AQI Automation, Statistical Analysis and Forecasting 24, March 2023

AQI Automation, Statistical Analysis and Forecasting

*Palak Goel*

*MSc Data Science*

# *Abstract*— India is the fifth-most polluted country in the world, and the rapid rise in pollution is having a significant impact on the environment in a number of important cities.This study will measure air pollution trends an also evaluate the air quality index with respect to various geographic locations in order to get a global perspective of the harm caused by air pollution and to design suitable methods to prevent it in the future.

# In this regard, the Central Pollution Control Board and the Ministry of Environment, Forestry, and Climate Change's data were used to create the polluted database (India). These statistics, which were gathered at several monitoring stations in the cities of India, show the annual growth of SO2, NOx, and particulate matter (PM) 2.5 in 2021.

# The findings indicate that multiple transport modes, small- and large-scale power generation (from diesel, coal, and gas plants), industries, constructions, and residential cooking were the sources of SO2, NOx, and PM 2.5.

# Overall, India had a trend that was becoming worse every day. According to the findings, that used as a trained dataset for Machine learning algorithms the locations under consideration fell into four categories: critically contaminated (CP), highly polluted (HP), moderately polluted (MP), and low polluted.

**Keywords**: Air Quality Index, 2021, Machine Learning, Automation, statistical analysis

# INTRODUCTION

***Air pollution has become a significant public health concern in India, with cities facing severe levels of pollution year-round. The deterioration of air quality has led to adverse effects on human health and the environment. In recent years, machine learning techniques have shown promise in predicting air quality levels, enabling timely interventions and effective policies. In this research paper, we present a novel approach for predicting air quality levels in Indian cities using machine learning and Long Short-Term Memory (LSTM) models. Our study focuses on the air quality data of 2021, which is critical in understanding the impact of the pandemic on air*** ***quality. Our findings can inform policymakers and stakeholders to implement targeted interventions to improve air quality and protect public health.***

* + 1. BACKGROUND

Automatic text summary aims to deliver the source text in a concise, semantically rich form. The

besides stop words are taken into consideration as keywords.

In the processing phase, factors that affect the

main benefit of adopting a summary is that it cuts relevance of phrases are selected, calculated, and then down on reading time. Informative summarizing given weights using the weight learning approach. systems deliver condensed information of the main Using the feature-weight equation, the final score of text and are 20 to 30 percent of the length of the entire each sentence is calculated. The final summary is

text. chosen from the top-ranked sentences. A crucial

Key processes in text summarization: Topic component of text summarization is summary identification, interpretation, and summary creation evaluation.

are the three primary steps that must be taken while summarizing documents.

1. The most important information in the text is identified as the topic. Different strategies for topic identification are utilized, including location, cue phrases, and word frequency. The most effective methods for topic identification are those that are based on the position of phrases.
2. Interpretation: The interpretation stage is necessary for abstract summaries. In this process, many topics are combined to create a broad content.
3. Summary Generating: The system employs a text generation approach in this stage.

Summary of extracted text: This procedure may be broken down into two steps, Pre-Processing and Processing. Pre-Processing is the original text's organized representation. Typically, it includes:

Most of the extractive summarizers in the past relied on rating sentences in the original text. The most popular and contemporary text summarizing methods either employ linguistic or statistical methods. The sentences are weighted using the high frequency words, standard keyword, cue method, title method, and location method. Most of the the current automatic text summarizing systems create a summary using an extraction strategy.

Summaries can typically be assessed using either intrinsic or extrinsic metrics. Extrinsic methods estimate summary quality through a task-based performance measure, such as an information retrieval-oriented task, while intrinsic methods aim to do so through human evaluation.

To create extraction summaries, sentence extraction algorithms are frequently utilized. Sentence

1. Sentence boundary recognition. The presence of a scoring, which assigns a numerical value to each dot at the conclusion of a sentence in English sentence for the summary, is one technique for finding

indicates the boundary of a sentence. appropriate sentences. The best sentences are then

1. Stop-Word Elimination: Usual words without chosen to make up the document summary based on the semantic significance. This technique gets rid of the compression rate. The compression rate is a crucial words that appear most frequently in documents, component of the extraction procedure that is used to such as articles, prepositions, conjunctions, determine the proportion between the length of the interrogatives, assisting verbs, etc. Due to their summary and the source text. The summary will grow minimal role in the sentence extraction process, stop and contain more irrelevant content as the compression

words are eliminated. rate rises. While the summary becomes shorter due to

1. Stemming: The goal of stemming is to identify the compression rate, more information is lost. In fact, each word's stem or radix, which highlights its the quality of the summary is acceptable when the semantics. the stemming procedure may have a compression rate is between 5 and 30%.

negative or negligible effect on the functionality of

systems involved in semantic analysis. Therefore, we tested the suggested method using both methods of pre-processing (with and without stemming).

1. Tags for Parts of Speech: It is the process of figuring out which words, such as nouns, adverbs, verbs, etc., belong in which part of speech in a phrase. The computational applications, on the other hand, typically employ finer-grained POS tags like "nonplural." The Stanford Log-linear POS tagger was employed in this case.
2. Keyword extraction: In this step, the keywords from a document are extracted. Here, all words
   * 1. OBJECTIVE
   * Optimize the time and space complexity, involved in processing by using the algorithm.
   * Optimize the processing speed, to get the same in less amount of time.
   * Ease of interface for usability
   * Convert audio to text, by a single click.
   * Summarize the output, in text, which helps to read the bulk of information in few lines
     1. PROBLEM STATEMENT

While working and developing the model, multiple issues came into picture and issues were identified while summarizing the document, some of them are:

1. Non-Readability: Summarized text should be readable at its best, which actually means that the text should be free of grammatical errors and must be related to the context.
2. Redundancy: Redundancy in any model plays a very crucial role to get the output. The non- redundancy refers to the novelty in a summary. The summary should be non-redundant to increase the coverage of information residing in a document. summary of the whole document is more important and more informative as it talks about important and key points. Existing model and approaches focuses on finding the relevant content and extract to generate the summary. If we work precisely then we can measure the similarity between the contents of a document, by following this redundancy in the model can be minimized.
3. Irrelevancy: The overall agenda of the summarizer is to get the relevant data or contents from a document. Typically, sentences or other textual units within a document are evaluated using human-engineered text features. Some features may tend to create irrelevant elements in the summary because it is not always possible to include all of the considered features. Thus, increasing complexity and irrelevant by taking into account all textual elements is conceivable. Knowing which attributes are responsible for producing a high-quality summary from the available data is vital in result.
   * 1. RELEVANCE

Depending on services, digitalization, common people who were unable to access the best possible IT services related to personal, social and cultural work. Peak COVID-19 forced people to work virtually, over multiple platform for every work including some official meeting, whereas attending every time or able to note down every key point was not much possible. This led the concept of Summarization. Summaries facilitate the selection of documents for research in less time and reduced space. Performance is improved via automatic summarization, by using some algorithms. Compared to human summarizers, automatic systems are less prejudiced. Because they offer individualized information, personalized summaries are helpful in question-answering systems.

* + 1. SCOPE

Present work increases many new researchers to develop some same or different algorithms as well as thinking of developing some more applications. Working with the optimized model presented here, the file it can be text or audio file must be uploaded model, it can be updated further such that it converts speech- to-text in real time and result as summary. By using this method, efficiency, and performance of the libraries along with the model can be easily evaluated. Further, working can be made much more efficient and optimized depending upon the updates in the modules, library, etc.

* 1. REVIEW OF LITERATURE

Since the middle of the 20th century, text summarizing research has been researched. Lun (1958) used word frequency diagrams as a statistical tool to explain the topic in public for the first time. There have been a lot of various strategies developed so far. There are single and multi-document summarizations depending on the document count. The extractive and abstractive outcomes, meanwhile, are based on the summary results.

Qiang et al., 2016, Ansamma et al., 2017, Widjanarko et al., 2018. Christian et al., 2016 made text summarizing in a single document using TF-IDF and (Sarkar, 2013) designed automatic text summarizing in a single document using the Main Concepts.

On the 2004 DUC dataset, Qiang et al. (2016) summarized multiple documents using the pattern- based summarization (Patsum) method, demonstrating that the findings beat both the term-based and the ontology-based methods. In terms of precision and recall, Ansamma et al(2017) .'s summary of multiple documents utilizing Latent Semantic Analysis (LSA) and Non-Negative Matrix Factorization (NMF) performed better than the state of the art.

Finding the place of the sentence and the frequency of terms in the text were the typical issues that emerged from the extractive summarization research at first (Khan and Salim, 2014). (Baxendale, 1958). The Information Extraction (IE) technique was used in the following experiment to address the extraction issue and create a summary with more precise results and greater accuracy. RIPTIDES is one illustration of an automatic summarizing system that was created by utilizing IE techniques and functions to summarize news based on circumstances selected by the user (White et al., 2001). Although it has not yet been tested on larger datasets, research by Naik and

Gaonkar (2017) employing a rule base yields the best average precision, f-measure, and recall values for Rule-Based Summarizers.

Furthermore, extractive summary research employing neural networks have grown more popular recently than traditional methods, including the following studies: (Mohsen et al., 2020, Anand and Wagh, 2019, Xu and Durrett, 2019, Chen et al., 2018a, Chen et al., 2018b, Alami et al., 2019). Anand and Wagh's research from 2019 used the Feed Forward Neural Network (FFNN), a deep learning technique, to summarize a single legal document. FFNN has the benefit of producing an extractive summary without the need for features or domain knowledge, performs well as measured by the Rouge score, and produces a coherent summary, but it struggles to condense complex and lengthy sentences.

The review study by Gupta and Gupta (2019) highlights popular elements directly related to abstractive summarizing, including research trends in the field, a broad explanation of the methods, tools, and assessments that are now in use. Abualigah et al. (2020) did additional reviews and provided a succinct overview of text summarizing methods specifically for Arabic. The strategy, tactics, and methods utilized in text summarizing are the focus of a survey on the topic by Nazari and Mahdavi (2018). The methods to statistics, machine learning, semantic-based, and swarm intelligence were grouped by Nazari and Mahdavi (2018). Elrefaiy et al(2018) .'s research on summarizing extractive texts that concentrate on unsupervised techniques includes a comparative table with a list of pros and shortcomings.

In order to create an algorithm that can summarize a document, Dr. Annapurna P. Patil et al. [2014] extract key text from the document and attempt to change this extraction using a thesaurus. Our fundamental objective is to maintain coherence and semantics while condensing a given volume of text to a fraction of its original size. Automatic summarization is the technique of employing a computer software to condense written content into a manageable format for human consumption.

Samrat Babar (2013a) in his essay emphasized how there is a wealth of textual content on the internet. As a result, finding relevant papers from the many that are available and learning useful information from them is a challenge. Automatic text summary is crucial for resolving the fore-mentioned two issues. Text summary is the act of locating the most crucial and meaningful information in a document or group of linked documents, then

condensing that information while maintaining its main ideas.

Computer networks have provided opportunities for it because N. Moratanch et al. [2017] discussed how most conventional educational forms are no longer suitable for the demands of social progress and educational development and are unable to keep up with changes in learning demands in a timely manner. However, traditional web-based learning modes place system development and upkeep inside of educational institutions or businesses, which causes a number of issues, such as the requirement for significant investment but a lack of funding. Due to the overabundance of data on the web, text summarization has recently gained greater attention. Consequently, this information overwhelms results in a significant need for more capable and trustworthy progressive.

Text summary is the act of creating a synopsis from a given text document while retaining the key details and meaning of it, as explained by Mega Satish et al. in their article from 2021. In order to quickly and efficiently locate important information in large amounts of text, automatic summarization has emerged as a key technique. We suggest implementing a web application that can sum up a text or a Wikipedia link in this project.

* 1. METHODOLOGY

Some existing algorithm using typical NLP working and modules such as wordnet, led their model to get the output i.e., the summarized document in time period of 4-7mins with the accuracy of around 75% estimating in the terms of accuracy, the summarized document was measured to be around 50% of the original document i.e., the input.

The model that we are displayed and shown in this paper has an accuracy of 85.9% with the convert rate of almost 30%. The modules we used are NLP, wordnet, stopwords playing crucial role and also extension is made by hosting it on the web and using it as a web application.

* 1. RESULT & DISCUSSION

World-widely the meeting are being conducted virtually on online mode multiple challenges came into picture such as taking down every points discussed, making notes then to summarize. And, also organizers main ask from the participants was to be in the meet rather than just taking down notes. To overcome such types of problems, we can use NLP to work and summarize the meet. Artificial Intelligence and Machine Learning, one of the fastest hot trend domain and field in the market let us to dive in the

ocean of technology and helps us to overcome every problem and have multiple solution using multiple ways. Transcription s is given by most of the software but then too the user face some issues, here those transcriptions can be handled to work with python again one of the most preferred language by using libraries such as PyAudio, SpeechRecognition, Librosa, PYTTSX3, etc. But then also some issue may be addressed, so to overcome this direct used of the text can be made to convert into summarized text.

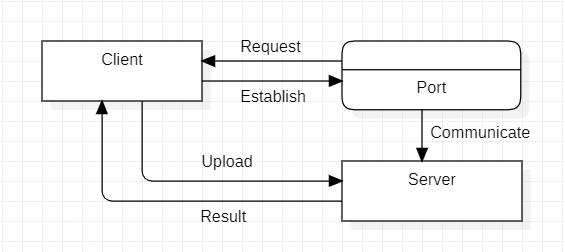
* + 1. DIAGRAM

The model has two modules, and each has its own working:

Server: Server help us to communicate over the browser, by hosting the model over web browser Client: Client will post the query by uploading the

* + 1. DATA FLOW DIAGRAM

They can be used to analyze an existing system or model a new one. Like all the best diagrams and charts, a DFD can often visually “say” things that would be hard to explain in words, and they work for both technical and nontechnical audiences (data flow diagram).



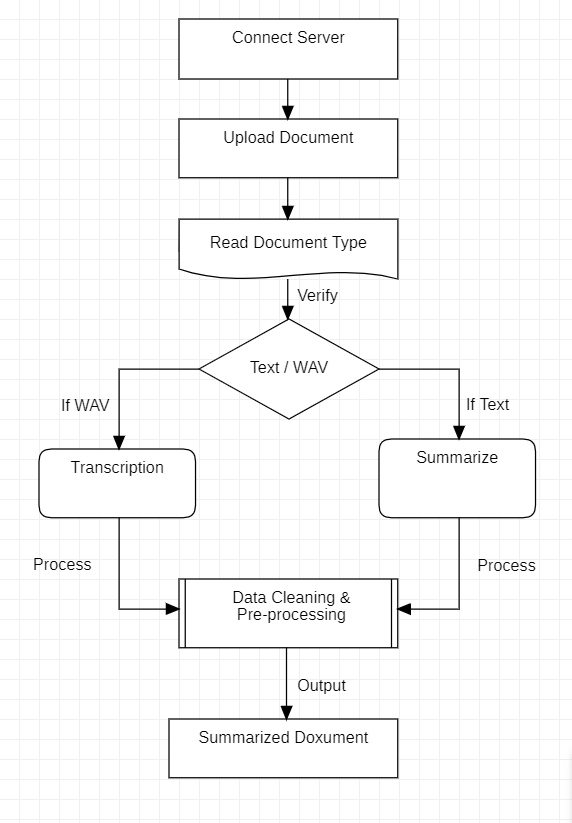
* + 1. USE CASE DIAGRAM

A use case diagram can summarize the details

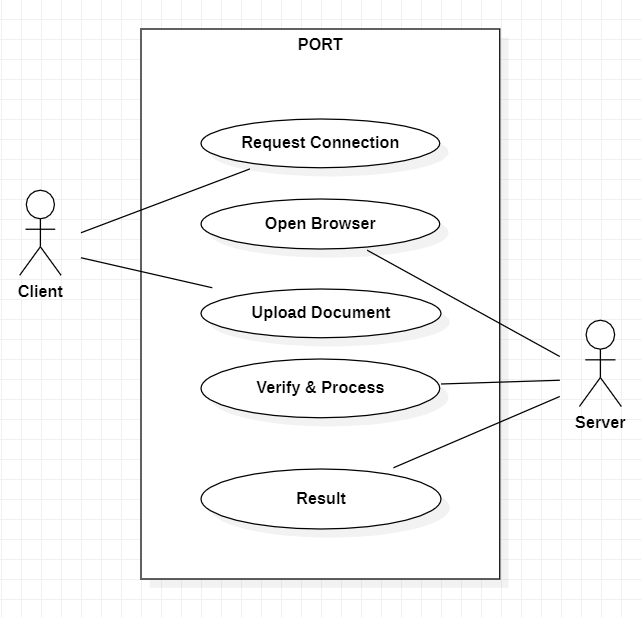
document by selecting Transcription or of your system's users (also known as actors) and their Summarization, based on the selection, model will interactions with the system. To build one, you'll use a

give suitable output.

4.1.1 FLOW CHART

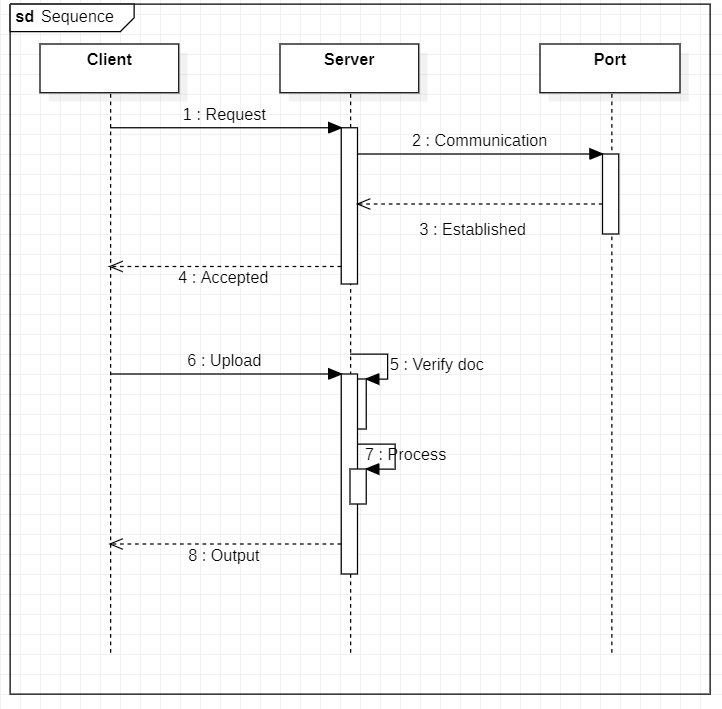
A flow diagram is a visualization of a sequence of actions, movements within a system and/or decision points. They are a detailed explanation of each step in a process, no matter the level of complexity of that process.

set of specialized symbols and connectors.



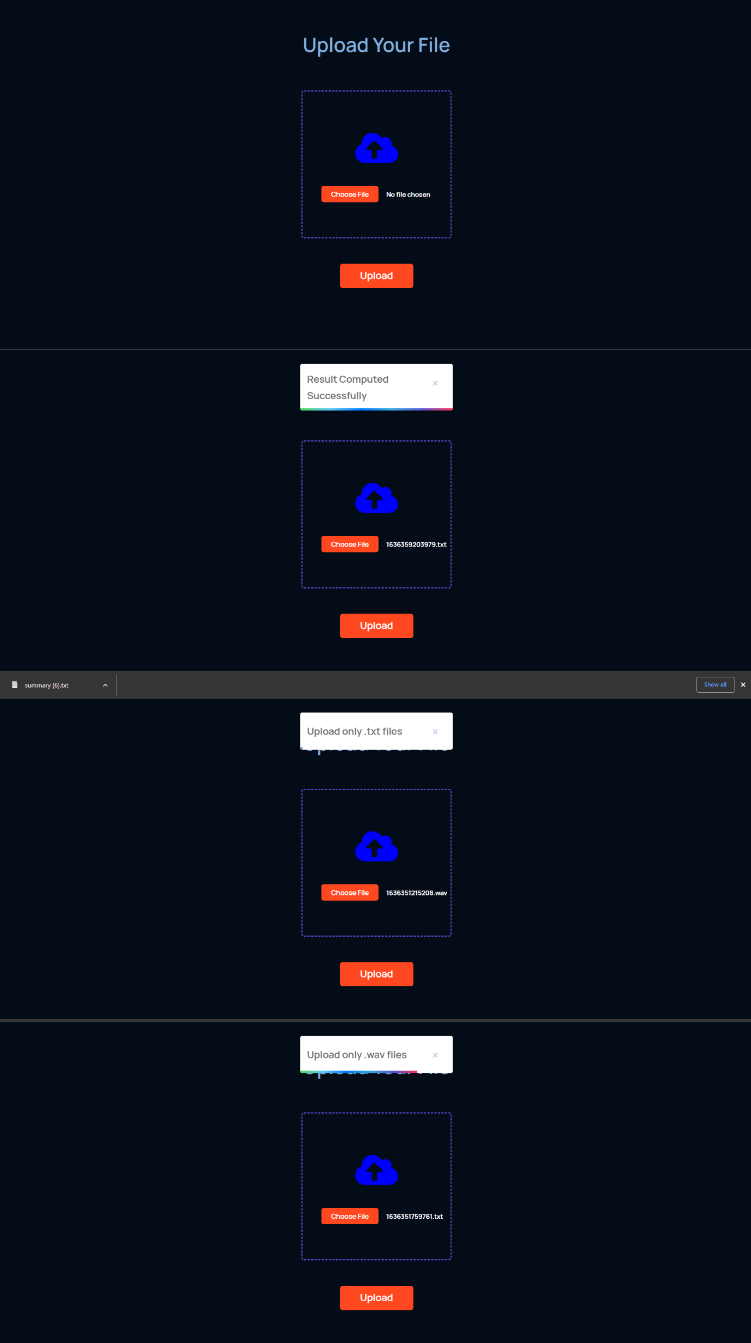
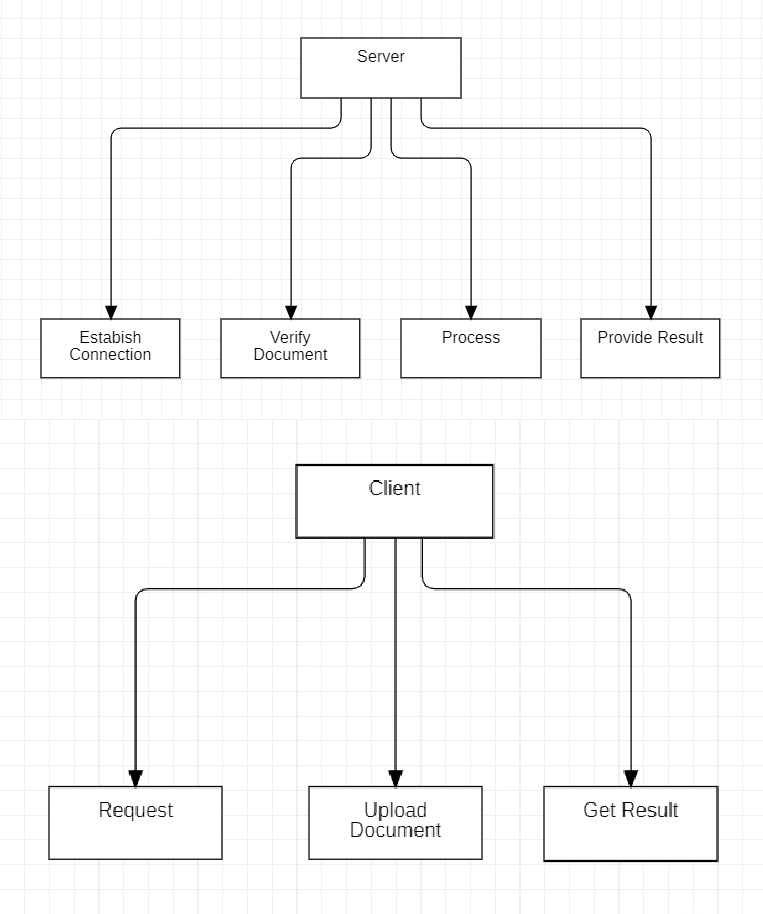
* + 1. SEQUENCE DIAGRAM Sequence Diagrams are time focus and they

show the order of the interaction visually by using the vertical axis of the diagram to represent time what messages are sent and when.

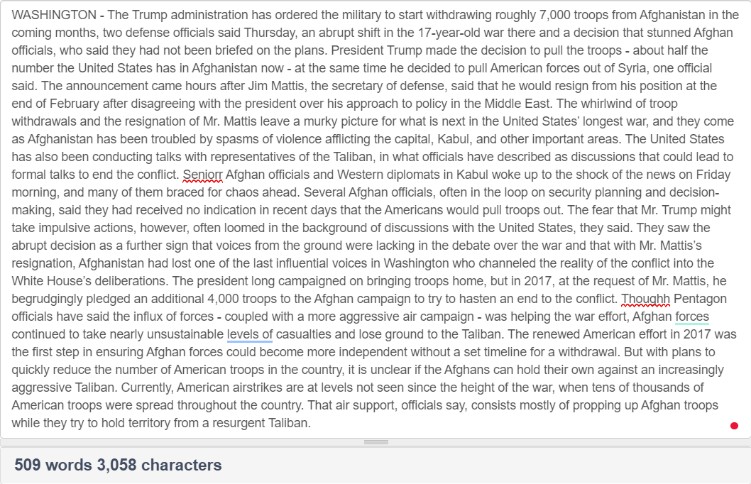


* + 1. MENU TREE

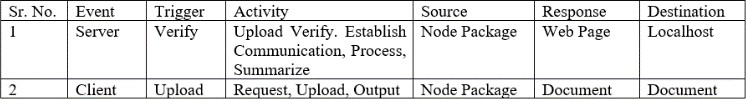
A menu tree known as a tree menu presents data in a hierarchical order. It helps you to group data objects according to the types of tasks, goods, and functions, among other criteria.

* + 1. SERVER
    2. Client

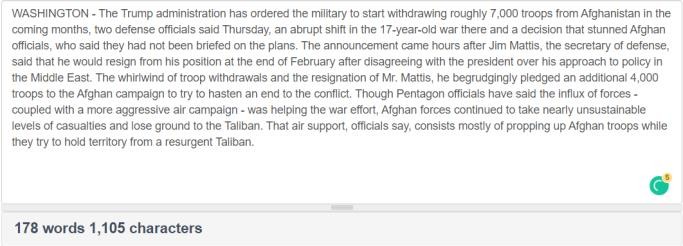
Original Document

* + 1. EVENT TABLE

The event table describes the type of change

made to an application table, and also contains an Summarized Document identifier for the changed row.

4.2 OUTPUT



* 1. CONCLUSION

Hahn & D. Harman (Eds.), Proceedings of the workshop on automatic summarization, 2002, pp.30–38.

[9]. Samrat Babar , Text Summarization:An Overview, published in the year 2013 at <https://www.researchgate.net/publication/257947528>

[10]. N. Moratanch, Chitrakala Gopalan, A survey on abstractive text summarization published by [*https://www.researchgate.net/publication/305912913*](https://www.researchgate.net/publication/305912913)

[11]. Amey Thakur, Mega Satish, “Text Summarizer”, published by

Although text summary is a tried-and-true *https:*[*//w*](http://www.researchgate.net/publication/357152089)*ww*[*.researchgate.net/publication/357152089*](http://www.researchgate.net/publication/357152089)problem, the present study focusses the creation of [12]. Paulus, R., Xiong, C., and Socher, R. (2017). *A deep* meeting minutes, as well as attention to the significant *reinforced model for abstractive summarization*. arXiv

preprint arXiv:1705.04304.

details and summarization of the text that occurred [13]. Rautray, R. and Balabantaray, R. C. (2017). *An* throughout the meeting to a level of 30% of the *evolutionary framework for multi document summarization* original text. Apps, languages, and algorithms *using cuckoo search approach*: Mdscsa. Applied Computing employed in our applications are all specifically and Informatics.

focused on the context in which they are used. Our [14]. Salton, G. and McGill, M. J. (1986). Introduction to

modern information retrieval.

focus is also on new developments in the fields of [15]. Sanchez-Gomez, J. M., Vega-Rodr´ıguez, M. A., and biomedicine, product reviews, education, emails, and Perez, ´ C. J. (2018). Extractive multi-document text blogs. This is a result of the information glut in many summarization using a multi-objective artificial bee colony domains, particularly on the Internet. An important optimization approach. Knowledge-Based Systems, 159:1–

topic of NLP (Natural Language Processing) research

8.

[16]. Sankar, K. and Sobha, L. (2009). An approach to text is automated summarization. It involves assembling a summarization. In Proceedings of the Third International summary of one or more texts automatically. Workshop on Cross Lingual Information Access: Extracted document summarizing aims to choose a Addressing the Information Need of Multilingual Societies,

few representative sentences, chapters, or paragraphs pages 53– 60. Association for Computational Linguistics.

from the source material automatically.

REFERENCES

[1]. N. Moratanch, Chitrakala Gopalan, A survey on abstractive text summarization published by [*https://www.researchgate.net/publication/305912913*](https://www.researchgate.net/publication/305912913)

[2]. Prerana Das, Kakali Acharjee, Pranab Das and Vijay Prasad, “*VOICERECOGNITION SYSTEM: SPEECH-TO-*

*TEXT”* published by Journal of Applied and Fundamental Sciences.

[3]. Annapurna P Patil, Shivam Dalmia, Syed Abu Ayub Ansari,Tanay Aul,Varun Bhatnagar, “*Automatic text summarizer*”, published by International Conference on Advances in Computing, Communications and Informatics (ICACCI) in year 2014.

[4]. Ishitva Awasthi; Kuntal Gupta; Prabjot Singh Bhogal, “*Natural Language Processing(NLP) based Text Summarization - A Survey*” published by 2021 6th International Conference on Inventive Computation Technologies (ICICT).

[5]. A. Khan, N. Salim, and Y. Jaya Kumar, “*A framework for multi document abstractive summarization based on semantic role labelling*,” Appl. Soft Comput., vol. 30, no. C, pp. 737–747, May 2015.

[6]. Dolière Francis Somé, “*EmPoWeb: Empowering Web Applications with Browser Extensions*” published by 2019 IEEE Symposium on Security and Privacy (SP).

[7]. Abbasi-ghalehtaki, R., Khotanlou, H., and Esmaeilpour, M. (2016). Fuzzy evolutionary cellular learning automata model for text summarization. Swarm and Evolutionary Computation, 30:11–26.

[8]. S. H. Finley and S. M. Harabagiu, “Generating single and multidocument summaries with gistexter,” in In U.

[17]. Shareghi, E. and Hassanabadi, L. S. (2008). Text summarization with harmony search algorithm-based sentence extraction. In Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology, pages 226–231. ACM

ABBREVATIONS: - AI- Artificial Intelligence, ML- Machine Learning, NLP- Natural Language Processing, PyAudio- Python Audio, PYTTSX3- Python Text-To-Speech, WAV- Waveform Audio File Format.