# Text Summarizer – On The GO

*Palak Goel, Rithik Raj Vaishya*

*MSc Data Science*

***Abstract*—We were confronted with the biggest problem of the 21st century at the beginning of 2020 – The Pandemic of COVID-19. Online education and digital globalization, where the role of education is almost a daily or hourly phenomenon extended from media to educational institution to any domain of knowledge. The major challenge is that the summarization of the text along with the dynamics of the video is not well structured which is an urgent need for the society. The present work attempted to address a problem related to educational training and knowledge assimilation programs, Online and audio meetings to support and facilitate almost every domain from economy to education to business processing of every domain. To provide a summary of the online meetings, we designed a web and desktop application for this paper. While conducting the validation of existing summarization tools we found that certain applications offer us with the transcript but does not provide a summary of the meeting. In order to illustrate the information gathered across different fields, people frequently miss crucial details on account of the difficulties in recording the video session and they found it laborious, annoying, and boring. Therefore, automatic text summarization is the order of the day.**

**Our methodology involves NLP and related function. The computational devices are being planned to use Spyder, Web Scrapping. The target beneficiaries are from school students to trainers, professionals of any domain. Validation: live validation -the present work cab be validated through a live session and cross validation could extended to extension of any length.**

**Keywords**: Natural Language Processing, Web scrapping, Summarization, Transcript.

## 1. INTRODUCTION

Due to the pandemic, the majority of offices, education began working online, which forced all operations—including holding meetings—to be done electronically. The notes that are taken during a meeting are called meeting minutes. They highlight the important topics being addressed, the motions being made or voted on, and the upcoming tasks. This forced the creation of meeting reports manually, which frequently resulted in inaccurate reports and extremely large reports that made readers lose interest. Therefore, while writing reports, it is important to keep them brief and to include just the information that the STM can accomplish. Since most of the meetings are held online, taking notes, and summarizing them might present a number of difficulties. For instance, a network difficulty, a sound issue, etc. The major goal of the organizer should not be to record everything, but rather to participate more actively in the meeting. To solve this issue, we can use natural language processing to write a meeting summary. We can either provide an exhaustive summary or make suggestions to the organizer to fill in any blanks or omissions.

Computer networks have created opportunities for education because most traditional educational models are currently unable to meet the demands of social advancement and educational development or to quickly adapt to changes in learners' needs.

The major goal of this project is to create a speech-to-text converter that is automated, avoiding human involvement in notetaking and making it simpler and more dependable for meetings to acquire an accurate and condensed summary of the meeting. Utilizing the file or website link for the specific information, we hope to offer several capabilities for summarizing any document. This is useful for summarizing college notes, abstracts, and other important materials. For humans, manually extracting the summary from a lengthy written source is exceedingly challenging. On the internet, there is a wealth of textual content. 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 mentioned two issues.

In this research, we offer an autonomous text summarization system that creates concise summaries using techniques for natural language processing. Both text and speech content can be used to create summaries. Our web program offers both non-transcript files and a summary of the transcript. Automatic summarization is a well-known technique that is used to distil a document down to its essential ideas.

## 1.1 BACKGROUND

Automatic text summary aims to deliver the source text in a concise, semantically rich form. The main benefit of adopting a summary is that it cuts down on reading time. Informative summarizing systems deliver condensed information of the main text and are 20 to 30 percent of the length of the entire text.

Key processes in text summarization: Topic identification, interpretation, and summary creation are the three primary steps that must be taken while summarizing documents.

A. 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.

B. Interpretation: The interpretation stage is necessary for abstract summaries. In this process, many topics are combined to create a broad content.

C. 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:

a) Sentence boundary recognition. The presence of a dot at the conclusion of a sentence in English indicates the boundary of a sentence.

b) Stop-Word Elimination: Usual words without semantic significance. This technique gets rid of the words that appear most frequently in documents, such as articles, prepositions, conjunctions, interrogatives, assisting verbs, etc. Due to their minimal role in the sentence extraction process, stop words are eliminated.

c) Stemming: The goal of stemming is to identify each word's stem or radix, which highlights its semantics. the stemming procedure may have a 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).

d) 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.

e) Keyword extraction: In this step, the keywords from a document are extracted. Here, all words besides stop words are taken into consideration as keywords.

In the processing phase, factors that affect the relevance of phrases are selected, calculated, and then given weights using the weight learning approach. Using the feature-weight equation, the final score of each sentence is calculated. The final summary is chosen from the top-ranked sentences. A crucial component of text summarization is summary evaluation.

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 scoring, which assigns a numerical value to each sentence for the summary, is one technique for finding appropriate sentences. The best sentences are then chosen to make up the document summary based on the compression rate. The compression rate is a crucial component of the extraction procedure that is used to determine the proportion between the length of the summary and the source text. The summary will grow and contain more irrelevant content as the compression rate rises. While the summary becomes shorter due to the compression rate, more information is lost. In fact, the quality of the summary is acceptable when the compression rate is between 5 and 30%.

## 1.2 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

## 1.3 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:

i. 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.

ii. 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.

iii. 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.

## 1.4 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.

## 1.5 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.

## 2. 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.

## 3. 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.

## 4. 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.

## 4.1 DIAGRAM

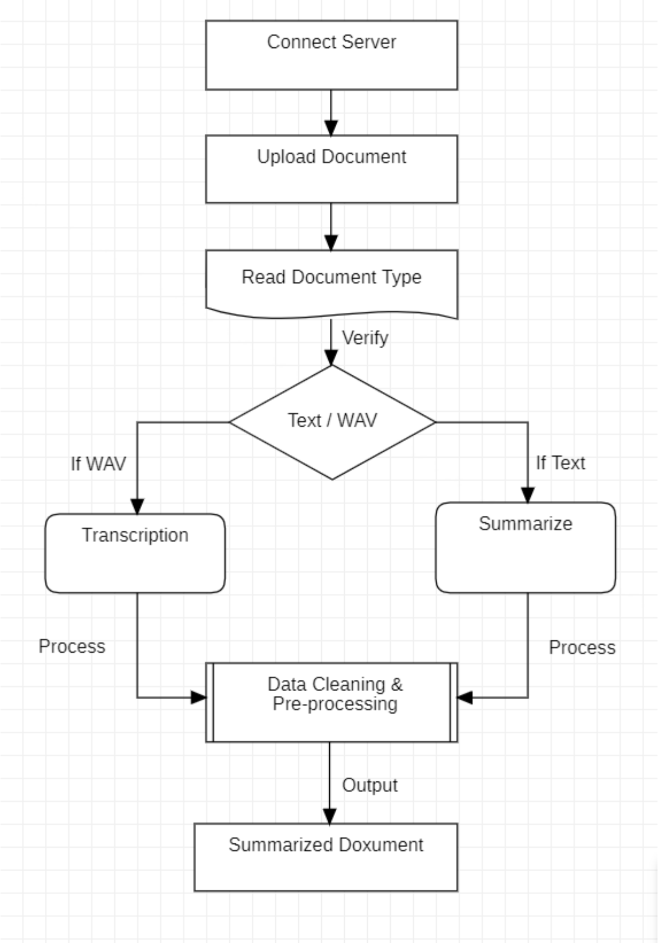
The model has two modules, and each has its own working:

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 document by selecting Transcription or Summarization, based on the selection, model will give suitable output.

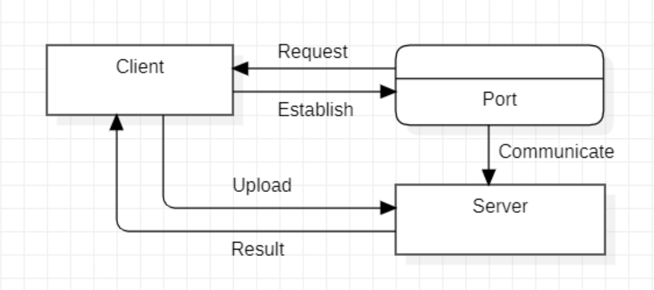
### 4.1.1 FLOW CHART

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.



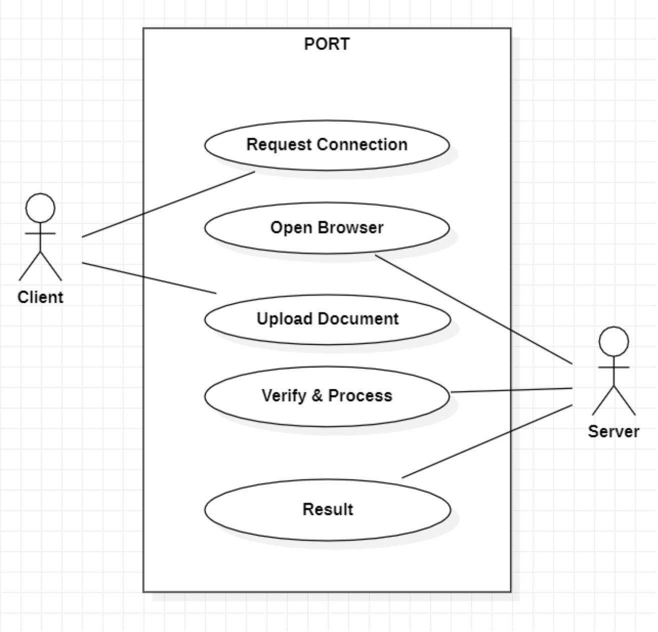
### 4.1.2 DATA FLOW DIAGRAM

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).



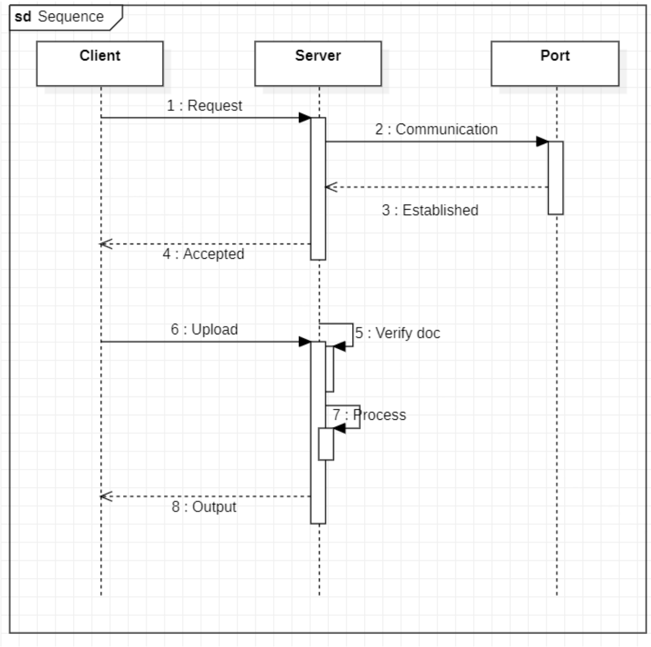
### 4.1.3 USE CASE DIAGRAM

A use case diagram can summarize the details of your system's users (also known as actors) and their interactions with the system. To build one, you'll use a set of specialized symbols and connectors.



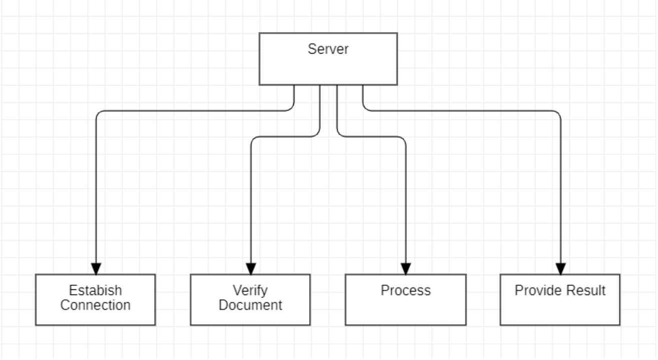
### 4.1.4 SEQUENCE DIAGRAM

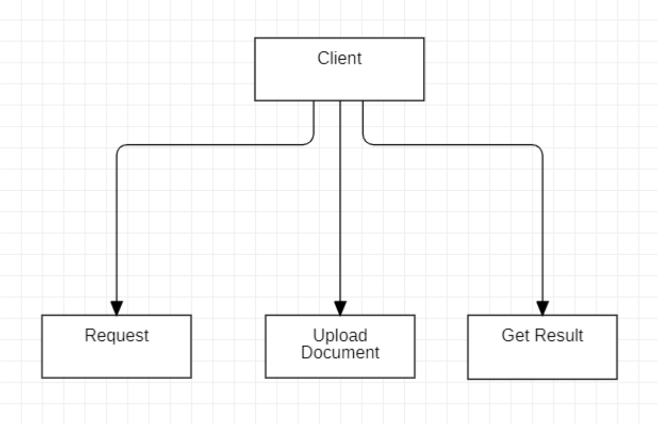
Sequence Diagrams are time focus and they show the order of the interaction visually by using the vertical axis of the diagram to represent time what messages are sent and when.



### 4.1.5 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.





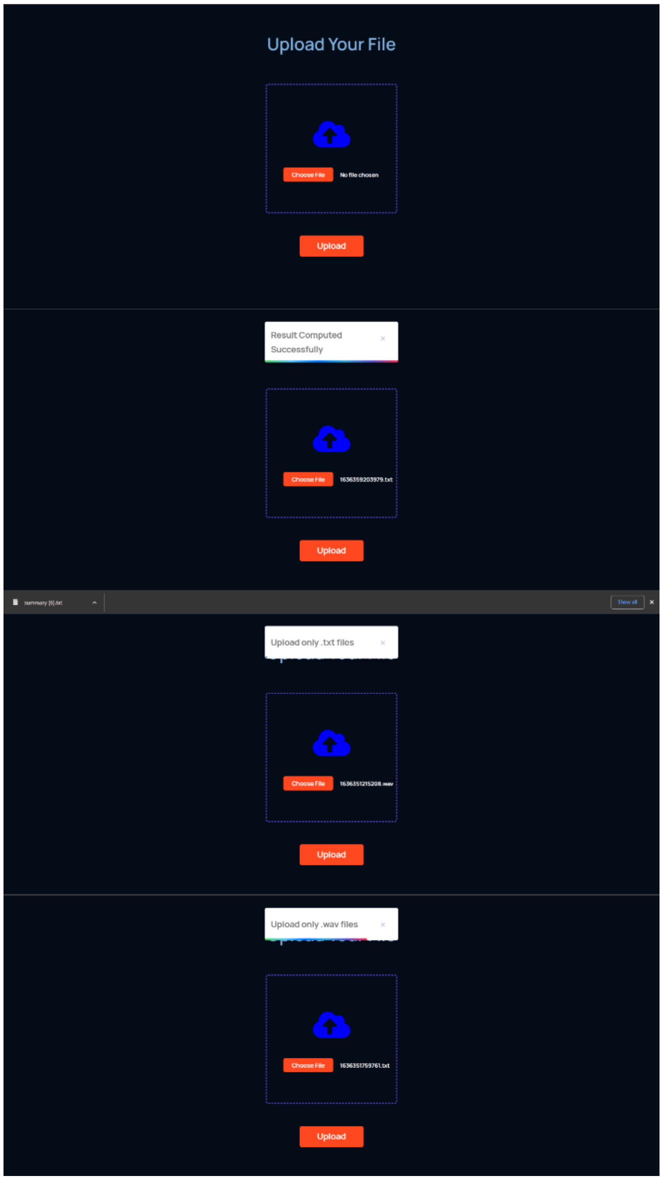
### 4.1.6 EVENT TABLE

The event table describes the type of change made to an application table, and also contains an identifier for the changed row.

| Sr. No. | Event | Trigger | Activity | Source | Response | Destination |
|---|---|---|---|---|---|---|
| 1 | Server | Verify | Upload Verify. Establish Communication, Process, Summarize | Node Package | Web Page | Localhost |
| 2 | Client | Upload | Request, Upload, Output | Node Package | Document | Document |

### 4.2 OUTPUT

#### 4.2.1 SERVER



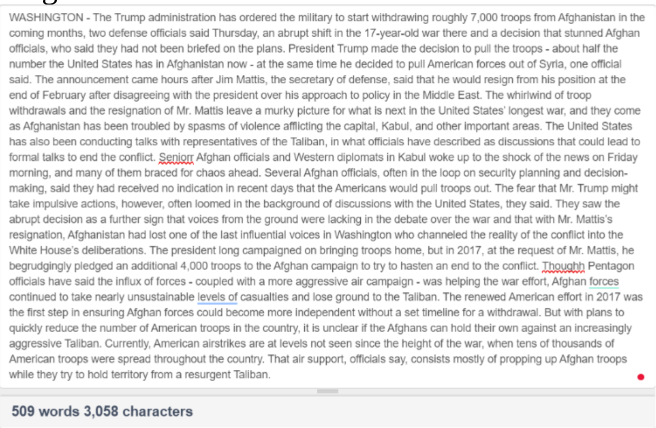
Result Computed Successfully

Upload only .txt files

Upload only .wav files

#### 4.2.2 Client

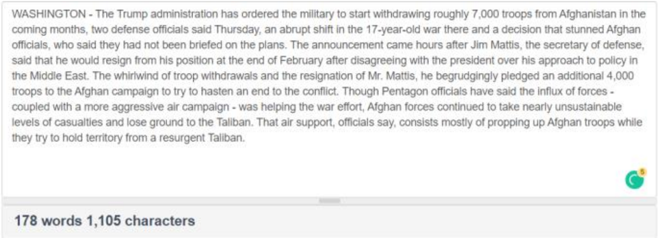
Original Document



WASHINGTON - The Trump administration has ordered the military to start withdrawing roughly 7,000 troops from Afghanistan in the coming months, two defense officials said Thursday, an abrupt shift in the 17-year-old war there and a decision that stunned Afghan officials, who said they had not been briefed on the plans. President Trump made the decision to pull the troops - about half the number the United States has in Afghanistan now - at the same time he decided to pull American forces out of Syria, one official said. The announcement came hours after Jim Mattis, the secretary of defense, said that he would resign from his position at the end of February after disagreeing with the president over his approach to policy in the Middle East. The whirlwind of troop withdrawals and the resignation of Mr. Mattis leave a murky picture for what is next in the United States' longest war, and they come as Afghanistan has been troubled by spasms of violence afflicting the capital, Kabul, and other important areas. The United States has also been conducting talks with representatives of the Taliban, in what officials have described as discussions that could lead to formal talks to end the conflict. Seniorr Afghan officials and Western diplomats in Kabul woke up to the shock of the news on Friday morning, and many of them braced for chaos ahead. Several Afghan officials, often in the loop on security planning and decision-making, said they had received no indication in recent days that the Americans would pull troops out. The fear that Mr. Trump might take impulsive actions, however, often loomed in the background of discussions with the United States, they said. They saw the abrupt decision as a further sign that voices from the ground were lacking in the debate over the war and that with Mr. Mattis's resignation, Afghanistan had lost one of the last influential voices in Washington who channeled the reality of the conflict into the White House's deliberations. The president long campaigned on bringing troops home, but in 2017, at the request of Mr. Mattis, he begrudgingly pledged an additional 4,000 troops to the Afghan campaign to try to hasten an end to the conflict. Thoughh Pentagon officials have said the influx of forces - coupled with a more aggressive air campaign - was helping the war effort, Afghan forces continued to take nearly unsustainable levels of casualties and lose ground to the Taliban. The renewed American effort in 2017 was the first step in ensuring Afghan forces could become more independent without a set timeline for a withdrawal. But with plans to quickly reduce the number of American troops in the country, it is unclear if the Afghans can hold their own against an increasingly aggressive Taliban. Currently, American airstrikes are at levels not seen since the height of the war, when tens of thousands of American troops were spread throughout the country. That air support, officials say, consists mostly of propping up Afghan troops while they try to hold territory from a resurgent Taliban.

509 words 3,058 characters

Summarized Document



WASHINGTON - The Trump administration has ordered the military to start withdrawing roughly 7,000 troops from Afghanistan in the coming months, two defense officials said Thursday, an abrupt shift in the 17-year-old war there and a decision that stunned Afghan officials, who said they had not been briefed on the plans. The announcement came hours after Jim Mattis, the secretary of defense, said that he would resign from his position at the end of February after disagreeing with the president over his approach to policy in the Middle East. The whirlwind of troop withdrawals and the resignation of Mr. Mattis, he begrudgingly pledged an additional 4,000 troops to the Afghan campaign to try to hasten an end to the conflict. Though Pentagon officials have said the influx of forces - coupled with a more aggressive air campaign - was helping the war effort, Afghan forces continued to take nearly unsustainable levels of casualties and lose ground to the Taliban. That air support, officials say, consists mostly of propping up Afghan troops while they try to hold territory from a resurgent Taliban.

178 words 1,105 characters

## 5. CONCLUSION

Although text summary is a tried-and-true problem, the present study focusses the creation of meeting minutes, as well as attention to the significant details and summarization of the text that occurred throughout the meeting to a level of 30% of the original text. Apps, languages, and algorithms employed in our applications are all specifically focused on the context in which they are used. Our focus is also on new developments in the fields of biomedicine, product reviews, education, emails, and blogs. This is a result of the information glut in many domains, particularly on the Internet. An important topic of NLP (Natural Language Processing) research is automated summarization. It involves assembling a summary of one or more texts automatically. Extracted document summarizing aims to choose a few representative sentences, chapters, or paragraphs 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*

[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. 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*

[11]. Amey Thakur, Mega Satish, "Text Summarizer", published by *https://www.researchgate.net/publication/357152089*

[12]. Paulus, R., Xiong, C., and Socher, R. (2017). *A deep reinforced model for abstractive summarization*. arXiv preprint arXiv:1705.04304.

[13]. Rautray, R. and Balabantaray, R. C. (2017). *An evolutionary framework for multi document summarization using cuckoo search approach*: Mdscsa. Applied Computing and Informatics.

[14]. Salton, G. and McGill, M. J. (1986). Introduction to modern information retrieval.

[15]. Sanchez-Gomez, J. M., Vega-Rodr´ıguez, M. A., and Perez, ´ C. J. (2018). Extractive multi-document text summarization using a multi-objective artificial bee colony optimization approach. Knowledge-Based Systems, 159:1–8.

[16]. Sankar, K. and Sobha, L. (2009). An approach to text summarization. In Proceedings of the Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies, pages 53– 60. Association for Computational Linguistics.

[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.