iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** V **Month of publication:** May 2022

DOI: https://doi.org/10.22214/ijraset.2022.42574

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue V May 2022- Available at www.ijraset.com

Survey on Content Modeling Using Natural Language Processing

Bhavana S A[1], Harshitha Vibhu N V[2], Neha N[3], Ms. Namratha B S[4]
[1, 2, 3]*Department of Computer Science and Engineering, Jyothy Institute of Technology, Bangalore*
[4]*Assistant Professor, Department of Computer Science and Engineering, Jyothy Institute of Technology, Bangalore*

Abstract: ***Content modelling is the procedure of formation is a compressed version of a text document which makes sure the definition or the meaning of the information isn't changed from the meaning of original text. Automatic text summarization turns into a beneficial method to findsignificant information accurately in lengthy text in a little amount of time with negligible trouble. Automatic text summarization has taken part in an essential part in assisting users to acquire important key details by adding a large amount of data while also has an advantage of developed technology. Previously, some of the other papers are interconnected in solving the issue to stir up brief outline of the content by using machine learning (ML). Content modelling using natural language processing, it is a prominent requisition exert to bring out suitable details with narrowing huge amount of text. Prevailing observation show that it usescodeword rooted methods to classify text, who doesn't give the file a real idea. We granted effective entity modelling method joins basic information about the data along with the outline inherited text. Besides, merging the features of the text abstract, shows competent outcome contrast to the synopsis given by controlled replicas of abstractive and extractive summarization.***

I. INTRODUCTION

At these times, between the time of owning immense knowledge, written data is swiftly booming which is attainable in various different dialects. Usually as a consequence of time constraints we won't be able to go through each part of the data that that can be easily acquired. In this accelerating planet, it's inconvenient to browse all the literary information. detect appropriate information, the user should put up with entire file which cause excessive data issues which ends up wasting a large amount of hard work and time. For decoding this scenario, the necessity of a programmed content summarization when it arrives to obscurity. This process we will be ability to summarise the situation information by which makes it lesscomplicated to take in the information, abiding with the material, and acknowledging the content. This paper brings forward a network of Optical Character Recognition (OCR) that pulls out information from the transmitted data. The main motive of Optical Character Recognition the is making an updatable information from documents or picture files that are existent. The Optical Character Recognition works on analysing the sentences to save the form of the document. So, a client can take a picture of the text that the user needs to print, uploads the picture in an Optical Char acter Recognition (OCR) program and the model would create an editable content record for the client which is amendable. This data can be taken advantage todistribute or print the necessary text. Now and again, unaltered data which is being handled by analysing and understanding the process.

Our devised plan deals with three types of data being entered, which are classified as:

1) Image Content.
2) Raw text Content.
3) Document Content.

II. LITERATURE SURVEY

1) It's impossible to ingest all of the vast volumes of information available online due to timeconstraints. As a result, it's critical to be able to summarise a document so that it's easy to read while keeping the information's essence and quality.

a) *Preprocessing:* Text is split by punctuation, stop words, and remaining words, which arethen held in a list with the matching sentence.

b) After that, we have a weighted score for each sentence in the paper. We tend to notice the sentences with the finest weights and display them in the document in the order in which theywere written. The current extractive outline is beneficial for certain document formats. The abstraction improves readability and length by a small margin. It doesn't require any special procedures and runs entirely on algorithms.

c) The algorithmic program TextRank has been implemented.

2) Content writing using automatic text report is a vital method of locating pertinent information data in a colossal font in a short period of time very minimal effort. This paperconsists of a survey of assorted ways identified. ways identified:

a) *Extractive Methodology*

- Symbolic logic in text report
- Inverse Document Frequency methodology
- Frequency of Term Cluster-Based technique
- Text report with Neural Network

b) *Methodologies that are Abstract*

- Tree-Based methodology
- Structured primarily based Approach
- Ontology-Based Method
- Example primarily based Method

This paper gives us extractive and theoretic summarization methods. Report system ought tomanufacture a practical outline in a very less amount of time with negligible redundancy having accurate sentences.

3) With the accelerated advance of technology, several literatures are relating to solving theproblem of creating short summaries by utilising artificial intelligence to assist people in receiving essential information from expanding large volumes of data (AI).

Creating an artificial intelligence (AI) text accounting system using three models: applied math, machine learning, and deep learning, as well as measuring the performance of the threemodels. Titles and abstracts of essays are used to train computer science deep learning models to create potential titles, which are then evaluated by ROUGE.

This page benefits from being projected by an AI automatic text account, which uses deep learning to build fast abstracts and summaries from the current titles

4) Matter knowledge is currently gaining popularity and is available in a variety of languages. It's difficult to explore all of the literature info in today's environment whereeveryone is moving at a breakneck speed. As a result, the text report is getting a lot of attention.

Process, extraction, and pre-processing are the three primary aspects of this system. The pre-processing section includes segmentation, tokenization, and the elimination of stop words. Sentence position, sentence length, numerical data, presence of quote, and keywords in the sentence are examples of feature extractions that can be employed in thissystem. During the extraction phase, a machine learning algorithm is used to determine ifthe sentence should be wrapped in outline or the coaching set was not supported. The sentence is then assigned a rank based on its sentence score. The summary considers the most hierarchical sentences first.

This document provides an overview of text report formats for a variety of Indian and foreign languages, including European, French, and others. It also hinted that some reading will be done in order to summarise Indian language content using machine learning approaches.

5) Enhanced technique of automatic text report for net contents mistreatment semanticrelated terms proposes associate in nursing improved extractive text summarization techniques using lexical chain for documents by enhancing the traditional lexical chain method to form finer applicable data strategies identified:

a) The NEWSUM algorithm: To extend accuracy we've got to use agglomeration so weare able to avoid unrelated documents.

b) Position score rule to rank the sentences: To rank the extracted sentences we have atendency to use position score rule The counselled framework is undergoing development. presently taken results are giving positive final result from planned system.

c) Clustering with trigonometric function similarity rule for sentence extraction: To avoid single domain summarization that's algorithm solely works on specific documents we have a tendency to suppose to use cosine similarity algorithm whichsupplies regardless of the type of document or its size, superior sentence extractionresults are obtained.

6) Due to the vast amount of data that has become inflated on the internet, it is difficult to allow the user to browse all of the material available on the internet. Summarization techniques have grown ubiquitous, allowing users to save time by avoiding the need to readall of the information available on the internet.

After the initial pre-processing, give each sentence in a text with supported features (both linguistic and statistical) the following score:

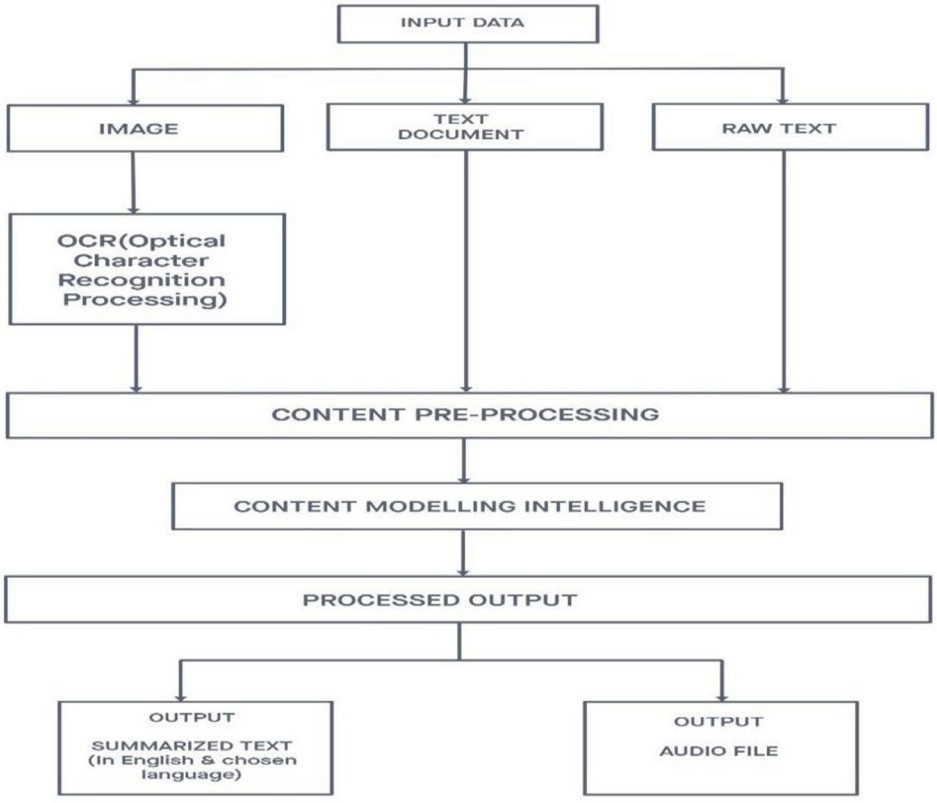
a) We next use all feature weights to calculate the whole weight of each word (Twt) in a phrase. TermFre quencywt(word) Twt(word)=(TermFre quencywt(word)) /Total number ofpossibilities

b) Assign a weight to each word based on how important it is.c.Sum the total weights of eachword to get the sentence's overall score. (Twt() + Twt() +Twt()+)/ Score
(sentence)=(Twt() + Twt() +Twt()+ .....)/ (Total no of words)

c) Calculate the mean weight of all sentences to determine the edge price for selecting essential sentences. We have a tendency to only select sentences that meet a certain thresholdvalue. We offer an automatic text summarisation technique that uses ordered thresholds to locate the outline, or essential sentences, from a given input text source. The number of sentences in the outline would be limited only by the number of paragraphs in a text document, which could be achieved using our ordered threshold approach.

III. PROPOSED METHODOLOGY

The working methodology of this model uses many sub-modules as shown in Figure. Oncethe user chooses the mode of input, the data is first pre-processed, then fed to the content modelling intelligence to generate the summary and then the processed output is translatedand finally converted into an audio file.



Content summarisation is that the condensation of the authentic report right into a compact and important precise reflective the vital content material of the report while not dynamic thewhich implies of the content material with within the report. The term "content summarization" is extensively classified into the extraction of summarization and theoreticalsummarization.

1) *Abstractive Summarization:* Abstractive account produces a generalized outline by growing last sentences sort of a man or ladies that' apothegmatic and brief. outline could comprise cutting-edge expressions that aren't accessible withinside the availability content material. theoretical content material summarization is that the rephrasing of the sentences in the supply matter content within which it receives the foremost important principles within the document and then passes on the most intensive knowledge from the preliminary content material report.
2) *Extractive Summarization:* The Extractive altogether} totally account technique chooses instructive sentences from the document as they precisely show up in offer primarily based totally on distinctive standards to create a summary. the foremost necessary venture before the purpose of extractive summary is to work out that sentences taken from enter are the documents noteworthy and all told chance to be protected within the outline For this function, sentence evaluation is applied primarily based totally on traits of the sentences. initial it' aiming to assign a rating to each sentence primarily based totally on traits then it positions the sentences agreeing to their rating. Sentences with the maximum noteworthy rating are all told chance to be integrated within the ultimate summary.

We use extractive method over the abstract method for the following reasons:

a) Summary illustration is that the sensible sized problem of the theoretic method.
b) In many instances AN abstractive outline will currently not specific the linguistics appealbetween final phrases withinside the record.
c) For growing generalized precis language Era laws are deeply required.
d) Abstractive summaries from time to time originate incoherence.
e) Abstraction entails semantic interpretation of text.
f) The most popular of abstractive precise depends upon at the deep linguistic skills. The CMI version makes use of Extractive account approach within which The following sentences have been chosen as informative sentences. a precis primarily based entirely on precise criteria.

The important venture in terms of extraction account is intended to select that sentence derived from enter record is noteworthy and should be blanketed within the precis. For thisapproach of summarization, the version makes use of the subsequent techniques and packages.

IV. CONCLUSION

The goal of the initiative is to achieve a comprehensive model to acquire a compelling summary with very little repeatability and sentences that are grammatically sound from the reference document. This method has proven effective for all summarization purposes. This Content Modelling Intelligence model primarily demonstrates 3 modules include raw data, image data, and a single text document. The user first enters the raw input data, which must be summarised into a more straightforward and condensed version in both English and language derived from the drop-down menu. The length of the raw input data is mentioned and displayed and the length after the summarization is also displayed . The summary in the chosen language is additionally converted into an audio file which can be accessed by the enduser.

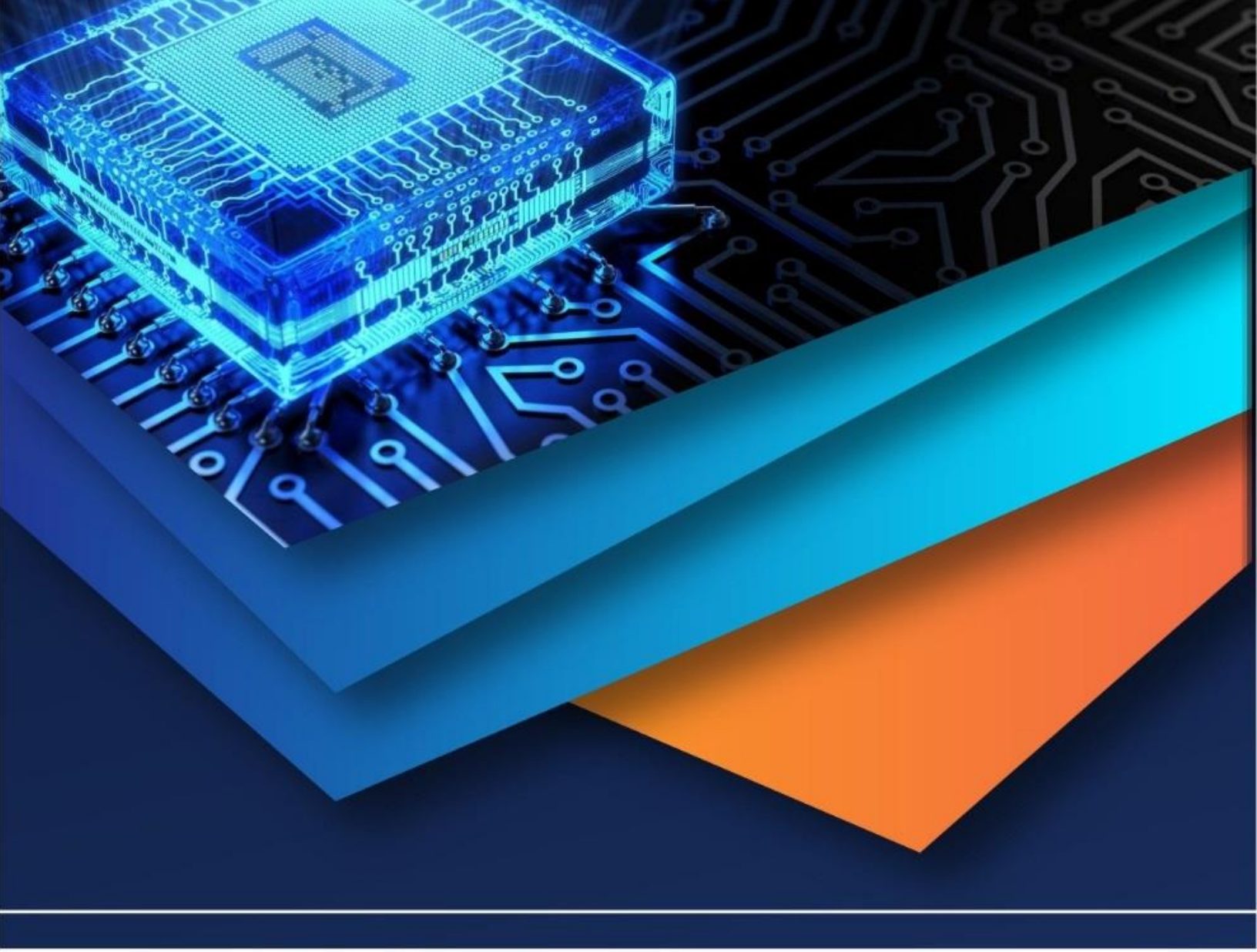
Secondly, Using Optical Character Recognition the uncooked image that the user has uploaded is then condensed and a condensed version is provided in both English and any other language other language option accessible within the model. The condensed summarized version can the later be converted into an audio file which can also be downloaded by the user therefore showcasing text-to-speech feature.

The user will have the option and flexibility to choose any language required for summarization and uploaded document will be summarized to that particular language. Afterthe modelling of the document and the data contained within it the output language will be the language chosen by the user and also in English. This can also be obtained in the form ofan audio file which is available to the user to download.

The present process of summarization favours particular document formats. This particular framework should produce an apt summary of the given input data with proper grammar andlittle to no repeatability within a short period of time. Our next objective is to be able to read and take in multiple input raw data and to also provide summaries of the different links that are copied by the user. As there are advances in the area of Artificial intelligence, Machine learning and Natural language processing necessary changes can be incorporated in the modelfor increasing efficiency and the accuracy of the process.

REFERENCES

[1] Dr. Annapurna P Patil, Shivam Dalmia, Syed Abu Ayub Ansari, Tanay Aul, Varun Bhatnagar, " Automatic Text Summarizer ", International Conference on Advances in Computing, Communications and Informatics, IEEE (2014).
[2] Narendra Andhale, L.A.Bewoor, " An Overview of Text Summarization Techniques " ICACCI (2016).
[3] Min-Yuh Day, Chao-Yu Chen," Artificial Intelligence for Automatic Text Summarization ", IEEE International Conference on Information Reuse and Integration for Data Science (2018).
[4] Prachi Shah, Nikitha P. Desai "A Survey of Automatic Text Summarization Techniques for Indian and Foreign Languages ", International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) (2016).
[5] E.PadmaLahari, D.V.N.Siva Kumar, S. Shiva Prasad, " Automatic Text Summarization with Statistical and Linguistic Features using Successive Thresholds ", IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT) (2014).
[6] Keerthana P "Automatic Text Summarization using Deep Learning" EPRA International Journal of Multidisplinary Research (2021)
[7] Prakhar Sethi, Sameer Sonawane, Saumitra Khanwalker, R. B. Keskar, "Automatic Text Summarization of News Articles ", International Conference on Big Data, IoT and Data Science (BID) Vishwakarma Institute of Technology, Pune, Dec 20-22 IEEE (2017).
[8] Pandu Prakoso Tartan, Alva Erwin, Kho I Eng, Wahyu Muliady, Akon Teknologi," Automatic Text Summarization Based on Semantic Analysis Approach for Documents in the Indonesian Language ", Institute of Electrical and Electronics Engineers IEEE (2013).
[9] Ishitva Awasthi, Kuntal Gupta, Prabjot Singh Bhogal, Sahejpreet Singh Anand,Piyush Kumar Soni "Natural Language Processing (NLP) based Text Summarization" Published by IEEE(2021)



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY