International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

TEXT SUMMARIZATION OF DOCUMENTS

1Afifa Khan, M.Tech (CSE), MuffakhamJha College of Engineering and Technology, Hyderabad

2Mrs. Naimoonisa Begum, Assistant Professor (CSED), MuffakhamJha College of Engineering and Technology, Hyderabad

ABSTRACT: There is a pressing need in the modern era to provide a more effective system for extracting information from the vast amounts of data available on the internet. Summarizing a lengthy piece of text by hand is a difficult task for humans. There is a plethora of written content available online. It is thus difficult to sift through the mountain of available documents in search of the information you need. These two issues can only be addressed effectively through the use of automatic text summarization. Text summary is the process of extracting the most relevant details from a lengthy document or collection of related papers and condensing them into a concise version while preserving their overall meanings. Automatic text summarization using deep learning methods is an active area of study. The majority of previous attempts in extractive text summarization had to rely on manually built features. However, it's a tedious and lengthy process. In this study, we used a data-driven approach to building extractive summaries with the help of deep learning.

1. INTRODUCTION

In the big data age, there has been a meteoric rise in the volume of textual information available from a variety of sources. There is a tremendous deal of data and expertise contained in this massive body of literature, and it needs to be summarized properly before it can be put to good use. Natural language processing (NLP) for automatic text summarization calls for substantial study because of the proliferation of available materials. Automatic text summarizing is the process of creating a condensed and fluent summary of text without the help of a human while retaining the meaning of the original text. There are several natural language processing (NLP) jobs that can benefit from text summarization, such as categorization, question answering, legal text summarization, news summarization, and headline generating.

Summarization that is extracted from larger texts is of most interest to us. The method focuses on getting items from the entire

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

collection in a non-destructive manner. Extractive summarization algorithms take in texts and output a vector of potential outcomes.

A summary is not typically thought of as a string of straight quotations from a work, but it should nevertheless strive to convey the main points of the original. Though not immediately apparent, extractive summaries serve the objective of delivering the most important chunk of material, which can give a good idea of what the book is about and specific sentences that can be cited or referred to for various purposes. To get at the heart of extractive summaries, this work use a technique called paraphrase detection. The initiative also aims to evaluate the efficacy of data-driven methods for extract generation in Indian languages.

This initiative's ultimate goal is to provide readers with a condensed version of an article that contains just the most vital details.

Methods for Summarizing Content Create summaries that present the main points of the material as people would after reading them.

It uses generative methods that can produce natural-sounding sentences without changing the meaning of the source material. As a result of Deep Learning's success, various novel approaches have been proposed to tackle this difficult issue.

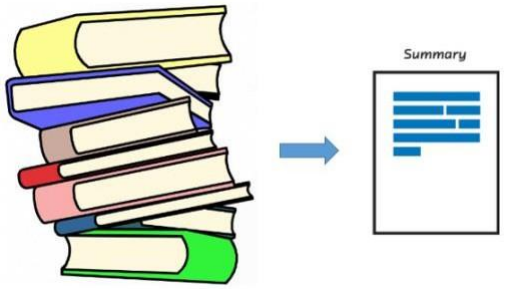

Fig.1: Example figure

To generate summaries, extractive text summarization takes the given material and selects key sentences from it to display to the user. With a score assigned to each sentence, only the highest-scoring ones are included in the final extract. This is easier than abstractive summaries, which require the generation of phrases and words, the organization of those words into intelligible sentences, and the presentation of an interpreted gist of the text. As such, it is a far more challenging task as it would require extensive use of natural language processing. The purpose of this study is to help with text summary by creating extractive summaries with a data-driven approach, made possible by deep learning methods. This entails sorting through the mass of content and producing a list of sentences that are likely to

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

be the most helpful and contain the essential gist of the text.

The most obvious gain from using a summary is the reduction in time spent reading. Two common methods for condensing text are extraction and abstraction. An extractive summarizing method involves selecting key phrases, paragraphs, and other pieces from the original document and stringing them together into a condensed version. This is easier than abstractive summaries, which require the generation of phrases and words, the organization of those words into intelligible sentences, and the presentation of an interpreted gist of the text.

An integral part of any text summarizing project is an analysis of the summary's effectiveness. Both internal and external measures can be used to evaluate it. Summary quality is evaluated in two ways: intrinsically, through human judgment, and extrinsically, through task-based performance measures like the information retrieval-oriented test. Biomedical text summarizing aids information seekers in the biomedical field by shortening documents without changing their meaning. For example, the situation in which the summary is created could be instructive in evaluating its significance. Sentence choice can also be affected by other factors, such as if the document in question is a news article, email, scientific paper, etc.

2. LITERATURE REVIEW

Automatic Text Summarization Model using Seq2Seq Technique

Increasing acquisition of digitization over the information storing and processing in our daily lives has increased the demand of digitization in multiple facets including in investigation processes as well. In fact, for crimes involving computer systems requires the adoption of best practices for the process of evidence extraction from acquired devices from the crime scenes. Over the past years, summarization has become a topic of research. Various techniques of Natural Language Processing (NLP) enabling researchers to generate efficient results for a wide spectrum of documents. In the proposed work Seq2Seq Architecture with RNN is used to perform summarization tasks for documents. The nature of the summary is abstractive and allows the generation of internal meaning by the model itself. With refinement and continual work, this model becomes a strong foundation to perform summarization on longer and legal documents. The results are efficient summary

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

generation and ROUGE scores in the range of 0.6 - 0.7.

Extractive text summarization using sentence ranking

Automatic Text summarization is the technique to identify the most useful and necessary information in a text. It has two approaches 1) Abstractive text summarization and 2) Extractive text summarization. An extractive text summarization means an important information or sentence are extracted from the given text file or original document. In this paper, a novel statistical method to perform an extractive text summarization on single document is demonstrated. The method extraction of sentences, which gives the idea of the input text in a short form, is presented. Sentences are ranked by assigning weights and they are ranked based on their weights. Highly ranked sentences are extracted from the input document so it extracts important sentences which directs to a high-quality summary of the input document and store summary as audio.

Extractive summarization of Telugu documents using TextRank algorithm

Reading large and lengthy documents is a tedious and time-consuming task. A summary of the same document gives us an overall idea of what the document is all about. Automated summaries can be generated using various algorithms. The summary can be generated for single or multiple document inputs. Here the proposed system carries out extractive summarization of multiple Telugu text documents. The algorithm applied here is text rank algorithm.

Extractive Text Summarization from Web pages using Selenium and TF-IDF algorithm

To obtain an overview of the content present in numerous documents, is a time-consuming task. Similarly, searching for specific information online, from multiple websites and webpages is a monotonous task. To avoid this, automatic text summarization is one of the most widely adopted techniques today to get a concise and brief outline of the information. In this paper, a novel process is proposed to generate an extractive summary of the information based on the user's query by extracting data from multiple websites over the internet. Web-scraping through Selenium is also discussed. The Term Frequency-Inverse Document Frequency (TF-IDF) algorithm is applied for text

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

summarization. The proposed approach is unique and efficient for generating summaries as per the user's request.

Extractive Summarization of Text from Images

The modern age has brought with it an abundance in the inflow of information. The consumption of such large chunks of information leads to a latency in the gathering of relevant information. The condensation of such information becomes necessary as the volume of this inflow keeps expanding. The most efficient way for the retrieval of the most important contents of the data is to summarize the data such that only non-redundant and useful information is contained in it. The manual summarization of textual content is a laborious task and is not very efficient. Therefore, automatic summarization is of utmost importance. Text summarization is the process of identifying the most relevant information and discarding the unnecessary and the irrelevant information. In this paper, textual content is extracted from images using optical character recognition and the extracted text is subjected to various extractive summarization algorithms. The summary is evaluated with respect to the summary generated by an abstractive summarization model using Rouge-N and Rogue-L metric to calculate the Precision, Recall and F-Score.

Text Document Summarization Using POS tagging for Kannada Text Documents

Humongous amount of data available on the world wide web has been a constant issue pertaining to better Information Retrieval (IR) techniques. Text document summarization is there around for the past several decades, but providing a succinct summary has been challenging as ever. This work focuses on extractive summarization techniques using POS tagging, where the goal is to tag individual words with its parts of speech in a document and do the extractive summarization with more grammatical meaning. The Hidden Markov Model (HMM) is used for tagging the dataset. The idea is to use sentence ranking to produce the summary of a given document. This method of summarization uses the key phrase extraction, where the goal is to select individual words or sentences to tag a document to create text document summary.

Text summarization using text frequency ranking sentence prediction:

In the era of information technology, data plays significant role. The data which prevails on the internet are unstructured and

are not in a concise manner. To make the raw data into a structured, readable, coherent and concise and to extract the summary of data, the text summarization concept is introduced. The text summarization involves in providing asummary of the useful information from the raw data without dissolving the main theme of the data. Nowadays readers face the challenge of reading comments, reviews, news articles, blogs, etc., as they are too informal and noisy. Retrieving the correct gist of the text which is necessary for all the readers is a quite difficult task. In order to overcome the problems faced by the readers, TFRSP (Text Frequency Ranking Sentence Prediction) algorithmis proposedto generate a precise summary that uses supervised and unsupervised learning algorithms. The proposed approach uses the combination of TF-IDF-TR (Term Frequency – Inverse Document Frequency – Text Rank) as an unsupervised learning algorithm and Seq2Seq (Sequence to Sequence) model as a supervised learning algorithm to obtain the benefits of both extractive and abstractive summarization. The results of the proposed TFRSPapproach is compared with the existing methods of text summarization using the Recall Oriented Understudy of Gisting Evaluation (ROUGE) and attains a high ROUGE score, hence achieves high accuracy ofsummary.

An Adaptive Normalized Google Distance Similarity Measure for Extractive Text Summarization:

No doubt, for each clustering algorithm running improper similarity calculation, that can lead to reduce the clustering accuracy. Hence, several applications that employ such algorithm are affected negatively and generate improper results. In a previous work, we found that employing the Normalized Google Distance (NGD) similarity measure to cluster document's sentences for text summarization problem is unreasonable; since NGD was basically designed to work with large databases. On the other hand, a term-weighting approach is used widely to define document's contents. In this paper, a term-weighting approach is integrated with the NGD similarity measure to adopt the latter from being able to work in small database (single document). Differential Evolution (DE) algorithm is used to train and test the proposed method. The DUC2002 dataset is preprocessed and used as a test bed. The results showed that our proposed method could outperform the previous work in terms of F-score evaluation measure as well as outperformed the standard

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

baseline methods Microsoft Word and Copernic Summarizer.

3. IMPLEMENTATION

The first step in implementing machine learning-based methods was to train simple classifiers, such as naive-bayes classifiers, decision trees, clustering, and hidden markov models, using a feature vector that was manually crafted using the aforementioned parameters. Genetic algorithms, which describe optimization issues and solve them with strategies typically observed in nature in natural selection procedures like mutation, cloning, and cross-overs, have also been examined in some of this work.

Weaknesses:

Machine learning algorithms that simulate optimization problems and find solutions employing methods observed in nature.

In this project, we used deep learning to automatically generate data-driven extraction summaries. The proposed approach uses paraphrase strategies to evaluate whether a sentence should be included in a summary.

Benefits:

As a result of the momentum that Deep Learning has brought, several novel solutions have been presented for what was previously considered an intractable problem.

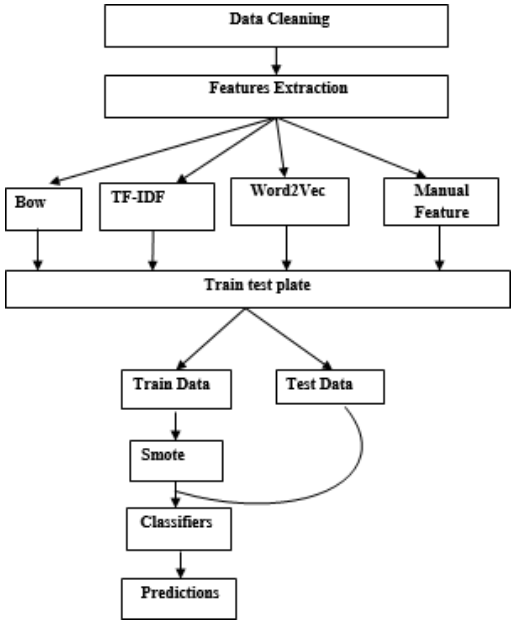

Fig.2: System architecture

MODULES:

- Import libraries
- Data exploration
- Data processing
- Feature selection
- Data splitting
- Tokenization
- Model generation
- Build LSTM classifier
- Accuracy comparison
- Model build
- Create flask object
- Load model

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

- Sighnup & signin
- Upload test reviews
- Predict text summarization result

CHARACTERISTIC OF TEXT SUMMARIZATION:

- Words, phrases, and sentences are singled out and used to create a summary through text summarizing techniques.
- Unlike single-document summarizing, which can only take in a single source document, multi-document summary can take in multiple documents so long as they are all connected to the same topic.
- In order to develop a summary, relevant information is taken from each document and then compiled and structured.
- Feature engineering-based models have proven to be significantly more effective when considering domain- or genre-specific summarization (such as medical reports or certain newspaper articles), as classifiers may be trained to recognize specific types of information.
- To do a good job of summarizing a work of literature, we often read the whole thing first so that we can understand it better, and then we write a summary highlighting the key themes. Automatic text summarization is a labor-intensive process because computers can't fully mimic human language and comprehension.

4. ALGORITHMS

Predicting and recognizing patterns in order to provide acceptable results after understanding them is what machine learning is all about. Algorithms using ML seek out and learn from data patterns. An ML model will learn and get better with each new attempt. It is necessary to split the data into training and test sets before evaluating the performance of a model. To train our models, we split the data into two sets, with the former making up 70% of the dataset and the latter making up 30%. The results from our model were to be evaluated using a variety of quality indicators.

The performance of the designed system has been verified by testing with a test set. Evolution analysis refers to the study and modeling of recurring patterns or trends in the behavior of objects across time. The confusion matrix is used to measure precision and accuracy. Building a predictive model

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

with a regular LSTM model is the most crucial step.

RNN:

The recurrent neural network (RNN) architecture known as long short-term memory (LSTM) is used in deep learning. LSTM differs from standard feedforward neural networks because it has feedback connections. Data sequences, not just individual data points (like photos), can be processed (such as speech or video). Unsegmented, connected handwriting recognition, speech recognition, anomaly detection in network traffic, and intrusion detection are only some of the applications of LSTM (intrusion detection systems).

LSTM:

LSTM networks are well-suited to categorizing, processing, and making predictions based on time series data since there might be lags of uncertain duration between critical occurrences in a time series. The vanishing gradient problem is one that arises during training standard RNNs, and LSTMs were developed to solve this issue. Compared to RNNs, hidden Markov models, and other sequence learning techniques, LSTM is advantageous in many contexts because it is less sensitive to gap length.

An LSTM-based RNN can be trained supervised on a set of training sequences by employing an optimization algorithm like gradient descent in conjunction with backpropagation through time to compute the gradients required during the optimization process, thereby adjusting the values of each LSTM unit's weights in proportion to the derivative of the error (at the LSTM network's output layer) with respect to the corresponding weight.

As the delay between significant events grows longer, the error gradients produced by gradient descent for typical RNNs gradually disappear. Back-propagation of errors from the output layer to the LSTM units stores the error in the LSTM cell. Each gate in the LSTM unit receives feedback from the "error carousel" until it learns to stop accepting the error value.

TEXT RANK ALGORITHM:

- For the purpose of automated summarization, Text Rank represents any document as a network in which sentences serve as nodes.
- In order to construct intermediate edges, a function that calculates sentence similarity is necessary. The similarity between two phrases

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

determines the importance of the edge connecting them in the graph, hence this function is used to provide a weight to each edge in the graph.

- In order to determine how similar two sentences are, Text Rank looks at the words and phrases they include in common. This overlap is simply estimated as the number of shared lexical tokens between them divided by the total length of each phrase to avoid highlighting excessively long ones.
- First, we'd string together all the articles' text, and then we'd break it up into sentences.
  - Locating vector representations (word embeddings) for each phrase will be the next stage.
  - A matrix is then used to store the results of a calculation on the similarity of text vectors.
  - The similarity matrix is then transformed into a directed graph, with sentences serving as vertices and similarity scores as edges, and the result is used to determine the ranking of sentences.
  - The concluding summary is made out of a certain number of carefully selected sentences.

5. EXPERIMENTAL RESULTS

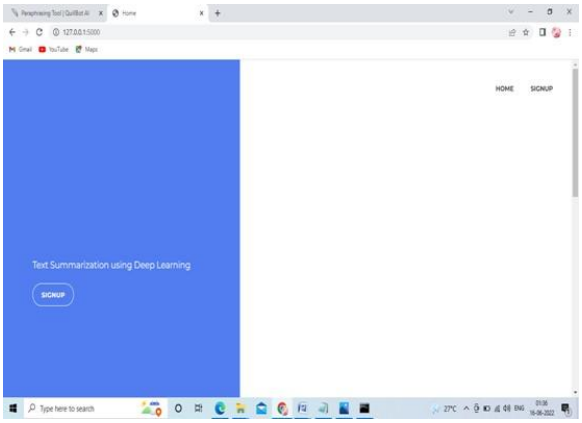

Fig.3: Home screen

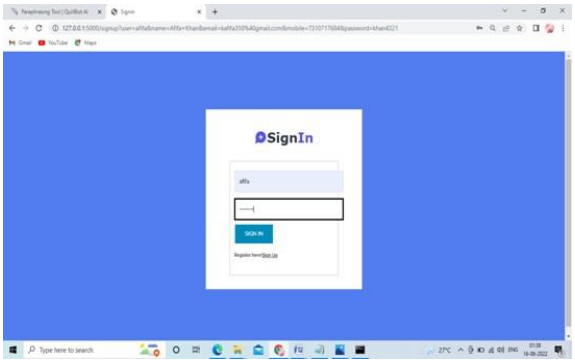

Fig.4: User login

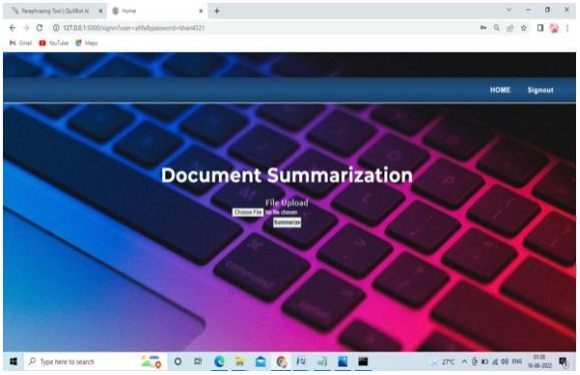

Fig.5: Main screen

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

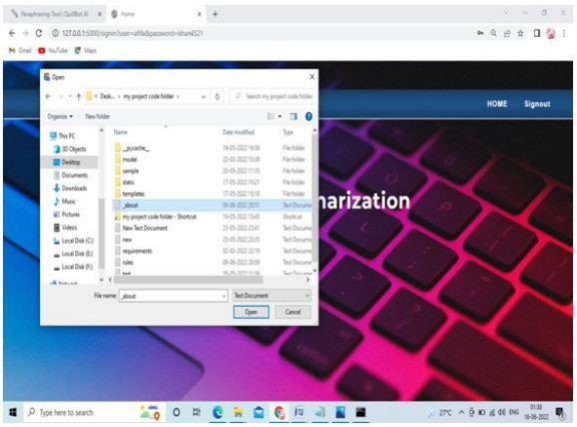

Fig.6: Input uploading

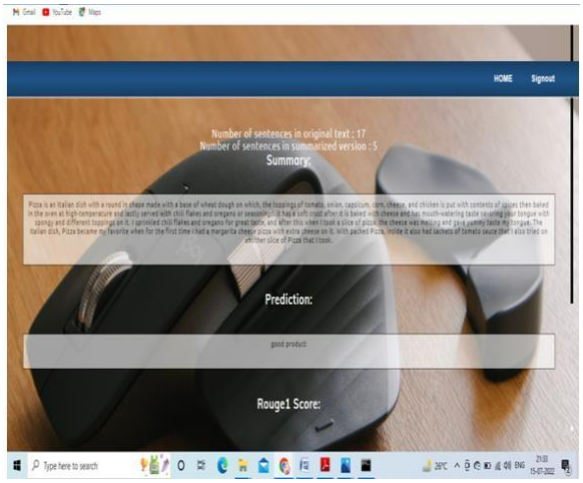

Fig.7: Prediction result

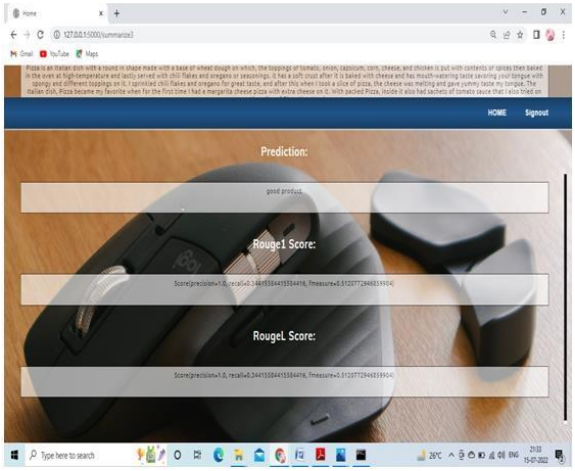

Fig.7: Prediction result

6. CONCLUSION

There have been issues with humans as a result of the explosion in the volume of written texts. Due to public and business requirements, humans have very little time to gather the essential information from various resources, so the system built here makes the duties easier by efficiently summarizing the documents. It's a technique for computingly condensing the text of a document so as to get at the most crucial information in the least amount of time. Extending the scope of feature extraction from the sentence level to higher abstractions like paragraphs, and in the case of multi-document summarization, to documents, is one potential future direction for the approach. An technique based on Recurrent Neural Networks and associated attention mechanisms can be zeroed in on. Given that vanilla RNNs and their variants, such as LSTMs, have been shown to be good models in situations where we require some sort of memory, using them in this kind of approach could be very useful because they would aid in classifying a sentence to belong to a summary by taking into account the contributions from a few previous sentences as well. They are able to select which inputs to prioritize by employing attention processes in tandem with gated memory cells. A useful evaluation criteria for

the abstracts produced by the system Human scoring is a potential next step, and while contributions to the project are always appreciated, those dealing with data set production and gathering are especially encouraged. This data-driven method of education necessitates a sizeable enough to implement it.

REFERENCES

1. C. Prasad, J. S. Kallimani, D. Harekal and N. Sharma, "Automatic Text Summarization Model using Seq2Seq Technique," 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2020, (pp. 599-604), IEEE.
2. Madhuri JN, Kumar RG. "Extractive text summarization using sentence ranking". In2019 International Conference on Data Science and Communication (IconDSC) 2019 Mar 1 (pp. 1-3). IEEE.
3. Manjari KU. "Extractive summarization of Telugu documents using TextRank algorithm". In2020 Fourth international conference on I-SMAC (IoT in social, mobile, analytics and cloud)(I-SMAC) 2020 Oct 7 (pp. 678-683). IEEE.
4. Manjari KU, Rousha S, Sumanth D, Devi JS." Extractive Text Summarization from Web pages using Selenium and TF-IDF algorithm". In2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184) 2020 Jun 15 (pp. 648-652). IEEE.
5. Kolle R, Sanjana S, Meleet M. "Extractive Summarization of Text from Images". In2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES) 2021 Sep 24 (pp. 1-4). IEEE.
6. Jayashree R, Anami BS, Poornima BK. "Text Document Summarization Using POS tagging for Kannada Text Documents". In2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence) 2021 Jan 28 (pp. 423-426). IEEE.
7. Meena SM, Ramkumar MP, Asmitha RE, G SR ES. "Text summarization using text frequency ranking sentence prediction". In2020 4th International Conference on Computer, Communication and Signal Processing (ICCCSP) 2020 Sep 28 (pp. 1-5). IEEE.
8. Abuobieda A, Osman AH. "An Adaptive Normalized Google Distance Similarity Measure for Extractive Text Summarization". In2020 2nd International Conference on Computer and Information Sciences (ICCIS) 2020 Oct 13 (pp. 1-4). IEEE.

9. Yadav D, Desai J, Yadav AK. "Automatic Text Summarization Methods: A Comprehensive Review". arXiv preprint arXiv:2204.01849. 2022 Mar 3.
10. Ghodratnama S, Beheshti A, Zakershahrak M, Sobhanmanesh F. "Extractive document summarization based on dynamic feature space mapping". IEEE Access. 2020 Jul 28;8:139084-95.