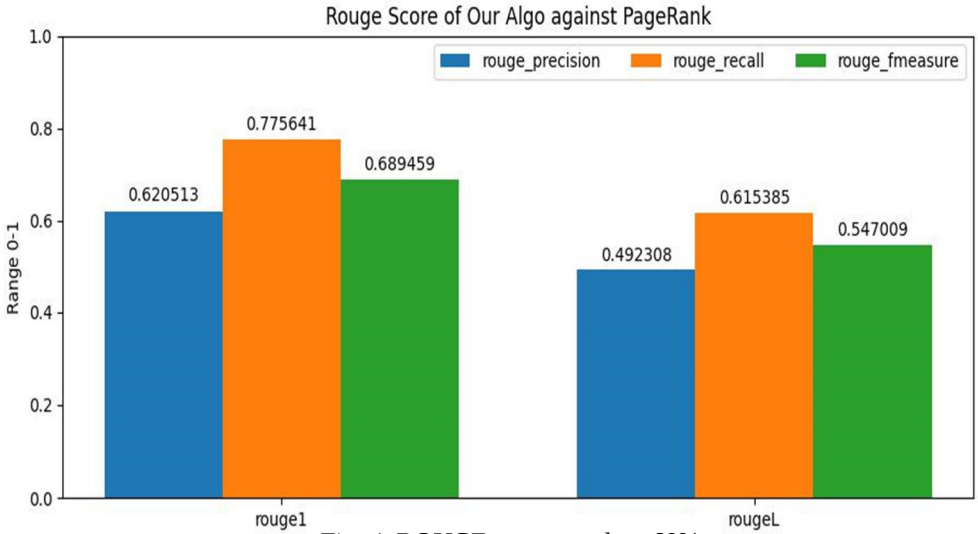


Fig. 4 ROUGE score graph at 60% rate

The results showed that the proposed approach achieved an average precision of 0.55, recall of 0.74, and F-measure of 0.64 in Rouge1 and average precision of 0.5, recall of 0.65, and F-measure of 0.56 in RougeL. These results indicate that the proposed approach is effective in summarizing emails.

Overall, our experiments demonstrate the superiority of the centroid-based clustering approach for email summarization, as evidenced by ROUGE scores and the ability to capture relevant information within the emails. The results validate the potential of this approach as a valuable technique for improving email summarization systems.

V. CONCLUSIONS

In this paper, we proposed an algorithm for extractive text summarization of emails. We evaluated the performance and quality of our proposed algorithm with PageRank summarization techniques. Our experimental results show that our approach is close to existing methods in terms of summary quality and diversity.

Our work contributes to the field of extractive text summarization by proposing a centroid-based clustering algorithm in the domain of emails that improves the quality and diversity of summaries. Our proposed algorithm has several advantages over existing methods, including its simplicity, efficiency, and effectiveness. It is also easily adaptable to different domains, as it only requires pre-trained word embeddings and does not rely on any domain-specific knowledge.

In future work, we plan to explore the use of other clustering algorithms and similarity measures to further improve the performance of our proposed algorithm. We also plan to investigate the use of other pre-processing techniques and feature engineering methods to enhance the quality and diversity of summaries. Finally, we plan to evaluate the performance of our algorithm on other domains and languages to demonstrate its effectiveness and generalizability.

Overall, our proposed algorithm shows great promise for extractive text summarization for emails and has the potential to be applied in a wide range of applications. We believe that our work will inspire further research in this area and contribute to the development of more effective and efficient summarization techniques.

VI. ACKNOWLEDGMENT

We would like to express our deepest gratitude to Dr. Avinash Kumar Sharma, the Head of the Computer Science Department, for their invaluable support and guidance throughout the course of this research. We would also like to extend our heartfelt appreciation to Ms. Shalu Mall, our esteemed project supervisor, for their unwavering commitment, mentorship, and expertise.

REFERENCES

[1] Biswas, S., Rautray, R., Dash, R., & Dash, R. (2018, September). Text Summarization: A Review. In 2018 2nd International Conference on Data Science and Business Analytics (ICDSBA) (pp. 231-235). IEEE.

[2] Alguliyev, R. M., Aliguliyev, R. M., Isazade, N. R., Abdi, A., & Idris, N. (2019). COSUM: Text summarization based on clustering and optimization. Expert Systems, 36(1), e12340.

[3] Koupaee, M., & Wang, W. Y. (2018). Wikihow: A large scale text summarization dataset. arXiv preprint arXiv:1810.09305.

[4] Liu, Y., Zhu, C., & Zeng, M. (2021). End-to-end segmentation-based news summarization. arXiv preprint arXiv:2110.07850.

[5] Hark, C., & Karcı, A. (2020). Karcı summarization: A simple and effective approach for automatic text summarization using Karcı entropy. Information processing & management, 57(3), 102187.

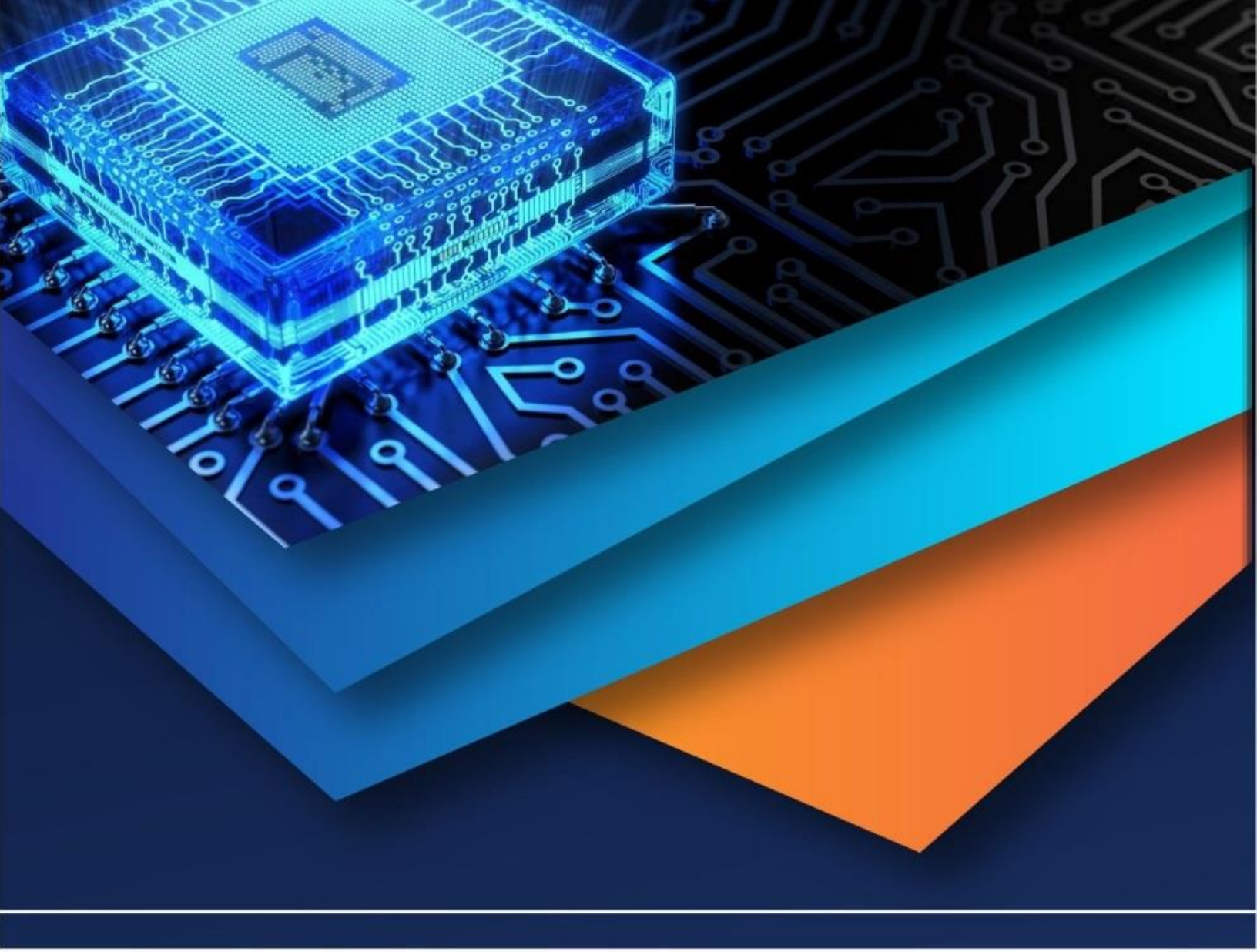
[6] Aries, A., & Hidouci, W. K. (2019). Automatic text summarization: What has been done and what has to be done. arXiv preprint arXiv:1904.00688.

[7] Sarwadnya, V. V., & Sonawane, S. S. (2018, August). Marathi extractive text summarizer using graph based model. In 2018 fourth international conference on computing communication control and automation (ICCUBEA) (pp. 1-6). IEEE.

[8] Moratanch, N., & Chitrakala, S. (2017, January). A survey on extractive text summarization. In 2017 international conference on computer, communication and signal processing (ICCCSP) (pp. 1-6). IEEE.

[9] Krishnaveni, P., & Balasundaram, S. R. (2017, July). Automatic text summarization by local scoring and ranking for improving coherence. In 2017 international conference on computing methodologies and communication (ICCMC) (pp. 59-64). IEEE.

[10] Rani, R., & Lobiyal, D. K. (2021). A weighted word embedding based approach for extractive text summarization. Expert Systems with Applications, 186, 115867



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY