



# TEXT SUMMARIZATION OF COVID DATA USING MACHINE LEARNING TECHNIQUES

Amitha S M, Amulya D M, Deekshitha V Reddy, Deepthi T E  
Computer Science and Engineering,  
K S School of Engineering and Management,  
Bangalore, India

Mrs. Sandhya A Kulkarni  
Assistant Professor,  
Computer Science and Engineering,  
K S School of Engineering and Management,  
Bangalore, India

**Abstract:** The present scenario of COVID-19 demands an efficient face mask detection application. The main goal of the project is to implement these systematic entrances of colleges, airports, hospitals, and offices where chances of spread of COVID-19 through contagion are relatively higher. Reports indicate that wearing face masks while at work clearly reduces the risk of transmission. While entering the place everyone should scan their face and then enter ensuring they have a mask with them. If anyone is found to be without a facemask, beep alert will be generated. As all the work places are opening. The number of cases of COVID-19 are still getting registered throughout the country. If everyone follows the safety measures, then it can come to an end. Hence to ensure that people wear masks while coming to work we hope this module will help in detecting it.

**Index Terms - Coronavirus, Covid-19, Machine Learning, NLP Techniques, Summarization Methods**

## I. INTRODUCTION

With development of technology and its usage the size of data has increased. This bulky amount of data needs to be processed and stored efficiently. In this era, a lot of information is available on the internet. The information is available in lengthy form which takes time to read. This can be stressful to read and is not efficient. Summarizing in a manual way is strenuous job. This is where text summarization technique that comes into play. Text summarization is producing a brief and concise summary by capturing the vital information. Different approaches are present for summarizing the text and few techniques can be used to implement it. There exist many techniques. One such technique is Natural Language Processing (NLP). NLP is a field in Computer Science that focuses on the study where human language and computer interaction is considered. NLP has capacity to summarize large volume text even in real time. There exists two different kind of approaches which are Abstractive summarization and Extractive summarization. In Abstractive summarization, it uses semantic representation and natural language generation techniques to generate summary. Such a summary might not contain words which are explicitly present in the original article. Extractive summarization technique is where words in summary are chosen from the existing original article.

## II. RELATED WORK

In text summarization technique the main requirement is the input article on which the technique can be implemented. The input can undergo various processing and produces summary. There are some of the fundamentals need to be highlighted which are as follows: First is the Text which is the article to be summarized. It can be in file format or URL. Second is the Summarization technique which creates a short summary of any text. Third is Text summarization technique where long summary is getting reduced to a short one still reflecting the same meaning of the original text. Finally, selection of the summarization technique is important. This paper is based on the Extractive summarization which uses scoring technique. Processing is happening in several steps which are described below. User input is received which undergoes pre-processing. Text cleaning takes place which removals unwanted characters, stop words. Tokenization is the next technique which can be categorized into two: Word Tokenization and Sentence Tokenization. Initially sentence tokenization takes place. Word tokenization occurs creating tokens. The frequency for these words is calculated. Then scores are assigned to the sentences. This can be termed as sentence scoring. Selection is a prominent step. From this stage the summary is generated by choosing the important words and sentences. Our work also provides a GUI which helps which takes input and generates summary. Further let us discuss about data input that can be taken, the methods used in our system us in forming a conclusion.

### III. DATA INPUT

Two kinds of data input can be possible. One is where data input is chosen from our file which can be document. This exists in text form. Fig. 1 is file taken as input for summarization.



Fig. 1. Data input added from file

Second is where just URL of the input is added. The data is fetched from there to summarize. In Fig. 2 we can observe URL added to give summary.

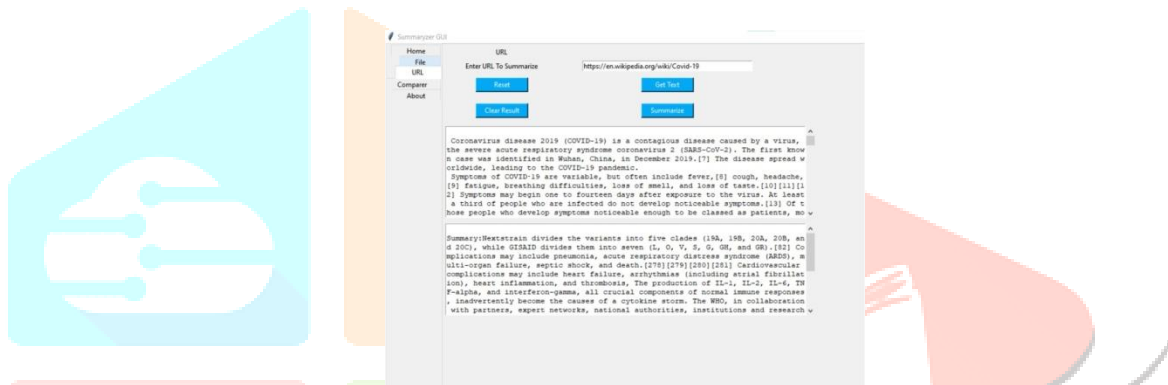


Fig. 2. Data input added using URL

### IV. INCORPORATING PACKAGES

#### A. NLTK

The Natural Language Toolkit (NLTK) is a platform used for building Python applications that work with human language statistics for applying in statistical natural language processing (NLP) it's far one of the maximum effective NLP libraries, which contains programs to make machines recognize human language and reply to it with an appropriate response. It consists of text processing libraries for tokenization, parsing, category, stemming, tagging and semantic reasoning.

#### B. TKINTER

Python offers multiple alternatives for growing GUI (Graphical user Interface). Out of all of the GUI methods, tkinter is the maximum normally used technique. It's far trendy Python interface to the Tk GUI toolkit shipped with Python. Python provides standard library. Tkinter for creating the graphical user interface for desktop based applications. Tkinter provides a effective object-orientated interface to the Tk GUI toolkit. Class matrix in data processing, it helps to compile the overall model.

#### C. Re

Regular expressions are textual content matching styles defined with a proper syntax. The patterns are interpreted as set of commands, which might be then done with a string as input to produce an identical subset or modified model of the original. Expressions can consist of literal text matching, repetition, pattern- composition, branching, and other sophisticated rules. ordinary expressions are typically used in programs that involve a lot of text processing.

### V. THE PROPOSED METHOD

The proposed method includes a Natural Language Processing techniques. It includes nltk and tkinter. the steps in textual content summarization is as follows:

#### A. TEXT CLEANING

Text cleaning consists of techniques to easy the text to prepare it to take care of to the models. The cleaning of text can be accomplished by removing unwanted characters, tokenization and casting off stop words. Stop words are the English words, which does not add a great significance to a sentence. They can be omitted without losing the real importance of a sentence. To remove the stop words, textual content may be divided into phrases and then extract it using the nltk library.

## B. TOKENIZATION

Tokenization is one of the common tasks in textual content processing. It is the process of separating a given text into smaller units known as tokens. An input textual content is a set of multiple phrases which make a sentence. We need to break the textual content in this type of way that machines can apprehend this article and tokenization facilitates us to achieve that. These tokens help in grasping the unique circumstance or fostering the model for the NLP. The tokenization helps in deciphering the importance of the text by examining the arrangement of the words. The most straightforward approach to tokenize text is to utilize whitespace inside a string as the "delimiter" of words. This can be achieved with Python's parted capability, which is accessible on all string object instances as well as on the string built-in in classes. Tokenization is divided into two steps, Word Tokenization and Sentence Tokenization.

## C. WORD TOKENIZATION

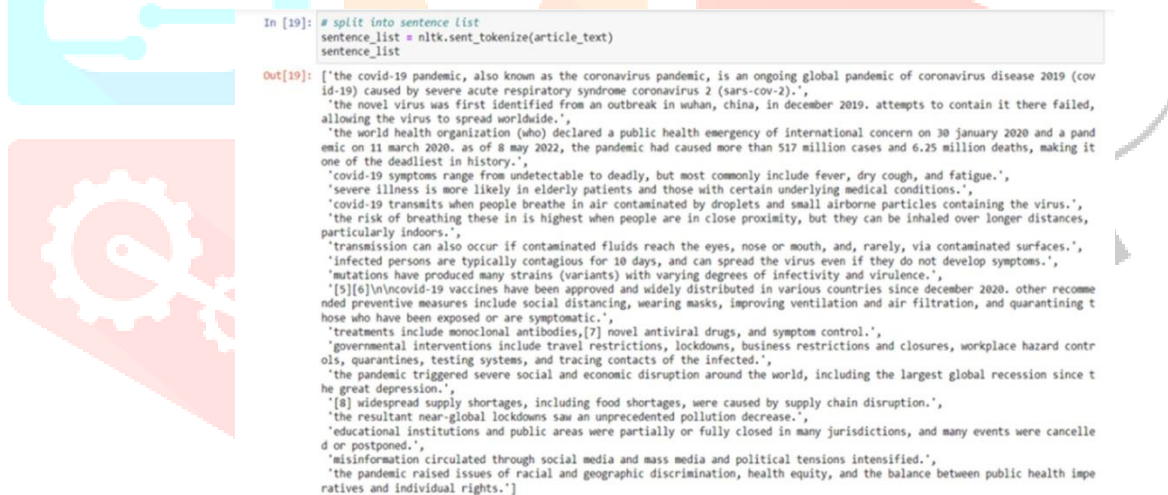
Word tokenization is the process of splitting a big pattern of textual content into words. That is a demand in natural language processing tasks where every word needs to be captured and subjected to similarly analysis like classifying and counting them for a selected sentiment and many others. The Natural Language Toolkit package (NLTK) is a library used to acquire this. For example, whilst you tokenize a paragraph, it splits the paragraph into words known as tokens. Phrases are separated by way of a space, so the manner of phrase tokenization reveals all the areas in a chunk of text to break up the information into words.

## D. SENTENCE TOKENIZATION

Fig 3 depicts the Sentence Tokenization. It means breaking a chunk of textual content into sentences. For example, while you tokenize a paragraph, it splits the paragraph into sentences known as tokens. In many Natural language processing problems, splitting text information into sentences may be very beneficial. Sentences are separated by a complete prevent, so the process of sentence tokenization finds all of the full stops in a piece of textual content to split the information into sentences.

## E. SENTENCE SCORING

The Sentence Scoring is quite possibly of the most involved process in the space of Natural Language Processing (NLP) while chipping away at text based information. It is an interaction to connect a numerical value with a sentence in view of the algorithm priority. This cycle is exceptionally utilized particularly on text outline. This interaction is profoundly utilized particularly on text outline. There are numerous famous techniques for sentence scoring like TF-IDF, Text Rank, etc. Thus by using these different methods of pre-processing techniques, summary could be obtained efficiently and thereby reducing the users reading time with a concise text without ignoring the useful information.



```
In [19]: # split into sentence list
sentence_list = nltk.sent_tokenize(article_text)
sentence_list

Out[19]: ['the covid-19 pandemic, also known as the coronavirus pandemic, is an ongoing global pandemic of coronavirus disease 2019 (covid-19) caused by severe acute respiratory syndrome coronavirus 2 (sars-cov-2).',
'the novel virus was first identified from an outbreak in wuhan, china, in december 2019. attempts to contain it there failed, allowing the virus to spread worldwide.',
'the world health organization (who) declared a public health emergency of international concern on 30 january 2020 and a pandemic on 11 march 2020. as of 8 may 2022, the pandemic had caused more than 517 million cases and 6.25 million deaths, making it one of the deadliest in history.',
'covid-19 symptoms range from undetectable to deadly, but most commonly include fever, dry cough, and fatigue.',
'severe illness is more likely in elderly patients and those with certain underlying medical conditions.',
'covid-19 transmits when people breathe in air contaminated by droplets and small airborne particles containing the virus.',
'the risk of breathing these in is highest when people are in close proximity, but they can be inhaled over longer distances, particularly indoors.',
'transmission can also occur if contaminated fluids reach the eyes, nose or mouth, and, rarely, via contaminated surfaces.',
'infected persons are typically contagious for 10 days, and can spread the virus even if they do not develop symptoms.',
'mutations have produced many strains (variants) with varying degrees of infectivity and virulence.',
'[5][6] unvaccinated covid-19 vaccines have been approved and widely distributed in various countries since december 2020. other recommended preventive measures include social distancing, wearing masks, improving ventilation and air filtration, and quarantining those who have been exposed or are symptomatic.',
'treatments include monoclonal antibodies,[7] novel antiviral drugs, and symptom control.',
'governmental interventions include travel restrictions, lockdowns, business restrictions and closures, workplace hazard controls, quarantines, testing systems, and tracing contacts of the infected.',
'the pandemic triggered severe social and economic disruption around the world, including the largest global recession since the great depression.',
'[8] widespread supply shortages, including food shortages, were caused by supply chain disruption.',
'the resultant near-global lockdowns saw an unprecedented pollution decrease.',
'educational institutions and public areas were partially or fully closed in many jurisdictions, and many events were cancelled or postponed.',
'misinformation circulated through social media and mass media and political tensions intensified.',
'the pandemic raised issues of racial and geographic discrimination, health equity, and the balance between public health imperatives and individual rights.']
```

Fig 3: Sentence tokenization

## F. SUMMARIZED TEXT

After the user gives the textual content in various strategies and plays out the further process for developing a possible and a helpful summary by use different methodologies of information pre-processing techniques include text cleansing, Sentence Tokenization, word Tokenization and Sentence Scoring. The concise and specific summary is acquired without losing the general meaning of the authentic textual content is the summarized text. Fig 4 represents the overall view of the text summarization architecture.

Fig 5 represents the summary that is generated in the GUI using the data pre-processing techniques and the required libraries such as NLTK for obtaining a brief and accurate summary more efficiently.

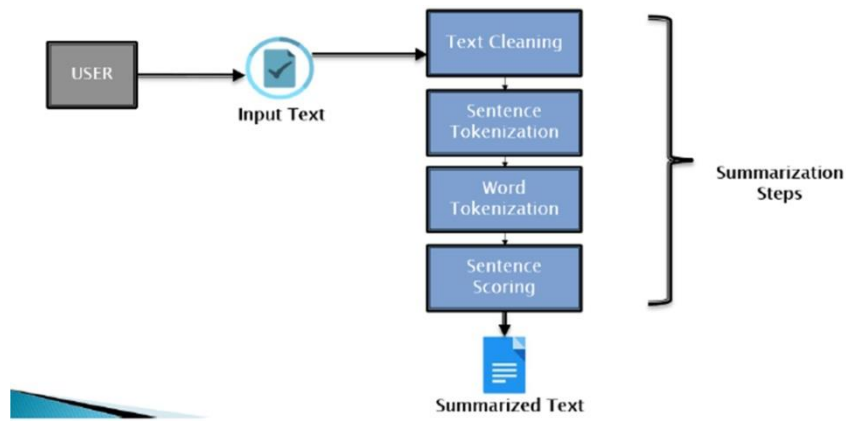


Fig 4: Overview of text summarization technique

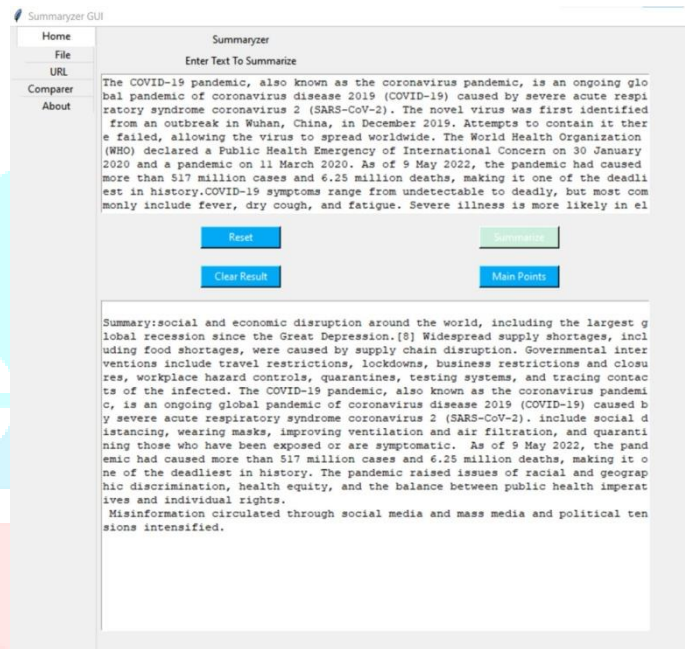


Fig 5: Summary Generated

## VI. RESULT AND ANALYSIS

This technique helps in the better use of the colossal measure of information accessible to us on the Internet and in various documents, outline is a significant strategy. Manual synopsis by specialists is an inordinately difficult and tedious action. Individuals couldn't access, read, or utilize such a major heap of data for their necessities. This paper presents a productive and subjective examination of the various processes utilized for obtaining a useful precise texts. An outline created ought to prevail with regards to consolidating the fundamental subtleties and the primary thoughts of the given text. Extractive synopsis methods can widely investigate the given message semantically, i.e., on sentences, words, catchphrases, and so forth. The main challenges faced by this method mainly comprises of varying topic identifications, assessment and generation of summary. This study also analyzed a few more factors that might be engaged with the composition of outlines and the effect of these variables on the utilization of summaries for perusing perception evaluation.

## VII. CONCLUSIONS

In this paper, we briefly explained the motivation of the work at first. Indeed, even before the commencement of web papers were perused and features were made utilizing markers. The fast-moving era of technology summarization helps user to save time and to make a quick read. An opinion can be made by reading the summary in short time for books and novels. Newsletter can be understood easily by reading summary which will have the highlights of the information. It helps the reader to decide if the original text is worth reading in full or not. Summary can be generated using the extractive approach which is simple and easy. The article can be of any field like medical, sports, book, website, technology etc. The whole summarization system is taking place online so there is no need for paper making it eco-friendly.

Summarization of report should be possible in two potential ways. Extractive summaries and Abstractive summaries. Abstractive procedure is which produces outline which is semantically related however it is hard to fabricate. Extractive method is basic, simpler to assemble and is more proficient. It is exceptionally intelligible, firm and less excess. NLTK library is utilized for handling text string by string. Computers have been made to gain data from on the web sources and apply what they have realized in genuine world. With blend with natural language generation, PCs have acquired capacity to get and give guidelines.

Text summarization is an interesting research topic among the NLP community that helps produce concise information. Man-made intelligence can assist people with focusing on their time really in the event that it can do an initial pass of perusing the

examination and giving a decent outline. Anyway synopsis will in general be a troublesome and emotional undertaking. To really sum up AI needs to comprehend the substance which is troublesome because of the enormous variety recorded as a hard copy styles of specialists. Text summarization can also be aided in various other divisions including medical history of patients especially during the Covid-19 crisis which had a worst effect all over the world, media observing, lawful agreement examination, monetary research, Question addressing and bots , meeting and video conferencing and video scripting.

## REFERENCES

- [1] JUGRAN, SWARANJALI, ASHISH KUMAR, BHUPENDRA SINGH TYAGI, and VIVEK ANAND. "Extractive automatic text summarization using SpaCy in Python & NLP." In 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), pp. 582-585. IEEE,2021.
- [2] Kumar,G.Vijay,ArvindYadav,B.Vishnupriya,M.NagaLahari,J. Smriti, and D.Samved Reddy. "Text Summarizing Using NLP." Recent Trends in Intensive Computing 39(2021):60.
- [3] Awasthi, Ishitva, Kuntal Gupta, Prabjot Singh Bhogal, Sahejpreet Singh Anand, and Piyush Kumar Soni. "Natural Language Processing (NLP) based Text Summarization-A Survey." In 2021 6th International Conference on Inventive Computation Technologies (ICICT), pp. 1310- 1317. IEEE,2021.
- [4] Awasthi, Ishitva, Kuntal Gupta, Prabjot Singh Bhogal, Sahejpreet Singh Anand, and Piyush Kumar Soni. "Natural Language Processing (NLP) based Text Summarization-A Survey." In 2021 6th International Conference on Inventive Computation Technologies (ICICT), pp. 1310- 1317. IEEE,2021.
- [5] Awasthi, Ishitva, Kuntal Gupta, Prabjot Singh Bhogal, Sahejpreet Singh Anand, and Piyush Kumar Soni. "Natural Language Processing (NLP) based Text Summarization-A Survey." In 2021 6th International Conference on Inventive Computation Technologies (ICICT), pp. 1310- 1317. IEEE,2021.
- [6] Widyassari, AdhikaPramita, SupriadiRustad, GuruhFajarShidik, Edi Noersasongko, Abdul Syukur, and AffandyAffandy. "Review of automatic text summarization techniques & methods." *Journal of King Saud University-Computer and Information Sciences*(2020).
- [7] Ermakova, Liana, Jean ValèreCossu, and JosianeMothe. "A survey on evaluation of summarization methods." *Information processing & management* 56, no. 5 (2019):1794-1814.
- [8] Aries,Abdelkrime,andWalidKhaledHidouci."Automatictextsummarization:Whathasbeendoneandwhathastobedone ." *arXiv preprint arXiv:1904.00688*(2019).
- [9] Allahyari, Mehdi, SeyedaminPouriyeh, Mehdi Assefi, SaeidSafaei, Elizabeth D. Trippe, Juan B. Gutierrez, and KrysKochut. "Text summarization techniques: a brief survey." *arXivpreprint arXiv:1707.02268*(2017).
- [10] Munot,Nikita,andSharvariS.Govilkar."Comparativestudyoftextsummarizationmethods." *InternationalJournalofComputerApplications* 102, no. 12(2014).

