International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 10 Issue 6 ǁ 2022 ǁ PP. 1818-1824

Extractive Text Summarization

Adwaith S[1],Amal Joy[2],Aysha Basheer[3],Blessy K Babu[4],Rajesh K S[5]

[1]*PG Student, Department of Computer Applications - Saintgits College of Engineering (Autonomous) PathamuttomKottayamKerala India-*

[2]*PG Student, Department of Computer Applications - Saintgits College of Engineering (Autonomous) PathamuttomKottayamKerala India -*

[3]*PG Student, Department of Computer Applications - Saintgits College of Engineering (Autonomous) PathamuttomKottayamKerala India*

[4]*PG Student, Department of Computer Applications - Saintgits College of Engineering (Autonomous) PathamuttomKottayamKerala India -*

[5]*Professor,DepartmentofComputerApplications- SaintgitsCollegeofEngineering(Autonomous)PathamuttomKottayam KeralaIndia –*

ABSTRACT
Inthisnewera,wheretremendousinformationisavailable on the internet, it is most important to providetheimprovedmechanismtoextracttheinformationquicklyandmostefficiently.Itisverydifficultforhuman beings to manually extract the summary of alargedocumentsoftext.Thereareplentyoftextmaterialavailableontheinternet.So,thereisaproblem of searching for relevant documents from thenumber of documents available, and absorbing relevantinformation from it. In order to solve the above twoproblems,theautomatictextsummarizationisverymuch necessary. Text summarization is the process ofidentifying the most important meaningful informationinadocumentorsetofrelateddocumentsandcompressing them into a shorter version preserving itsoverall meanings. As the name suggests, this techniquerelies on merely extracting or pulling out key phrasesfrom a document. It is then followed bycombiningthese key phrases to form a coherent summary.WehaveimplementedOpticalCharacterRecognition(OCR)textdetection technique. The experimental results show that Extractive Text Summarization with a pretrained encoder model achieved the highest values for ROUGE1, ROUGE2, and ROUGE-L (44.05, 20.43, and 39.88, respectively).

Key words: *Natural Language Processing, Abstractive Summarization,Extractive Summarization.*

Date of Submission: 15-06-2022 Date of acceptance: 30-06-2022

I. INTRODUCTION

Textsummarizationreferstothetechniqueofshortening long pieces of text. The intention is to createa coherent and fluent summary having only the mainpointsoutlinedinthedocument.Automatictextsummarizationisacommonprobleminmachinelearningandnatural languageprocessing (NLP).The extractive text summarization techniqueinvolvespullingkeyphrasesfromthesourcedocumentandcombining them to make a summary. The extraction ismadeaccordingtothedefinedmetricwithoutmaking anychangestothetexts.Before going to the Text Summarization, first wehave toknow that what a summary is. A summary is a textthat is produced fromone or more texts, that conveys importantinformationin the original text, and it is of a shorter form. The goalofautomatictextsummarizationispresentingthesource text into a shorter version with semantics. Themost important advantage of using asummary is, itreduces the reading time. Text Summarization methodscanbeclassifiedintoextractiveandabstractivesummarization. An extractive summarizationmethodconsists of selecting important sentences, paragraphsetc.fromtheoriginaldocumentandconcatenatingthem into shorterform.AnAbstractivesummarizationisanunderstandingofthemainconceptsinadocumentandthenexpressthoseconceptsinclearnaturallanguage. Therearetwodifferent groups of text summarization: indicative andinformative. Inductive summarizationonlyrepresentthe mainideaof the textto the user. The typicallength of thistypeof summarization is 5 to 10 percentof the main text. On the other hand, the informativesummarization systemsgives concise information ofthe main text. The length of informative summary is 20to30percentof themaintext. The technology we are using in the work is NLP with TextRank algorithm.

II. LITERATURESURVEY

For the development of this work, a review of two research papers are done.As discussed in[1], Automatic Text summarization is a better approach to bring out the useful informationfast and most effectively in a text using the procedures namely abstractive or extractivemethods. The text may be in single document or multiple documents on the same topic. In thisera, the amount of data present online is very difficult to summarize the entire content andrefine it into suitable form of information. The abundance of unstructured information increasesthe need for automatic systems that can "condense" information from various documents intoa shorter- length, readable summary. Such summaries may befurther required to cover aspecific information needed (e.g.,summarizing websearch results, medical records, questionanswering and more). In this study, they took data from various sources for a particular topicand summarize it for theconvenience of the people, so that they don't have to go through somultiple sites for relevant data. The entire system is focused on creating a system that getsconcisesummaries of technological topics. The implications of this would mean thatknowledge gathering would be easier and time saving. This system can have additional featuressuch as domain-unspecific so that it could be used for any applications, like travelling blogs,product review summarization and many more.

In [2], summarization is a complex task which contains many sub-tasks in it. Every subtask has an ability to get good quality summaries. 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 work, a novel statistical method to perform an extractive text summarization on single document is demonstrated. Theresearch has been conducted in the field of text summarization and there was various approaches and algorithms for summarizing the text. The idea presented in the system ETS is comprised of the methods used in the above-mentioned papers [1] and [2].

In our system, we use extraction-based approach using text rank with the help of NLP technique. Extractive method selects a subset of existing words, phrases or sentences in the original text to form the summary. Sentence similarity matrix is computed using cosine distance. The result of this process is a dense graph representing the document. From this graph, PageRank (PR) is used to compute the importance of each vertex. The most significative sentences are selected based on page rank score. Then, we extract top n sentences. This algorithm will model any document as a graph using sentences as nodes and similarity between the sentences as the weight of the edges between these nodes. The entire system is focused on creating summaries and reduces reading time. Consumption of information becomes a costly and time-consuming process as the information grows in size and with the presence of irrelevant material or noise. This project can be used as a technique to filter them out. Research is still occurring in this field. Various systems have been proposed and some of them have been implemented in the market. Notable social media platforms use this process to generate summaries for posts that are grouped based on the content called topics. These topics are used to engage users online. Google's home feed for example generates summaries based on the user's preferences. Search engines today directly answerthe provided query rather than providing links. The same concept is an application for voice-based assistants while answering the user's queries.

III. IMPLEMENTATION

The implementation is one phase of software development. Implementation is that stage where theoretical design is turned into working system. Implementation involves placing the complete and tested software system into actual work environment. Implementation is concerned with translating design specification with source code. The primary goal of implementation is to write the source code to its specification that can be achieved by making the source code clear and straight forward as possible. Implementation means the process of converting a new or revised system design into operational one. The implementation is the final stage and it's an important phase. It involves the individual programming, system testing, user training and the operational running of developed proposed system that constitute the application subsystems. One major task of preparing for implementation is education of users, which should really have been taken place much earlier in the project when they were being involved in the investigation and design work. During this implementation phase system actually takes physical shape.

Following figures (1) and (2) shows the working steps and phases of extractive text summarization. The task of extractive text summarization is composed of mainly three phases: thedata pre-processing phase, algorithmic processing phase, and post-processing phase.

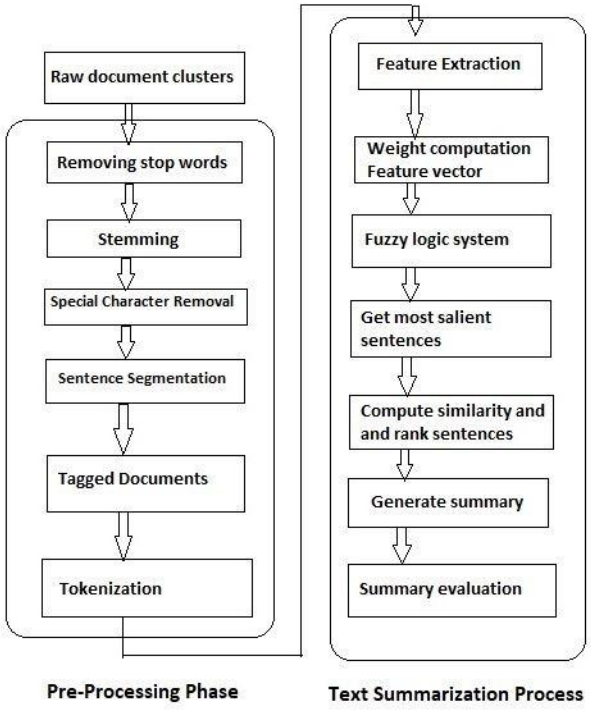

Figure 1: Working Steps of Text Summarization

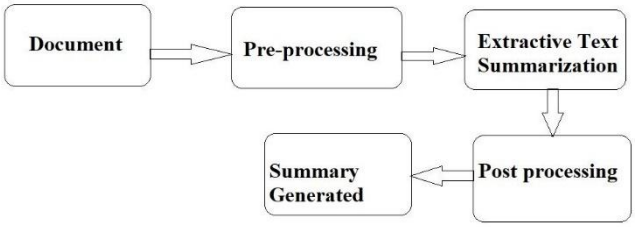

Figure 2: Overview of Text Summarization

3.1 TOOLS AND PLATFORMS USED

3.1.1 FLUTTER

Flutter is a cross-platform tool intended for creating Android and ios app from a single code base by using a modern reactive framework. The major components of Flutter include:

- Dart platform
- Flutter engine
- Foundation library
- Design-specific widgets
- Flutter Development Tools (DevTools)

3.1.2 MONGODB

MongoDB is an open-source NoSQL database management program. NoSQL is used as an alternative to traditional relational databases. NoSQL databases are quite useful for working with large sets of distributed data. MongoDB is a tool that can manage document- oriented information, store or retrieve information.

3.1.3 NODE JS

Node.js is an open-source, cross-platform, back-end JavaScript runtime environment that runs on the V8 engine and executes JavaScript code outside a web browser. Node.js lets developers use JavaScript to write command line tools and for server-side scripting—running scripts server-side to produce dynamic web page content before the page is sent to the user's web browser. Express is a minimal and flexible Node.js web application framework that provides a robust set of features to develop web and mobile applications. It facilitates the rapid development of Node based Web applications.

3.1.4 DEBIAN

Debian is a popular and freely-available computer operating system that uses the Linux kernel and other program components obtained from the GNU project. Debian GNU/Linux is the primary Debian distribution and the only distribution that has been officially released and considered ready for production.

3.1.5 AWS EC2

Amazon Elastic Compute Cloud (Amazon EC2) provides scalable computing capacity in the Amazon Web Services (AWS) Cloud. Using Amazon EC2 eliminates need to invest in hardware up front, so you can develop and deploy applications faster. Can use Amazon EC2 to launch as many or as few virtual servers as we need, configure security and networking, and manage storage. Amazon EC2 enables to scale up or down to handle changes in requirements or spikes in popularity, reducing need to forecast traffics.

In this paper, we have uses TextRank algorithm. The algorithm is a graphbased sorting algorithm for text. The algorithm is an extraction type unsupervised text summarization method.

3.1.6TEXTRANK ALGORITHM

Steps of TextRank algorithm:

- The first step is to integrate all the articles into text data.
- Next, split the text into individual sentences.
- Then, we'll find a vector representation for each sentence (The word vector).
- Calculate the similarity between sentence vectors and store them in the matrix.
- Then the similarity matrix is transformed into a sentence node. The similarity score is the graph structure of the edge, used in sentences TextRank Calculation.
- Last, a certain number of the highest ranked sentences make up the final summary.

IV. EXPERIMENTS AND RESULTS

Extractive summarization picks up sentences directly from the document based on a scoring function to form a coherent summary.

This method work by identifying important sections of the text cropping out and stitch together portions of the content to produce a condensed version.A typical flow of extractive summarization systems consists of:

1. Constructs an intermediate representation of the input text intending to find salient content.
2. Scores the sentences based on the representation, assigning a value to each sentence denoting the probability with which it will get picked up in the summary.
3. Produces a summary based on the top most important sentences.

Through this work, it was found out that the text summarization process is a lengthy and time-consuming process that involves having to read and analyze the entire text. This method summarizes the text by selecting the most important subset of sentences from the original text. As the name suggests, it extracts the most important information from the text. The most favorite features are sentence length and sentence position. In the sentence length feature, long sentences contain more important or relevant information. That means short sentences do not cover any relevant information, so short sentences are considered unimportant or ignored. For calculation of Sentence Length (SL) can be seen in the following equation, where the variable is the length of the sentence, No. of a word occurring in is the variable that shows the number of words in the sentence, and No. word occurring in the longest sentence is the variable that shows the number of words in the longest sentence.

$$SL = \frac{No.of\ a\ word\ occurring\ in\ S}{No.of\ words\ occurring\ in\ longest\ sentence}$$

Extractive Text Summarization method does not have the capability of text generation by itself and hence the output will always have some part of the original text.This would mean significant time lost and this would mean that the work mentioned in this paper would significantly reduce the time required for summarizing and would help improve the productivity ofusers.

From the analysis of the text summarization system, it was observed that most of theexisting systems have been built either on statistical approaches or on linguistic approaches.Here we have followed a statistical method, which started with shallow features such as termfrequency and gradually extended to positional features and domain-specific thematic featuresto improve the quality of summary. This approach is simple and faster in implementation. Theyworked efficiently with larger documents also. The statistical techniques lacked in semanticanalysis of the textual units and thus generated summary that lacked cohesiveness andcoherence. But the linguistic techniques explore the discourse structure of the document byusing semantic analyses of the text. This technique generates cohesive summary as comparedto statistical techniques using shallow features. It has high complexity level of implementationas compared to statistical techniques and works slower for large documents. It is not useful fordomain-specific summarization as it does not use domain- specific features. To

achieve thebenefits of statistical and linguistic methods a hybrid approach must be used to generate asummarization system that uses semantic analysis of document along with important featuresof textual units in news domain and for resolution of correlation of sentences.The experimental results show that Extractive Text Summarization with a pretrained encoder model achieved the highest values for ROUGE1, ROUGE2, and ROUGE-L (43.85, 20.34, and 39.9, respectively).

Figure 1 is the screen where we give the title and the text for summarization and Figure 2 is the screen where we get the summary.

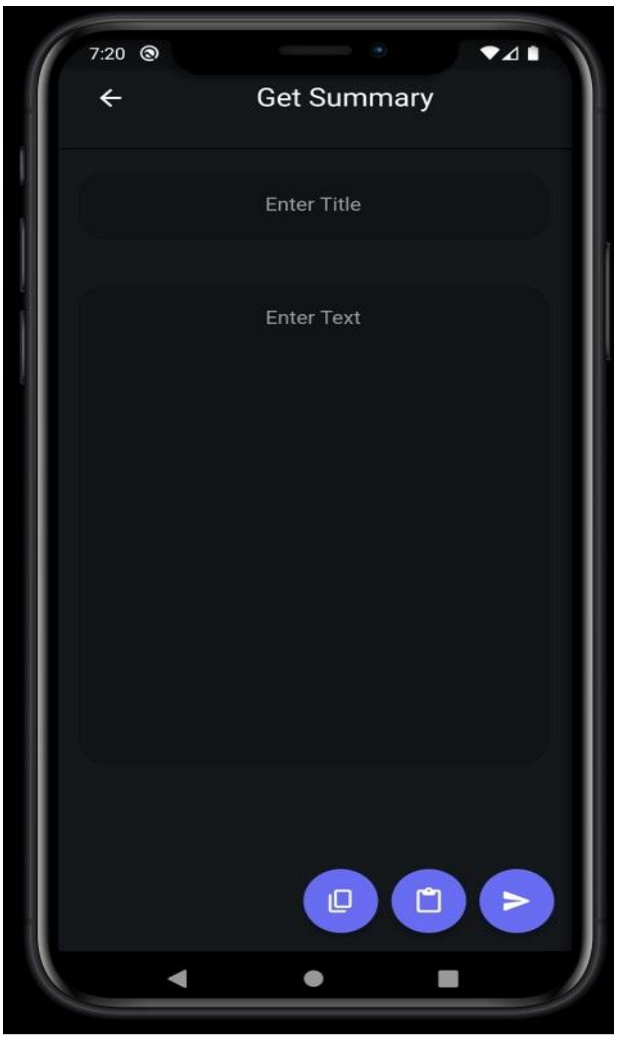

Figure 3 : Main Screen

Figure 4 : Result Screen

V. CONCLUSION AND FUTURE SCOPE

With the ever-growing text data, text summarization seems to have the potential forreducing the reading time by showing summaries of the text documents that capture the keypoints in the original documents. Applying text summarization on each article can potentiallyimprove customer experience and employees' productivity. With open-source software andword embedding packages becoming widely available, users are stretching the use case of thistechnology.The rate at which the information is growing is tremendous. Hence it is very important to build a multilingual summarization system and this research could be a stepping stonetowards achieving that goal provided there is availability of online lexical databases inother languages. The most effective and versatile methods used so far in automatic summarizationrely on extractive methods: they aim at selecting the most relevant sentences fromthe collection of original documents in order to produce a condensed text rendering important pieces of information.This survey paper covers different types of summarization processes based on extractive and abstractive techniques by using different algorithmsfor the summarization. Summarization processes have to produce a compelling summary in a brief time with less redundancy having linguistically correct sentences.All the above techniques used to give out good outcomes and also efficient summariesare obtained according to the context used. It canbe observed that a combination of the pre-processing and post-processing techniques discussed above could give an efficient modelcontaining all relevant features to the model user friendly.The Implemented system in this thesis can work as framework for the research communityto understand and extend the applicability of cognitive and symbolic approach in variousdomains of business needs.Research in summarization continues to enhance the diversity and information richnessand strive to produce coherent and focused answers to users information need.

REFERENCES

[1]. A. R. Mishra, V. K. Panchal and P. Kumar, "Extractive Text Summarization - An effective approach to extract information from Text," 2019 International Conference on contemporary Computing and Informatics (IC3I), 2019, pp. 252-255, doi: 10.1109/IC3I46837.2019.9055636.

[2]. J. N. Madhuri and R. Ganesh Kumar, "Extractive TextSummarization Using Sentence Ranking," 2019 International Conference on Data Science and Communication (IconDSC), 2019, pp. 1-3, doi: 10.1109/IconDSC.2019.8817040.

[3]. X. You, "Automatic Summarization and Keyword Extraction from Web Page or Text File," 2019 IEEE 2nd International Conference on Computer and Communication Engineering Technology (CCET), 2019, pp. 154-158, doi: 10.1109/CCET48361.2019.8989315.

[4]. D. Yang and A. N. Zhang, "Performing literature review using text mining, Part III: Summarizing articles using TextRank," 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 3186-3190, doi: 10.1109/BigData.2018.8622408.

[5]. Jaiswal, A.& Bhatia, N., (2016, January). Automatic textsummarization and it's methods-a review. In 2016 IEEE 6thInternational Conference-Cloud System and Big DataEngineering (Confluence) (pp. 65-72).

[6]. Moratanch, N., &Chitrakala, S. (2017, January). A surveyonextractive text summarization. In 2017 IEEE internationalconference on computer, communication and signalprocessing (ICCCSP) (pp. 1-6).

[7]. Mozhedehi, A. T., Rahimi,S. R.(2017,December). An Overviewon Extractive Summarization. IEEE International Conferenceon Knowledge based Engineering and Innovation (KBEI).

[8]. U. Hahn, I. Mani, "of Automatic Researchers areinvestigating summarization tools and methodsthat", IEEE Computer33. 11, pp. 29-36, November 2000.