IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 12, December 2022

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.118

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

|| Volume 11, Issue 12, December 2022 ||

| DOI:10.15680/IJIRSET.2022.1112015 |

Abstractive Document Summarization Using Machine Learning Approach

Anuja Parve [1], Purushottam Nimje [2], Mayur Kardile[3], Pratik Bachche[4], Kanchan Pradhan[5]

U.G. Students, Department of Computer Engineering, JSPM's Bhivrabai Sawant Institue of Technology and Research, Wagholi, Pune, Maharashtra, India[1,2,3,4]

Associate Professor, Department of Computer Engineering, JSPM's Bhivrabai Sawant Institue of Technology and Research, Wagholi, Pune, Maharashtra, India[5]

ABSTRACT: We present a novel divide-and-conquer strategy for the neural summary of large documents. Our solution makes use of the document's discourse structure and sentence similarity. Divide the task into smaller summarization challenges. We divide a lengthy document and its summary into numerous source-target pairs, which are then used to train a model. learns to sum up each section of the document individually After The partial summaries are then combined to form a final comprehensive summary. In contrast to the conventional approach, we may break down the challenge of long document summaries into smaller pieces and simpler problems, reducing computational complexity and increasing the number of training examples while also reducing There is noise in the target summaries. There are two publicly available datasets.

KEYWORDS: Complexity, Contrast, Summarization, Machine Learning, Classification, NLP(Natural Language Processing).

I. INTRODUCTION

In this present era, where huge quantity of information is generating on the internet day by day. So, it is necessary to provide a better mechanism to extract useful information fast and effectively. Text summarization is one of the methods of identifying the important meaningful information in a document or set of related documents and compressing them into a shorter version preserving its overall meaning. It reduces the time required for reading whole documents and also it is a space problem that is needed for storing a large amount of data. The automatic text summarization problem has two sub-problems that is a single document and multiple documents. In the single document, the single document is taken as the input and summarized information is extracted from that particular single document. In Multiple documents, the multiple documents of the single topic are taken as input and the output which is generated should be related to that topic.

Automatic Text summarization has two approaches 1) Abstractive text summarization and 2) Extractive text summarization. An extractive text summarization means important information or sentence are extracted from the given text file or original document. An extractive text summarization approach uses linguistic or statistical features for selecting useful informative sentences. An Abstractive text summarization will try to understand the input file or original file and re-generate the output in a few words by identifying the main concept of the input file. In many research papers, they have mentioned that extractive text summarization is sentence ranking. Extractive text summarization is divided into two phases: 1) Pre-processing and 2) Processing. In this paper, we are explaining extractive text summarization on a single document.

This thesis aims to replicate and extend knowledge of automatic text summarizing. The success of sequence-to-sequence (seq2seq) modelling in NLP has pushed rapid and significant growth in this field. However, the focus of the study has been on developing frameworks for short, single-document summarizing, rather than extended and multi-document summarization. The ultimate goal of this topic is to provide a framework that can produce high-quality summaries regardless of the length of the source documents, whether it's a single or multi-document assignment and irrespective of domain or language.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

|| Volume 11, Issue 12, December 2022 ||

| DOI:10.15680/IJIRSET.2022.1112015 |

Given the recent impressive progress in a short document summarizing, the next challenge will be to reproduce the results using longer documents. The goal of this thesis is to contribute to the knowledge of expertise in the field of long document summaries. The evaluation of summaries and systems for Abstractive document summarizing are two particularly relevant areas that we identify as core issues in this thesis.

II. RELATED WORK

1.Paper name: Text Summarization using Sentiment Analysis for DUC Data.
Author: Nidhika Yadav, Niladri Chatterjee
Abstract: The basic purpose of this study is to see if sentiments can be used to summarize material effectively. We describe a text summarization technique based on the emotions associated with key phrases that is computationally efficient. Sentiment analysis is already in use for large-scale text data interpretation and opinion mining in a variety of fields. The current research shows that sentiment analysis can be utilized to summarize text as well. Our findings were compared to those of the conventional DUC2002 datasets.A text summary is a technique for determining a text's main points. In the literature, various strategies have been presented, and many of these are currently in use in commercial systems
The work in this paper is divided into two stages. 1) Text- Detection 2) Inpainting. Text detection is done by applying morphological open-close and close-open filters and combining the images. Thereafter, the gradient is applied to detect the edges followed by thresholding and morphological dilation, erosion operation. Then, connected component labelling is performed to label each object separately. Finally, the set of selection criteria is applied to filter out non-text regions. After text detection, text inpainting is accomplished by using exemplar based Inpainting algorithm.

2. Paper name: Multi-document Summarization by using Text Rank and Maximal Marginal Relevance for Text in Bahasa Indonesia.
Author: Multi-document Summarization by using Text Rank and MaximalMarginal Relevance for Text in Bahasa Indonesia
Abstract: The text summarizer eliminates unnecessary information by selecting the most important sentences. There's a chance that two or more crucial statements in a multi-document summary will contain the same information. If those sentences are included in the summary result, the information will be redundant. The goal of this research is to condense similar lines from a variety of sources containing similar information into a more concise text summary. This study divides a variety of internet news articles into six groups to achieve its goal. The articles are concatenated and pre-processed to create clean text. Following the acquisition of clean text, the Text Rank algorithm is used to extract the important sentences based on similarity measurement. This procedure produced the summary text.

3. Paper name: Automatic Text Summarization by Local Scoring and Ranking for Improving Coherence
Author: P. Krishnaveni, Dr.S. R. Balasundaram.
Abstract: Due to a large amount of textual material available on the internet, machine-generated summarization has received a lot of attention. Manually summarizing these online text documents is difficult for most people. As a result, we will require an automatic text summarizer. Automatic Text Summarization (ATS) is defined as "shortening the original text while preserving its information value and overall meaning." Despite the fact that automatic text summarization has been studied since the 1950s, more cohesive and meaningful summaries still exist. The proposed technique generates an automatic feature-based extractive heading-wise text summarizer that im- proves the summary text's coherence and, as a result, its understandability. It just summarizes the input document using local scoring and ranking, and that's it.

4. Paper name: A Divide-and-Conquer Approach to the Summarization of Long Documents.
Author: Alexios Gidiotis and Grigorios Tsoumakas
Abstract: We propose a novel divide-and-conquer strategy for the neural summarization of large documents. Our solution takes advantage of the document's discourse structure and divides the challenge into smaller summarization problems using sentence similarity. We divide a lengthy text and its summary into numerous source-target pairs, which are then used to train a model that learns to summarize each section of the document separately. The partial summaries are then combined to create a final full summary. Using this strategy, we can divide the challenge of long document summarization into smaller and simpler problems, reducing computational complexity and increasing the number of training examples while reducing noise in the target summaries.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

|| Volume 11, Issue 12, December 2022 ||

| DOI:10.15680/IJIRSET.2022.1112015 |

III. METHODOLOGY

1. Project Scope

The rapid growth of the field of information retrieval is a result of the world's digitization. To stay informed, people use a range of resources. People like information that is brief and to the point. There is a huge issue with reading news stories and internet evaluations or feedback that makes it difficult to form conclusions unless they are read entirely. As a result, the concept of text summarization emerges as a means of improving information retrieval.

2. User Class and Characteristics

This is useful to generate document summaries.

3. Assumptions and dependencies

Assumptions:
We have used Python Technique. Input as Text as a paragraph.
Dependencies:
We have used python libraries. Output to generate document summary.

4. Functional Requirements

Functional Specification:
The application is user-friendly.
It provides an easy interface to the user.
The accessibility or response time of the application should be fast.
Dependency and Constraints:
End User application will be developed in Windows OS.
All scripts shall be written in Python.
The application design pattern shall be Singleton.

5. Functional Interface Requirements

User Interfaces:
Tkinter: Tkinter is Python's de-facto standard GUI (Graphical User Interface) package. Tools in Tkinter Window: It refers to a rectangular area somewhere on the user's display screen. Top-level window: A window that exists independently on the screen. It will be decorated with the standard frame and controls for the desktop manager. It can be moved around the desktop, and can usually be resized
HARDWARE INTERFACES:
System Processors: Core2Duo Speed: 2.4 GHz
Hard Disk: 150 GB
SOFTWARE INTERFACES:
Operating System: 64bit Windows 10 Coding Language: Python
Design Constraints: Spyder & PyCharm

6. Non-Functional Requirements

Performance Requirement:
The performance of the functions and every module must be well.
The overall performance of the software will enable the users to work efficiently.
The performance of detect Depression levels should be fast.
The performance of the providing virtual environment should be fast.
Safety Requirements:
The application is designed in modules where errors can be detected and fixed easily. This makes it easier to install and update new functionality if required.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

|| Volume 11, Issue 12, December 2022 ||

| DOI:10.15680/IJIRSET.2022.1112015 |

7. Software Quality Attributes

The software has many qualities and attributes that are given below:
Adaptability: This software is adaptable by all users.
Availability: This software is freely available to all users. The availability of the software is easy for everyone.
Maintainability: After the deployment of the project if any error occurs then it can be easily maintained by the software developer.
Reliability: The performance of the software is better which will increase the reliability of the Soft- ware.
User Friendliness: Since the software is a GUI application; the output generated is much user friendly in its behavior.
Integrity: Integrity refers to the extent to which access to software or data by unauthorized persons can be controlled.
Security: Users are authenticated using many security phases so reliable security is provided.
Testability: The software will be tested considering all aspects.

IV. ANALYSIS MODELS

The Waterfall Model is a step-by-step approach to software development. Each stage of the SDLC phase is intended to accomplish a specific goal. Winston Royce first presented it in 1970. This is how we will use it in our project. Because it is simple, straightforward, and uncomplicated, this paradigm is simple to understand and apply. It's simple to maintain because the model is tight, with clear deliverables and a review procedure for each phase. The waterfall paradigm works well for smaller projects with well-defined and well-known requirements.

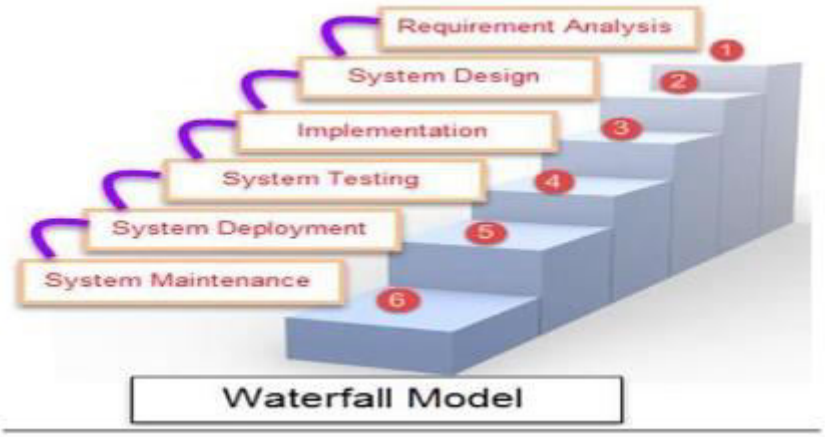

Figure: 1 Water fall Model

V. SYSTEM DESIGN

1. System Design

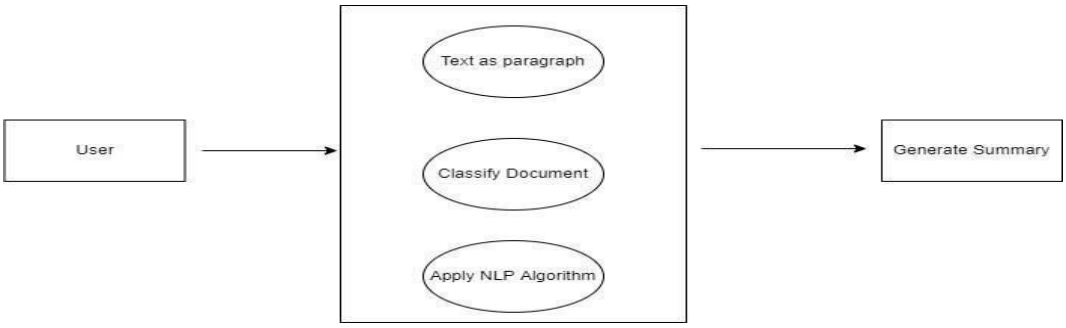

Figure: 2 System Architecture

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

|| Volume 11, Issue 12, December 2022 ||

| DOI:10.15680/IJIRSET.2022.1112015 |

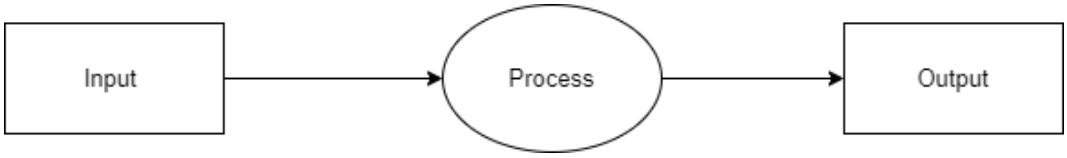

Figure: 3 DFD 0

1. **Data Flow Diagrams**

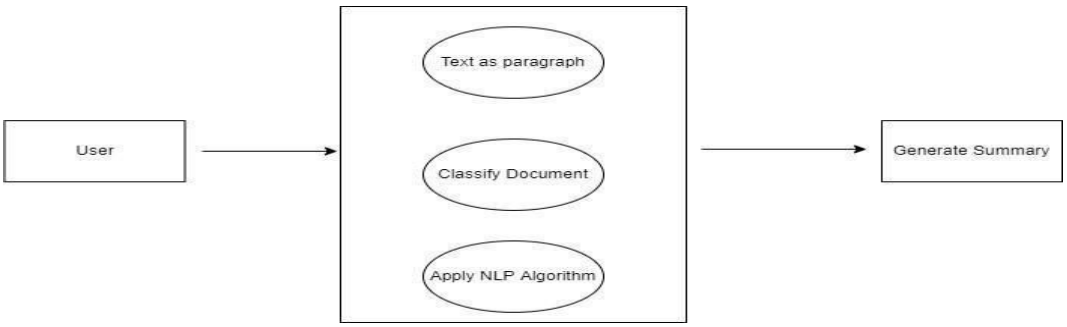

Figure: 4 DFD 1

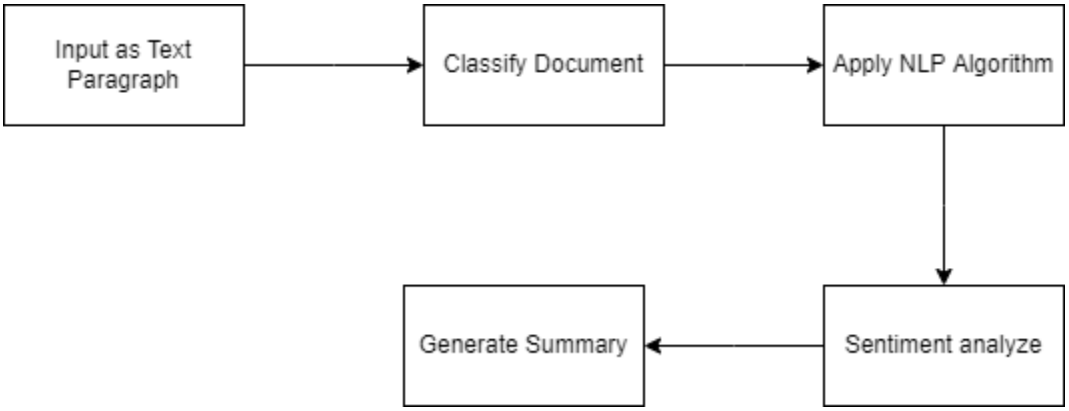

Figure: 5 DFD 2

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

|| Volume 11, Issue 12, December 2022 ||

| DOI:10.15680/IJIRSET.2022.1112015 |

2. UML Diagrams

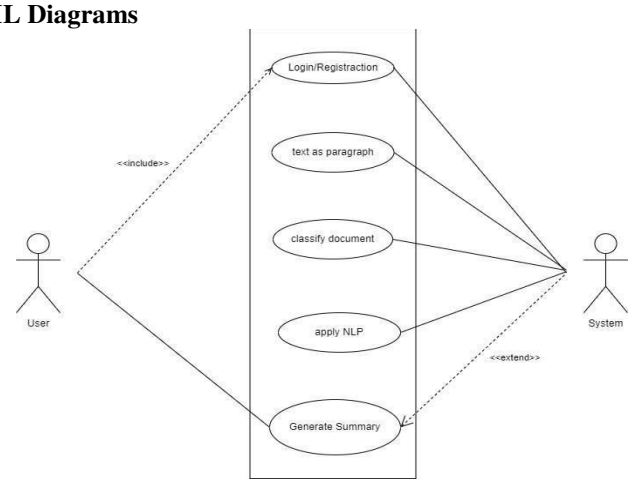

Figure: 6 UML Diagram

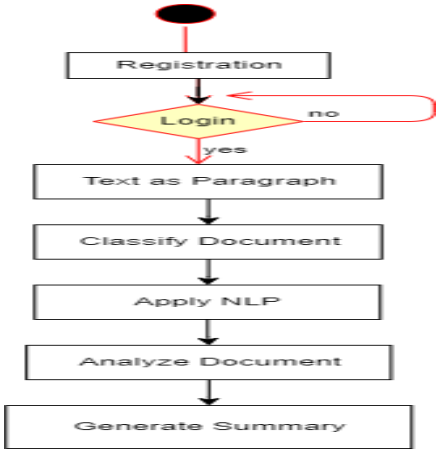

Figure: 7 Activity Diagram

The Unified Modelling Language (UML) is a programming language that is used to create software blueprints. The UML can be used to visualize, specify, build, and document software-intensive system artifacts. Although UML is process independent, it is best used in processes that are use case driven, architecture-centric, iterative, and incremental. There are a limited number of UML Diagrams available.

Use case diagram, Activity diagram, Sequence Diagram, and Class diagram.

VI. ADVANTAGES AND APPLICATIONS

- User-Friendly system
- Easy to generate document summaries.
- Improved best accuracy.

Abstractive methods choose words based on semantic understanding, even if the words do not appear in the source documents. It intends to create significant material in a novel manner. They interpret and analyze the text using advanced natural language techniques to generate a new, shorter text that conveys the most important information from the original text.

VII. CONCLUSION

Autonomy abstracted Document summarizing is a multi-step process with several sub-tasks. Every subtask has the ability to generate high-quality summaries. Abstractive document summarization requires identifying necessary paragraphs from a given document. We suggested an abstractive based text summary in this research, based on a statistical new approach based on sentence ranking, in which the summarizer selects phrases based on their rating. Abstract sentences are created as a summarized text that is then turned into audio. The proposed model outperforms the old technique in terms of accuracy.

REFERENCES

1. R. Socher, "Boiling the information ocean," 2020. [Online].Available: http://tiny.cc/45ohlz
2. S. Chopra, M. Auli, and A. M. Rush, "Abstractive sentence summarization with attentive recurrent neural networks," in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguist.: Human Lang. Technol., 2016, pp. 93–98.
3. A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," in Proc. Annu. Meet. Assoc. Comput. Linguist., 2017, pp. 1073–1083.
4. R. Paulus, C. Xiong, and R. Socher, "A deep reinforced model for abstractive summarization," in Proc. Int. Conf. Learn. Representations, 2018.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

|| Volume 11, Issue 12, December 2022 ||

| DOI:10.15680/IJIRSET.2022.1112015 |

5. Y. Liu and M. Lapata, “Text summarization with pretrained encoders,”in Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process., 2019, pp. 3721–3731.
6. K. Song, X. Tan, T. Qin, J. Lu, and T. Y. Liu, “MASS: Masked sequenceto sequence pre-training for language generation,” in Proc. Int. Conf. Mach.Learn., 2019, pp. 5926–5936.
7. L. Dong et al., “Unified language model pre-training for natural language understanding and generation,” in Adv. Neural Inform. Process. Syst., 2019, pp. 13 042–13 054.
8. Y. Yan et al., “ProphetNet: Predicting future N-gram for sequence- to sequence pre-training,” 2020, arXiv:2001.04063.
9. K. M. Hermann et al., “Teaching machines to read and comprehend,” inAdv. Neural Inform. Process. Syst., 2015, pp. 1693–1701.
10. R. Nallapati, B. Zhou, C. dos Santos, C. Gulcehre, and B. Xiang, “Abstractive text summarization using sequence-to-sequence RNNs and beyond,” in Proc. SIGNLL Conf. Comput. Natural Lang. Learn. Stroudsburg, PA,USA, 2016, pp. 280–290.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.118

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details