A MACHINE LEARNING APPROACH TO EXTRACTIVE TEXT SUMMARIZATION

[1]Dr.Venkata Kishore Kumar Rejeti, [2]M. Naga Lakshmi Devi, [3]N. Lakshmi Vasavi, [4]M. Vanaja, [5]M. Hasini

[1]Professor, Department of CSE, KKR & KSR Institute of Technology and Sciences, A.P, India.
[2,3,4,5]B.Tech Student, Department of CSE, KKR & KSR Institute of Technology and Sciences, A.P, India.

ABSTRACT
The main proposal of this project is to create an application which helps us to summarizes huge amount of text into maximum shortened form with in less amount of time. By using NLP several text summarization methods are being used to generate maximum shortened version of text. Proposed system takes input as text and pdf format where text is being extracted from it and a reference summary and range scale of text to include in summary. Initially extractive text summarization is applied on text. Based on result abstractive summarization is done by replacing words with synonyms from WordNet. Later interactive summarization is done on it. Based on input summarized text is displayed along with rouge and meteor score. Visual representation of important words is being displayed through word Cloud.

KEYWORDS – Extractive summarization, Abstractive summarization, Rouge and meteor score, WordCloud visualization, Stream lit UI.

I. INTRODUCTION

Imagine you have long document or article to study, and you want to get main points quickly without reading it completely. That's where text summarization concept arises. It is the process of converting huge amount of text into much shorter version of original text. Which in turn saves user time and effort. It produces maximum shortened summary which reflects original text. Thus, Machine Learning Approach is used to implement Text Summarization. Where it has computer algorithms and statistical methods to understand pattern. It has main components like data, training model, identifying patterns. Several training models like supervised and unsupervised are used to understand data. Among all methods, Extractive Text summarization method is used to summarize the text. By using BERT Model (bidirectional Encoder Representations from Transformers) sequence classification to score sentences. Identifies top ranked sentences based on scores obtained. Here specific sentences are being selected and extracted from original text to create condensed version that retains essential information. Where new text is not generated from it. Later abstractive summarization is done by replacing words with respective synonyms. WordNet is used for obtaining synonyms of respective words. Here the least length of obtained synonym is replaced with words in input text. NLTK is used for tokenization and comparison.

Along with input user need to provide reference summary for accuracy prediction. Based on summary and reference summary accuracy is being calculated. Through Rouge quality of generated summary is being compared to reference summary. Accuracy is being calculated and displayed to user through visual representation of rouge and meteor scores. User input can be text and pdf format where text is being extracted and summarized. We are providing three selection types of text summarization likely Summarization by BERT, Summarization by txtai and Summarization by document. In txtai is selected for text summarization without reference

summary. Here accuracy is not displayed. If BERT is selected, Wordcloud visualization is shown to user where it visually displays important words in given input text based on their usage in input text. Through this text summarization user can able to understand text completely without much effort and time consumption. The overall proposed system is the creation of a short, accurate and fluent summary of long text and pdf. This could help to consume relevant information faster and easier.

II. OBJECTIVES

It Provides user interface to summarize huge amount of text into maximum shortened version of input text by using NLP BERT Model.
It can able to take pdf format input, text is being extracted from it and gets summarized. It will display summarized text.
Based on reference summary and generated summary accuracy is calculated and displayed using rouge and meteor scale to user.
Not only generated summary, it can also able provide important words through WordCloud by using frequencies of words obtained from given text.
WordCloud visualization is displayed to user where important words are displayed with different sizes based on their usage.

III. LITERATURE REVIEW

[1] In 2019, an essential study was conducted by Prabhudas Janjanam et.., al on Text Summarization, as text compression techniques are advanced and improved by machine learning models, where abstraction deals the problem in different ways; it involves paragraphing salient text using Natural Language Generation; first understand the complete document and then apply sentence compression; and apply fusion techniques for summarization; semantic analysis is used to deal with the semantics; graph based methods will summarize the text by representing whole graph text in graphs; but comparison between graphs would be difficult.
[2] In 2020, a study was conducted by Surabhi Adhikari et.., al on NLP based machine learning approach for Text Summarizer helps in defining the content by considering the important words and helps in creating summaries that are in a human-readable format; EXT text summarization is a way of generating summaries by using the same sentences as in the document; ABS focuses on key concepts of the document; Machine Earning(ML), NNs, Reinforcement learning, Sequence to sequence modeling and fuzzy logic methods are used for Text summarization; but not much efficient.
[3] In 2020, comparative study by Sharmin Akte et.., al on Abstractive text summarization, where extractive text summarization(ETS) is used to summarizes text (Or) documents; In ETS extracted summary could turn out to be longer than average; graph based process, discourse based process, term frequency tactics, clustering methods are used for Text summarization; ATS is used where computer understand the given text first and then computer can generate a summary by its own; but main drawback of extractive text summarization is it can't always produce the expected outcome because of it nature and the work on ATS is not properly done yet because it's a long time process.
[4] In 2021, a study was conducted by Kuntal Gupta on NLP based text summarization on extractive, abstractive, reinforcement learning techniques; graph based, latent variable and term frequency based Extractive Unsupervised summarization techniques are used for Text Summarization; Extractive Supervised summarization strategies diminish the weight of summarization by choosing subsets of sentences; Extrinsic Evaluation and Intrinsic Evaluation methods like ROUGE(Recall Oriented Understudy of Gisting Evaluation) is used for automatic summary evaluation; but there exists causes Anaphora problem and compression results.

[5] in 2020, Dima Suleiman et..., al, author reviewed on Deep Learning Based Abstractive Text Summarization, neural networks with an attention mechanism and long short-term memory (LSTM) are the most prevalent techniques used; It provides an overview of the approaches, datasets, evaluation measures, and challenges of deep learning based abstractive text summarisation, several deep learning models like recurrent neural network (RNN), bidirectional RNN, attention mechanisms, long short-term memory (LSTM), gated recurrent unit (GRU), and sequence to-sequence models these models are used; but it faced challenge such as inaccurate sentences, where we are providing accurate sentences

[6] in 2021, Divakar Yadav et..., al, author reviewed on Automatic Text Summarization Methods, as extractive and abstractive approaches are used; based on number of input documents that is single document or multiple document, based on summarization methods like ETS and ATS summarization will be done, Generic text summarizers is used to fetch important information from single or multiple documents, Query based summarizer is used to handle the multiple documents and give solution as per the user's query summarization of multiple documents is difficult, where we are providing multiple documents summarization.

[7] in 2021, Manish Shinde et.., al, an essential survey conducted on Text Summarization, as text summarization techniques are extractive and abstractive are improved; Term Frequency and Inverse Document Frequency Approach are used to show significant a piece of data is in the given input; Attention Based Seq2Seq model is used to avoid attempting to learn a single vector representation for each sentence; summarization with Pointer Generator Networks is used to create new phrases and fresh words; Pre-training with extracted gap sentences for Abstractive text summarization is used;but evaluation of the summaries is difficult.

[8] in 2016,Deepali K et…, al, an essential reviewed on Text Summarization, as text summarization techniques are extractive and abstractive are improved; Abstractive Structured based Approach is used to encodes most important information from the document; Abstractive methods like Tree based method, Template based method, Ontology based method, Lead and body phrase method are used; It uses both semantic and Structured based Approaches; Extractive Text summarization techniques like Term Frequency Inverse Document Frequency Method, Cluster based method etc, Indian languages comparison Text summarizers are used; but less work is done using abstractive methods on Indian languages.

[9] in 2019, Fabio Bif Goularte et..., al, author reviewed on text summarization method based on fuzzy rules and applicable to automated assessment, as text summarization techniques are Fuzzy logics summarization are developed; And metrics for more abstract, higher-level or somewhat subjective attributes, preprocessing methods like Tokenization, Stop word removal are used for summarized text; ROUGE measures quality by counting overlapping units such as n-grams, word sequences and word pairs between the automatic summary and the reference summary; but Fuzzy logic did not produce promising results

[10] in 2020, Parth Rajesh Dedhia et…, al, author studied on Abstractive Text Summarization, as text summarization models are RNN and attention are improved; In this paper they use the models of RNN for development of attention models, pointer attention and how produce abstractive text summaries by synergy. And in this they didn't used the both extractive and interactive but they only focused on abstractive but in our proposed system we use the three methods Abstractive, Extractive and Interactive for desired results with our own respective queries like percentage of text included in summary. In this the RNN model will initialized by the sparse work; but does not work if multiple documents are passed to the model

[11] in 2023, DivakarYadav et…, al, author reviewed on Qualitative Analysis of Text Summarization Techniques and Its Applications in Health Domain, In this paper they applied assessed diverse extractive synopsis methods, and it only focused on improving the proficiency of rundown strategies but our proposed system will focused on accuracy and faster and easier summary by Txtai with understandable way like as human language. They use the PEGASUS

for abstractive text summarisation and text rank for extracting text summarisation for datasets, in our project we use the ROUGE score and meteor score for the results; as text summarization techniques are extractive and abstractive are improved but capture the clinical context is difficult.

[12] in 2020, Pradeepika Verma et..., al, author reviewed on Text Summarization Techniques, as text summarization techniques like Reinforcement Learning are improved; in this paper document summarization is done. But the accuracy of generated summary is not up to the mark. We implement interface with reference summary as input where we can able to find accuracy of generated summary compared to reference summary. It can't able to provide maximum concise summary. Where we can able to provide maximum concise summary of given input document.; but summarizing redundancy documents are difficult.

[13] in 2022, Anjana Kumari et…, al, an essential Real-Life Implementation of Text Summarization Technique, as text summarization techniques like Text mining models are improved. Different methodologies are used to study text mining are term based, phase based and pattern taxonomy models. This approach used to eliminate these arising grievances by discovering phrases that offer additional meaning. Whereas visualization of words is absent in this approach. Through this they can't able to provide accuracy to user. But no providing results up-to mark.

[14] in 2020, Abdul Syukur et…, al, conducted study on Review of automatic text summarization techniques & methods, where it uses latest technology like SLR but can't able to summarize complex language like dialogues. In this methodology fuzzy logic is often used to extract the final value of words or sentences included in the summary. the fuzzy systems work is to use multiple inputs from indices. In this there are main two steps one is preprocessing phase and text summarisation process and compute similarity and rank method will apply. And tokenization present in clustering methods. The score of each feature is then given to fuzzy inference system as input. We are implementing the tokenization will be placed by the BERT methods. Thus, accuracy need to be improved.

[15] In 2023, Review of Text Summarization Techniques of Documents was conducted by SHUBHAM et..., al where NLP and BERT are used to summarize text documents. But no providing results up-to mark. In this methodology they are uses the packages are NLTK for text processing, and used the tensor -flow hub for downloading the BERT pretrained model, and used sklearn library and other common libraries for importing the datasets and in this experiment the first 100 document ls have been executed from the news summary dataset and it also evaluate the results using rouge score method only. And in this page Rank model is used in the process of summary generation and also text rank method is used for extracting pertinent sentences from the lengthy texts using embedding. It has been pretrained using the unlabelled text. And it extracts the keywords from the TFIDF vectorizers.

IV. METHODOLOGY

The project starts by taking input as text or pdf format where text is being extracted from it. Several methodologies are being used to summarize text like Extractive, Abstractive, Evaluation Metrics etc.At first Extractive methodology is being used through BERT Model where important sentences are being extracted from it based on frequency and usage. Through sequence classification score of sentences are being calculated. Identifies top ranked sentences based on obtained scores for extractive text summarization. Later Abstractive compression is done by replacing words with synonyms of minimum length compared to words length in given input text. Through WordNet synonyms of words are being find and get replaced. As proposed system provides range scale from 0 to 100 of input text need to involve. Based on input, reference summary and range scale relevant output is provided to output.

After Text Summarization, Evaluation Metrics method is being applied on reference summary and summarized text. The purpose of Evaluation Metric method is to provide accuracy of summarized text through rouge and meteor score based on reference summary. Word Cloud Visualization is done on summarized text. Where it displays words of different sizes based on their frequency and usage in given input text. So that user can able to understand important words from it. Stream lit is the User Interface used to take input text, percentage of text need to involve in summary. It provides extractive summary, compressed summary and interactive summary along with evaluation metrics and WordCloud visualization.

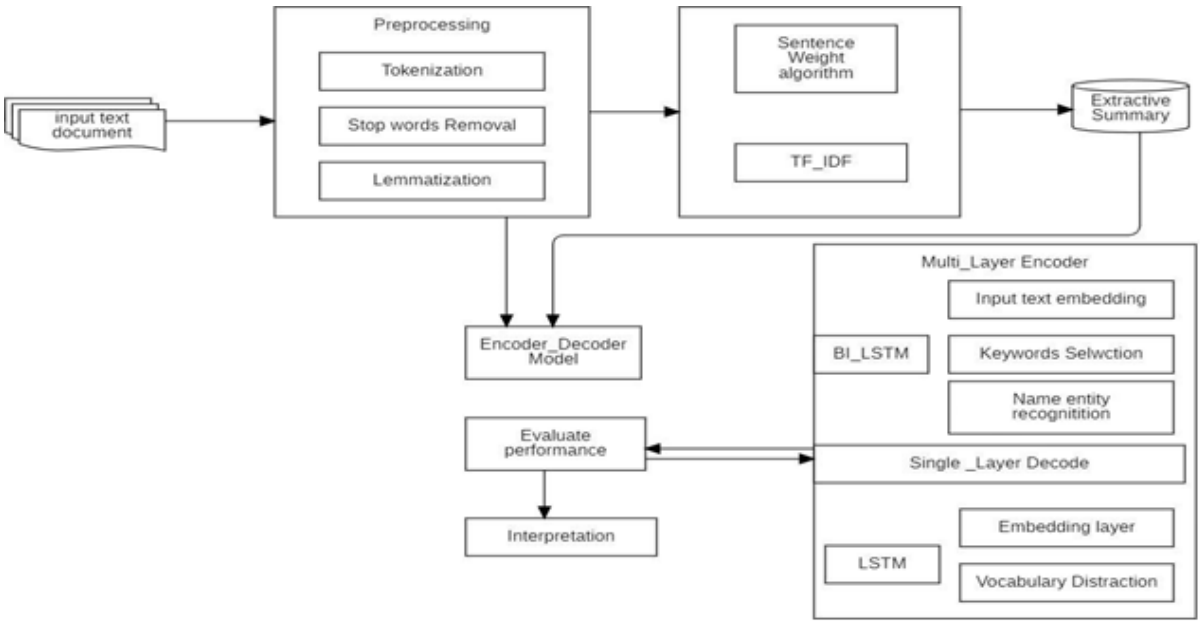

FIG 1: Process in Text Summarization

FIG 1 describes the methods that are being used in text summarization. Input can be text or document which need to be summarized. Methods are applied on input text. Based our output it visually represents rouge and meteor score along with WordCloud. In this the main steps are input document will preprocess and that input will process by the TF_IDE and extractive summary and different types of encoders.

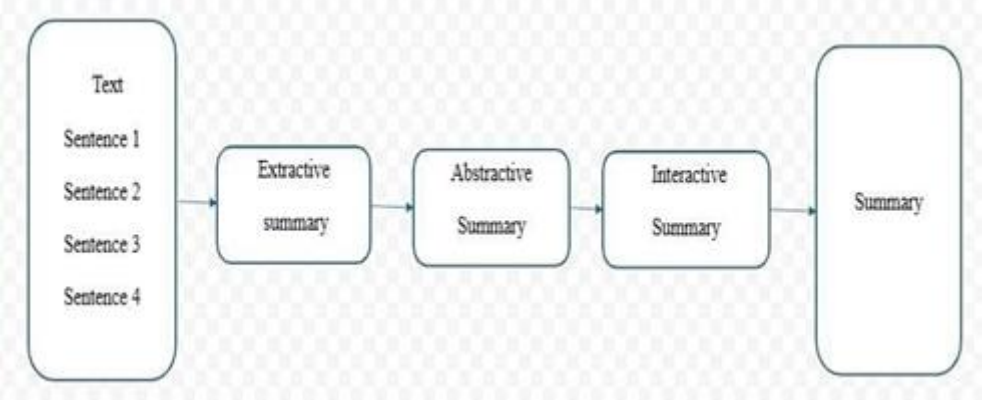

FIG 2: Methods in Text Summarization

FIG 2 describes the methods that are being used in text summarization. Input can be text or document which need to be summarized. Extractive and Abstractive methods are being applied on given input text which produces summary based on reference. In this the input text will

verify by the all three methodologies that gives desired results which performs their respective techniques to get the understandable and human based summaries

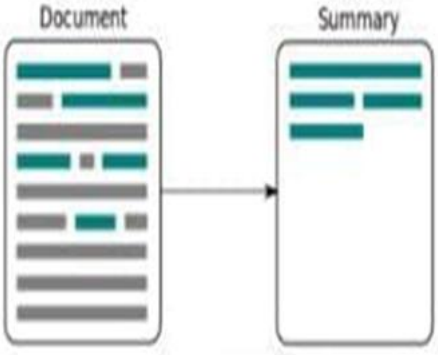

FIG 3: Text Summarization

FIG 3 shows document text summarization here we need provide pdf format input to given system. At first it extracts text from pdf, later methods are applied on the given text to generate summary of respective input which is given by user to the stream lit.

V. RESULTS

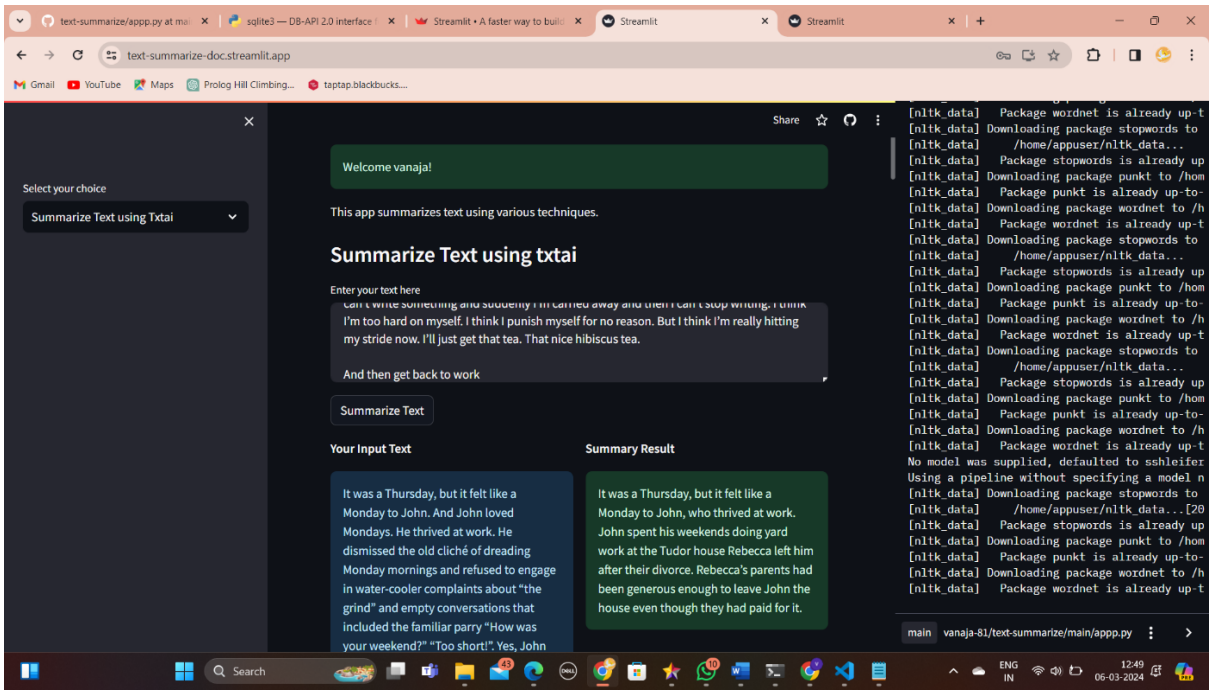

FIG 4: Summarizing text through Txtai

In this methodology we have to give the inputs like input text field only and then click on summarize button. there is no need of the reference summary, some people may have no need of the reference for their own respective results in that way this methodology is useful and easier and faster to get the desired results. The out of the wordcloud will be like this: It gives

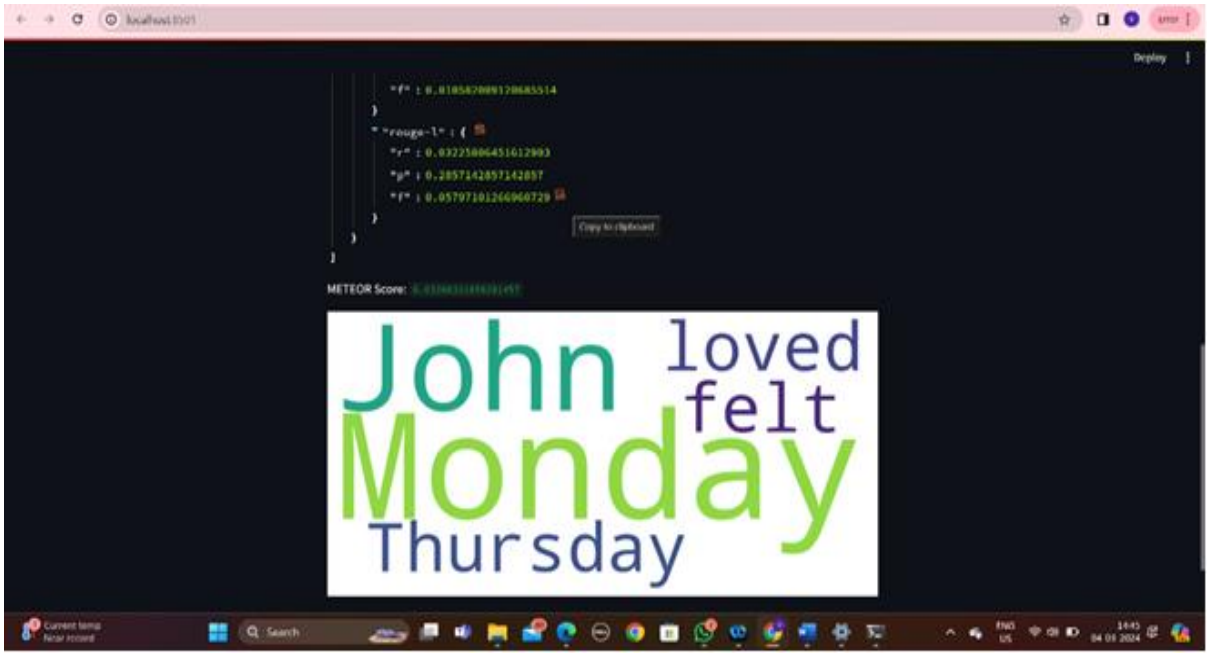

FIG 5: Word cloud output

In this output based on the wordcloud it gives the visualize output and gives the meteor score based on the rouge score,the overall theme will display like image of the text.the wordcloud is used to gives the visualization of the text within the summarized text summary.

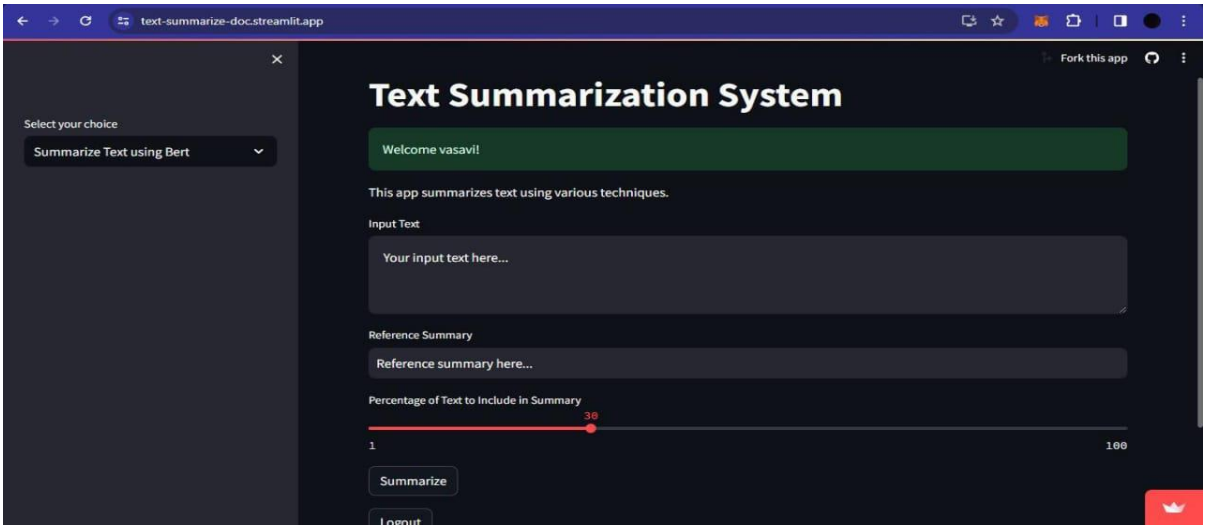

FIG 6: Summarizing text through Bert

FIG 6 depicts summarization of text with reference summary provided by the user. In this the system will takes the inputs like Input text, Reference summary, percentage of text include in summary, the field percentage of the text to include in summary will be based on our requirements, that if we have to include 50% of text then we will make that 50% by that range scale, after complete the input fields we click on the summarize button.it may takes some time based on our network capability that gives the output as summarized text like below:

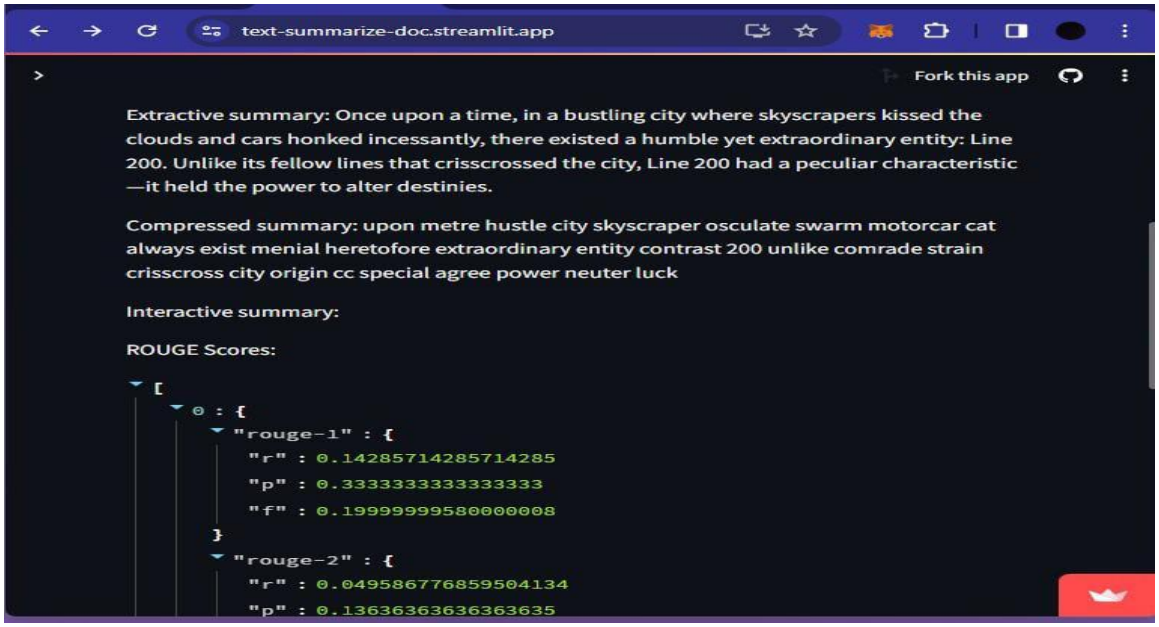

FIG 7: Output of text summarization

FIG 7 depicts the output of text summarization in terms of Extractive, compressed and interactive along with rouge and meteor score.

In this by using BERT methodology that based our reference summary and input field percentage of text include in summary the output will display the field likes Extractive summary, Compressed summary, Interactive summary summary, and it gives two types of scores are ROUGE score and Meteor scores

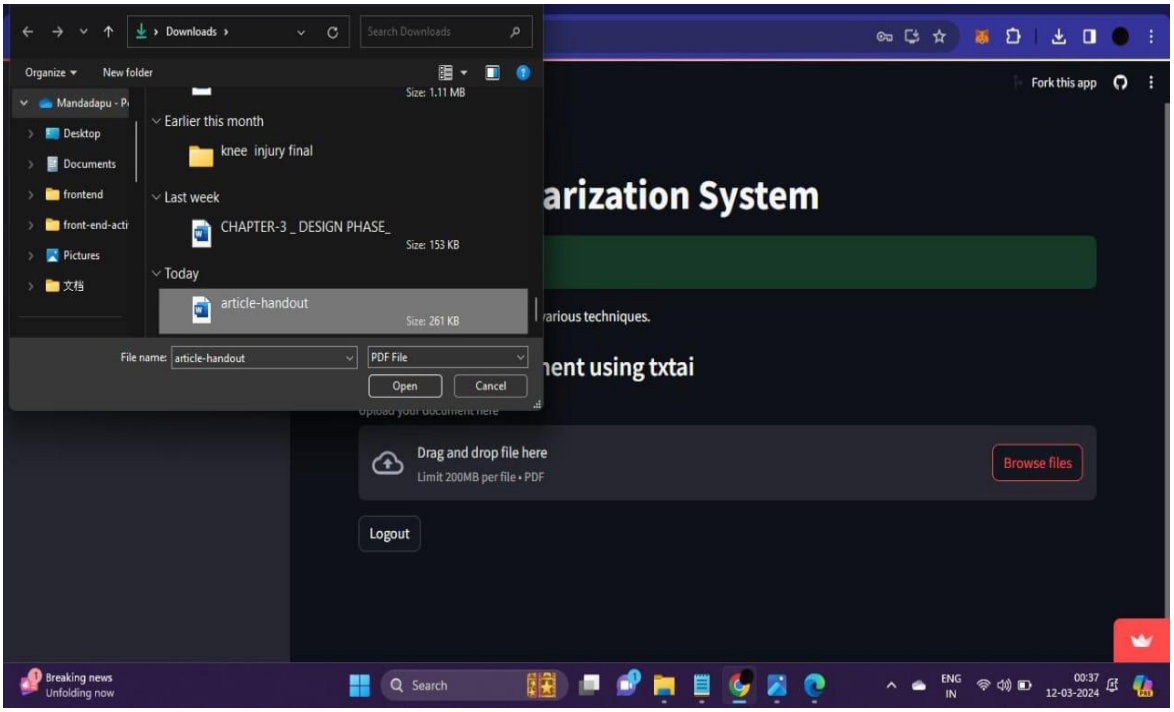

FIG 8: Uploading pdf for text summarization

FIG 8 shows how pdf text is getting summarized. Here we need to upload pdf which need to be summarized. Later text is getting summarized and displayed. The uploaded document must be in a pdf format and that pdf will be in the limit of 200 MB only. If we have put any number of papers in that pdf that maintain the size must be under 200MB.It gives the accurate summary without taking any reference summary and It did not give the rouge score it gives meteor score

based on the word cloud. Here in the time uploaded the document it visible summarize document button and then it gives text like document uploaded successfully.

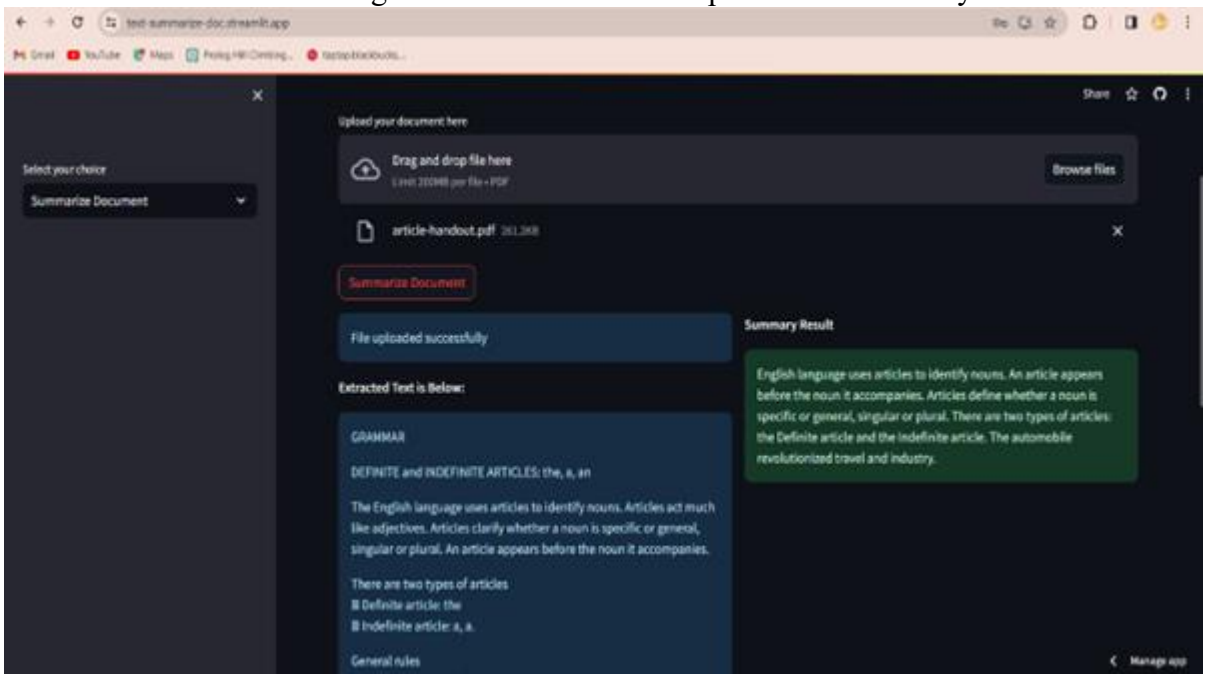

FIG 9: Document summary output

FIG 9 shows Summary results by providing with the Extractive summary and the entire minimized text will display in summary results. Based on frequency and usage of words in given input text importance is being calculated. And in this there is no need of the reference summary we can easily obtain the summary. Words sizes varies from one to other based on its usage. So that user can able to know most frequently used words through Wordcloud representation.

VI. CONCLUSION

This paper is the fundamental study of concepts, methods and algorithms associated to automatic text summarization. Machine -based extractive text summarization offers valuable insights into the effectiveness, versatility and usability of the summarization approach across different input formats like text and pdfs. This paper not only provides summary, it is able to provide Accuracy of generated summary to user by comparing with user reference summary. By displaying Rouge and Meteor Score to user. Wordcloud is used to generate visual representation of most frequent and important words in the given input data with different frequency words.

Text summarization techniques offer a solution for efficiently dealing with this data, providing summaries that are easier to understand and quicker to reduce the length of the given text. It helps to users for better understanding of input and not only provides only summary along with summary it provides worldcloud which shows the important keyword in the given input text. Through this keywords that produced by worldcloud which helps user to understand more quicker which in turn saves time and effort.

VII. FUTURE ENHANCEMENTS

- We can enhance this project in future by extending documents length compared to existing ones so that it can handle large amount of data easily.
- Extending summarization capabilities to multiple languages promoting accessibility to all users.

- Multi documents summarization can also be implemented so that it can handle more no of documents easily and effectively.
- Multimodal summarization can be implemented in such a way that it should integrate information from both text and other modalities like images or videos for better comprehensive summary.

REFERENCES

[1] Prabhudas Janjanam, CH Pradeep Reddy, "Text Summarization: An Essential Study", Second International Conference on Computational Intelligence in Data Science (ICCIDS-2019).
[2]Rahul, Surabhi Adhikari, Monika, "NLP based Machine Learning Approaches for Text Summarization'", Proceedings of the Fourth International Conference on Computing Methodologies and Communication (ICCMC 2020) IEEE Xplore Part Number: CFP20K25-ART; ISBN:978-1-7281-4889-2.
[3] Md Ashraful Islam Talukde, Sheikh Abujar, Abu Kaisar Mohammad Masum, Sharmin Akter, Syed Akhter Hossain,"Comparative Study on Abstractive Text Summarization", IEEE - 49239.
[4] Ishitva Awasthi, Kuntal Gupta, Prabjot Singh Bhogal, Sahejpreet Singh Anand, Prof. Piyush Kumar Soni," Natural Language Processing (NLP) based Text Summarization - A Survey", Proceedings of the Sixth International Conference on Inventive Computation Technologies [ICICT 2021]IEEE Xplore Part Number: CFP21F70-ART; ISBN: 978-1-7281-8501-9.
[5] Dima Suleiman and Arafat Awajan,"Deep learning based abstractive text summarization:approaches ,datasets,evaluation merasures,and challeges",Hindawi Mathematical problems in engineering volume 2020,article id 9365340.
[6] Divakar Yadav, Jalpa Desai, Arun Kumar Yadav, "Automatic Text Summarization Methods:", ORCID.
[7] Manish Shinde, Disha Mhatre, Gaurav Marwal, "Techniques and research in text summarization-A survey",2021 International conference on advanced computing and innovative technologies in engineering.
[8] Deepali K. Gaikwad, C. Namrata Mahender, "A Review Paper on Text Summarization", International Journal of Advanced Research in Computer and Communication Engineering. [11:53 pm, 11/03/2024] Sri lakshmi Mandadapu: [11]. Fábio Bif Goulartea, Silvia Modesto Nassar a, Renato Filetoa, Horacio Saggionb, "Expert Systems With Applications", / Expert Systems With Applications 115 (2019) 264–275.
[9]Fabio Bif Goularte,Silvia Modesto Nassar,Renato Fileto,Horacio Saggion,"A text summarization method based on fuzzyrules and applicable to automated assessment",ELSEVIER 2018.
[10] Parth Rajesh Dedhia, Hardik Pradeep Pachgade, Aditya Pradip Malani, Nataasha Raul, Meghana Naik, Study on Abstractive Text Summarization Techniques,2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE).
[11] DivakarYadav, NamanLalit, RiyaKaushik, YogendraSingh, Mohit, Dinesh, Arun Kr. Yadav, Kishor V. Bhadane, "Qualitative Analysis of Text Summarization Techniques and Its Applications in Health Domain", Hindawi. Pradeepika Verma, Anshul Verma2,"A Review on Text Summarization Techniques", Journal of Scientific Research, Volume 64, Issue 1, 2020.
[12] Pradeepika Verma, Anshul Verma2,"A Review on Text Summarization Techniques", Journal of Scientific Research, Volume 64, Issue 1, 2020.
[13] Anjana Kumari, Arnav Raviraj, Dr.Ajay Kumar Murari, Dr.I. Mukherjee, "Real Life Implementation of Text Summarization Technique", 2022 9th International Conference on "Computing for Sustainable Global Development". Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM).
[14] Abdul Syukur,Adhika pramita widyassari,Supriadi rustad,guruh fajar shidik,edi noersasongko,affandy affandy,de rosal Ignatius moses Setiadi,"Review of automatic text summarization techniques and methods",Journal of king saud university-computer and information sciences.
[15] Shubham U. Pawar, OM S. Behare, Sumit D. UMAP, Akshay K. Adhav, Bhushan B. Pawar, Ram S. Thakare, PROF. Harshada M. RaghuWanshi, "Review of Text Summarization Techniques of Documents",2023 IJCRT | Volume 11, Issue 3 March 2023 | ISSN: 2320-2882