iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VIII **Month of publication:** August 2022

DOI: https://doi.org/10.22214/ijraset.2022.46126

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 10 Issue VIII Aug 2022- Available at www.ijraset.com

Abstractive Document Summarization Using Divide and Conquer Approach

Pranav Sharma[1], Nitesh Sharma[2], Manoj Sakat[3], Vaibhav Shinde[4]
[1, 2, 3, 4]*Department of Computer Engineering, AISSMS Institute of Information Technology, Pune, India*

Abstract: Document summarising is a multi-step process with numerous subtasks. Each subtask has the ability to provide high-quality summaries. Identifying necessary paragraphs from the given document is a vital aspect of abstractive document summarization. In this paper, we describe an abstractive text summary strategy based on a statistically innovative sentence rating technique, with the summarizer selecting the sentences. The abstraction sentences are generated as a summary text and then transformed to audio. In terms of accuracy, the proposed model outperforms the standard technique.
Keywords: Summarization, Machine Learning, Classification, NLP (Natural Language Processing) etc.

.

I. INTRODUCTION

The purpose of this research is to reproduce and extend knowledge on automatic text summarization. The success of sequence-to-sequence (seq2seq) modelling in NLP has spurred rapid and significant growth in this field. However, the focus of research has been on standards and specifications for quick, single-document summarization rather than prolonged and multi-document summarization. The ultimate goal of this topic is to provide a framework for writing high-quality summaries regardless of the length of the source documents, whether it is a single or multi-document assignment, and regardless of domain or language. Given the recent excellent success in short-document summarization, the next goal will be to reproduce the results in lengthier resources. The goal of this thesis is to add to the body of knowledge in the field of long document summary. The evaluation of summaries and methods for abstractive document summarising are two especially significant subjects that we identify as critical problems in this thesis [4]

A. Problem Statement

Our technique breaks down the problem into smaller summary tasks by using the document's discourse structure and analysing sentence similarity. A huge text and its summary are separated into Several source-target pairs are then used to train a model that learns to summarise each section of the document individually. The component summaries are then integrated to form a thorough final overview.

B. Motivation

Data compression and information understanding, both of which are essential for information science and retrieval, are closely related to summarization. The ability to generate engaging and well-written document summaries has the potential to boost the success of both information discovery algorithms and human users looking to swiftly scan large amounts of paper for important information. Despite being one of the least solved natural language processing (NLP) challenges, automatic summarization has recently been recognised as one of the most significant.

II. LITERATURE SURVEY

Dani Gunawan, Siti Hazizah Harahap, Romi Fadillah Rahmat "Multi-document Summarization by using TextRank and Maximal Marginal Relevance for Text in Bahasa Indonesia" [1] According to simulated data from a ray tracing tool for use in specific metropolitan contexts in the 28-GHz band, using random 1 base station (BS) angles in a directional search operation may not even be ideal for users in non-line-of-sight scenarios. The frequency of the angle used and the intensity of the current street canyon propagation appear to be reliant on performance improvement. To discover the most effective BS angles that avoid power emissions, a simple ray-tracing technique is offered.

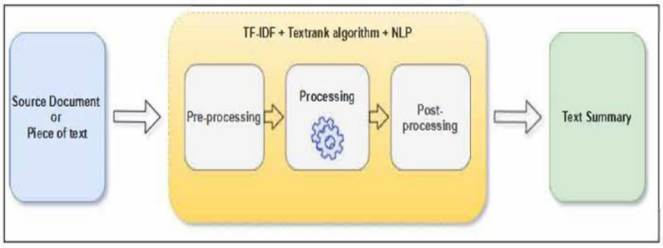
S huxia R en,K a ijie G uo," Text Summarization Model of Combining Global Gated Unit and Copy Mechanism" [2] Text summarization is a common NLP activity. Text summarization decreases document length automatically. Attention seq2seq neural networks provide summarization, according to a recent study. However, guaranteeing that the summary is accurate is tough. Furthermore, OOV has a significant impact on the quality of the final summary. We created a text summarization model with a global gated unit and a copy mechanism to address these concerns (GGUC). The model outperforms existing text summary algorithms on LCSTS datasets.

P.Krishnaveni, Dr.S. R. Balasundaram ”Automatic Text Summarization by Local Scoring and Ranking for Improving Coherence” [3], Text summarizers are a common automated software tool for processing massive amounts of web data. Its purpose is to produce a shorter version of the given text. without losing the source text's overall information with the use of this software programme, people can obtain a better knowledge. a big number of texts in a short amount of time and deciding whether or not to read the complete document The ATS approach should address the challenges of choosing relevant text excerpts and writing cohesive summaries. Automated summaries differ in general. more than summaries created by humans as a result of In humans, reasoning Text summarization can be divided into two categories. A process of abstractive summarization After thoroughly understanding the original text, the summary text is generated by rephrasing it. The summary is created through an extractive summarization process.

J.N. Madhuri, Ganesh Kumar. R “Extractive Text Summarization Using Sentence Ranking”,[4] The goal of this paper is to show how to use a novel statistical method to derive text summary from a single document. A sentence extraction method is provided, which provides a concise representation of the supplied text's notion. Weights are assigned to phrases, and the values of the sentences are utilised to rank them. The extracted essential sentences from the input document are led to a high-quality summary of the input document, which is preserved as audio.

Meena S M, Ramkumar M P “Text Summarization Using Text Frequency Ranking Sentence Prediction” [5] In the age of information technology, data is extremely valuable. The data that is widely available on the internet is disorganised and incoherent. The notion of text summarising is provided to turn raw data into a structured, accessible, coherent, and succinct manner, as well as to extract data summaries. Text summarization is the process of extracting relevant information from the raw data without diluting the material's main theme. Readers today face a challenge when it comes to reading comments, reviews, news pieces, blogs, and other forms of informal and noisy communication. It is difficult to find the exact gist of the text, which is required by all readers. To address the problems that the readers are having

III.SYSTEM ARCHITECTURE



.Fig. 2.1. System Architecture

IV.ALGORITHM/ METHODOLOGY

To make the statement, the words are taken completely out of context and slightly modified. Machine learning algorithms are frequently employed in NLP algorithms.

Pluralism Machine learning may be used in NLP to automatically learn these rules by studying a group of occurrences (i.e., reducing a large quantity to a collection of phrases) and drawing a statistical conclusion. (NLP) is used in textual categorization to analyse text and assign predefined tags or categories based on its content. Key words and phrases from the original text are extracted and combined to form the summary.[1][3]

1) Natural language processing (NLP) is a linguistics branch that combines linguistics, computer science, and artificial intelligence to research computer-human language interactions, particularly how to construct machines that can process and analyse large amounts of natural language data. The ultimate goal is for a computer to be able to interpret speech. These technologies include Natural Language Processing and Machine Learning.[3][4][6]

2) Machine Learning and Natural Language Processing are two key subfields of Artificial Intelligence that have recently gained attention. Computer learning and natural language processing are critical components of intelligent machine development. On the other side, capability of a computer system to interpret languages as a result of technical breakthroughs. A computer can only comprehend 0s and 1s; it is incapable of comprehending human languages such as English or Hindi. The computer system was able to comprehend English and Hindi using Natural Language Processing.

A. Document Summarization steps

1) Gather Information.
2) Text Preparation.
3) Transform paragraphs into sentences.
4) Sentence tokenization
5) Determine the weighted frequency of occurrence.
6) In sentences, replace words based on their weighted frequency.
7) Sort sentences in descending weight order.
8) Recapitulating the Article

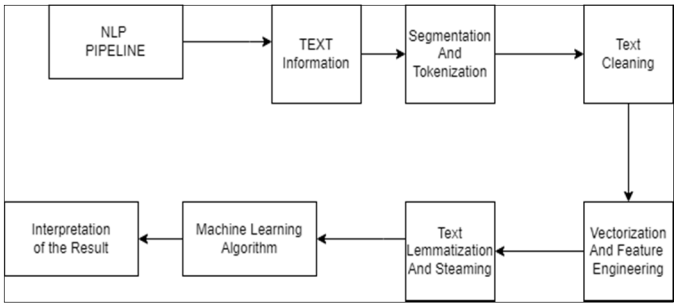


Fig 4.1

To evaluate the similarity of two string values, the Edit Distance technique analyses the number of operations required to alter one value to another (word, word form, word composition).

B. As A Result, The Following Text Operations Are Included In This Method

In a string, a character is inserted. A technique for removing (or replacing) a character from a string is character substitution.
Popular NLP applications for Edit distance include automatic spell-checking (correction) systems, bioinformatics - assessing the similarity of DNA sequences (letters view), and text processing - defining all the proximity of words that are near certain text objects.

One operation - Edit Distance: 1
Str1= “string”, str2 = “strong”

Several operations – Edit distance: 3
Str1 = “Sunday”, str2= “Saturday”

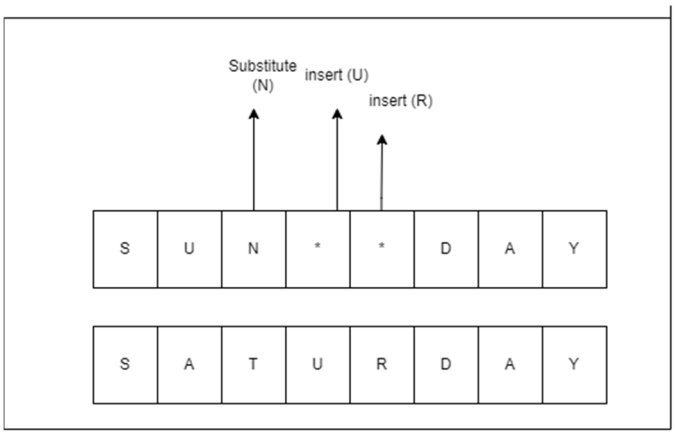


Fig. 4.2

C. Similarity in Cosine

Cosine similarity is a statistic for comparing text in different texts. This metric is calculated using the cosine vectors formula, which is based on vector similarity measures:

$$\text{COS}\Theta = \frac{\bar{A} \cdot \bar{B}}{||\bar{A}||.||\bar{B}||}$$

Text vectorization techniques, for example, can be used to express this text as a vector utilizing a variety of text attributes or qualities. The cosine similarity computes the differences between such vectors for three terms, as seen on the vector space model below.

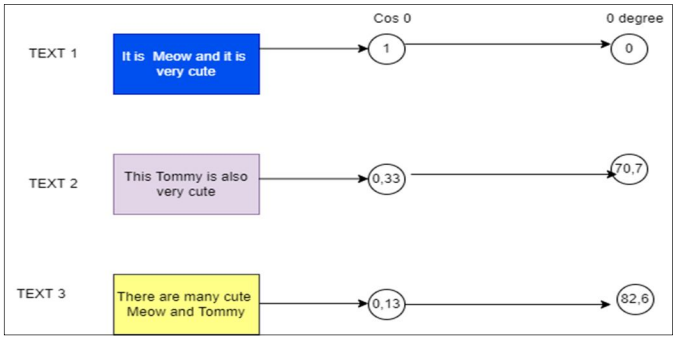


Fig. 4.3

The easiest and most natural method for vectorizing text data is as follows:

If you want a dictionary that only contains the terms that appear in the dictionary, you can create an index for each one and then count the number of times that word appears. In the end, we obtain a vector with an index value and frequency of occurrence for every word in the text

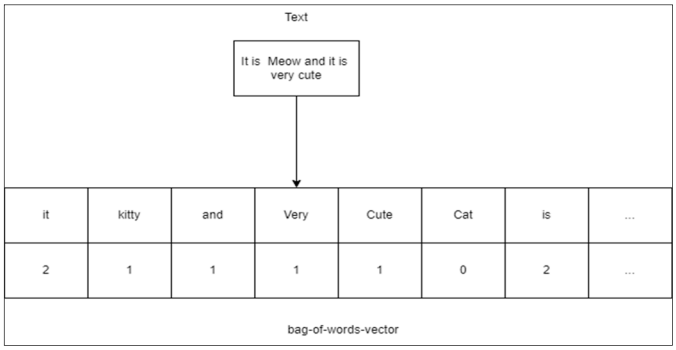


Fig. 4.4

D. TF-IDF

TF-IDF, which stands for term frequency and inverse document frequency, is one of the most widely used and successful NLP strategies. This method allows you to compare the importance of a phrase (or a group of words) to all other phrases in a document.

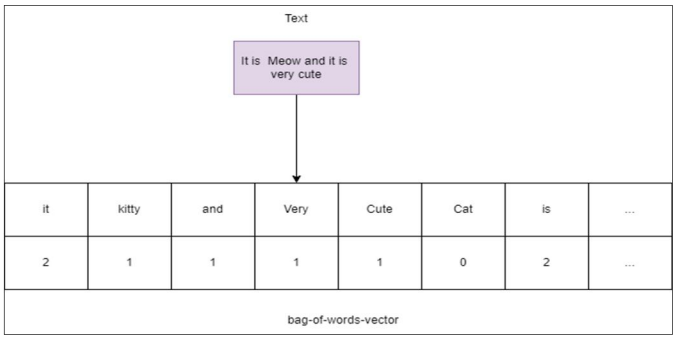


Fig. 4.5

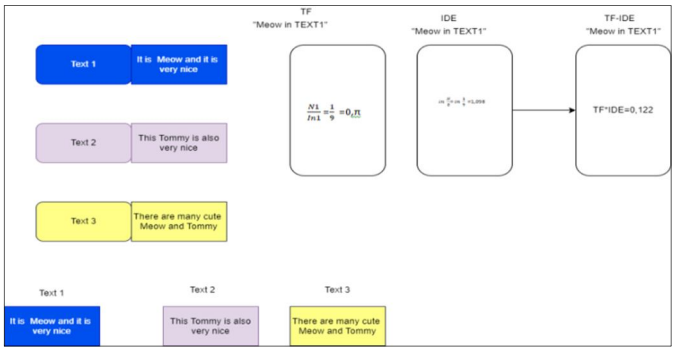


Fig. 4.6

V. RESULT

Here, we have a few input datasets, which are considered as our input, courrently input should be in .txt format. After providing some input, our algorithm starts working, it takes a few seconds and output is generated. As in the following image, we output is a short summary of given input.

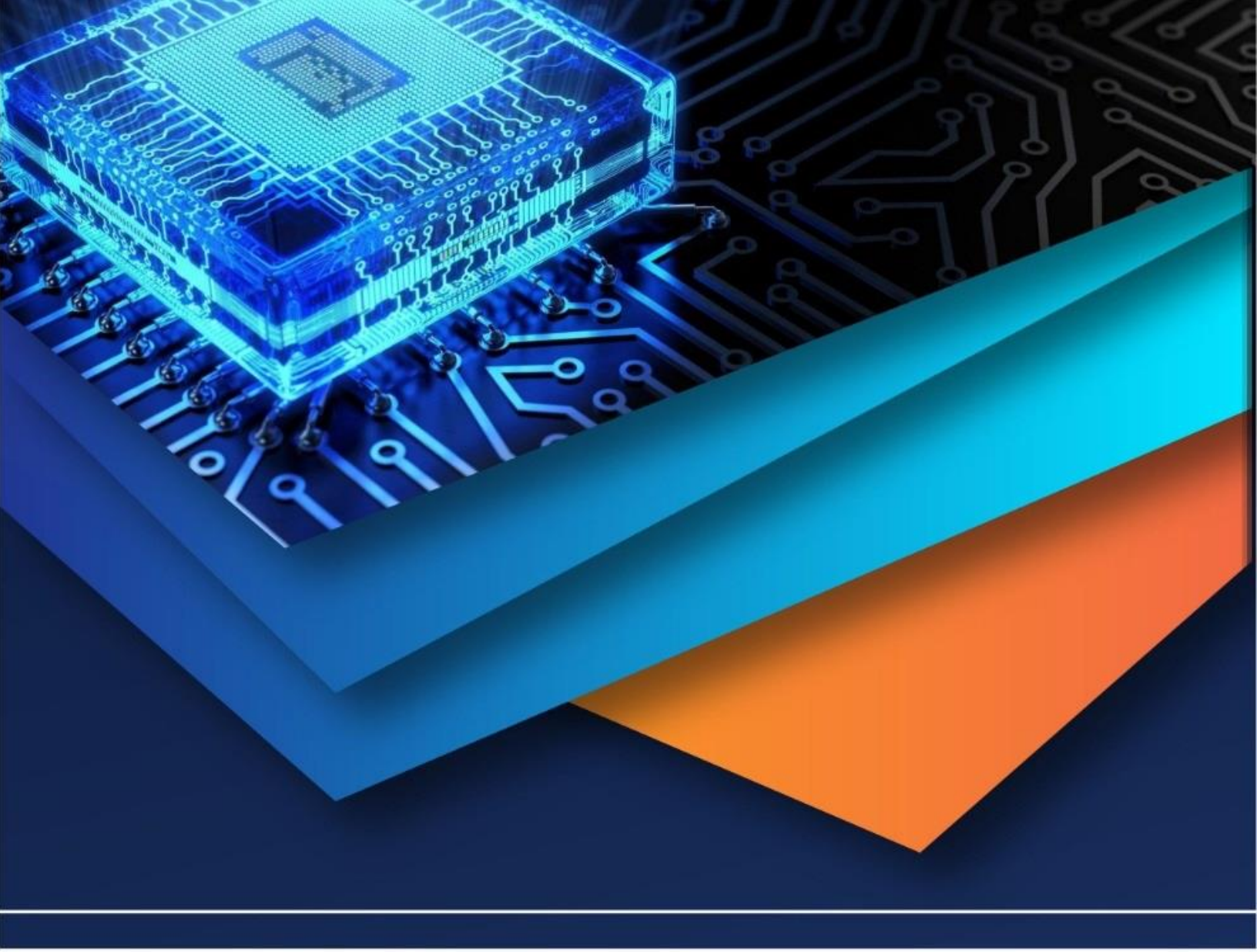
Fig. 5.1

VI.CONCLUSION

In the age of information technology, data is extremely valuable. The data that is widely available on the internet is disorganized and incoherent. The notion of text summarizing is provided to turn raw data into a structured, accessible, coherent, and succinct manner, as well as to extract data summaries. Text summarization is the process of extracting relevant information from the raw data without diluting the material's main theme. Readers today face a challenge when it comes to reading comments, reviews, news pieces, blogs, and other forms of informal and noisy communication. It is difficult to find the exact gist of the text, which is required by all readers. To address the problems that the readers are having,

REFERENCES

[1] Dani Gunawan, Siti Hazizah Harahap, Romi Fadillah Rahmat “Multi-document Summarization by using TextRank and Maximal Marginal Relevance for Text in Bahasa Indonesia” [2020] DOI: 10.1109/ICISS48059.2019.8969785
[2] S huxia R en,K a ijie G uo,” Text Summarization Model of Combining Global Gated Unit and Copy Mechanism” [2019] DOI: 10.1109/ICSESS47205.2019.9040794
[3] P.Krishnaveni, Dr.S. R. Balasundaram ”Automatic Text Summarization by Local Scoring and Ranking for Improving Coherence”[2018] DOI: 10.1109/ICCMC.2017.8282539
[4] J.N. Madhuri, Ganesh Kumar.R” Extractive Text Summarization Using Sentence Ranking” [2019], DOI: 10.1109/IconDSC.2019.8817040
[5] Meena S M, Ramkumar M P “Text Summarization Using Text Frequency Ranking Sentence Prediction” [2020], DOI:10.1109/ICCCSP49186.2020.9315203
[6] E. Sandhaus, “The New York Times annotated corpus.” Linguistic Data Consortium, Philadelphia 6.12, 2008, Art. no. e26752.
[7] C. Napoles, M. Gormley, and B. Van Durme, “Annotated gigaword,” in Proc. Joint Workshop Autom. Knowl. Base Construction Web-Scale Knowl. Extraction, 2012, pp. 95–100.
[8] M. Grusky, M. Naaman, and Y. Artzi, “Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguist.: Human Lang. Technol., 2018, pp. 708–719.
[9] A. Cohan et al., “A discourse-aware attention model for abstractive summarization of long documents,” in Proc. Conf. North Amer. Chapter Assoc.Comput. Linguist.: Human Lang. Technol., 2018, pp. 615–621.
[10] S. Subramanian, R. Li, J. Pilault, and C. Pal, “On extractive and abstractive neural document summarization with transformer language models,” in Proc. 2020 Conf. Empirical Methods Natural Lang. Process., 2019, pp. 9308–9319.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY