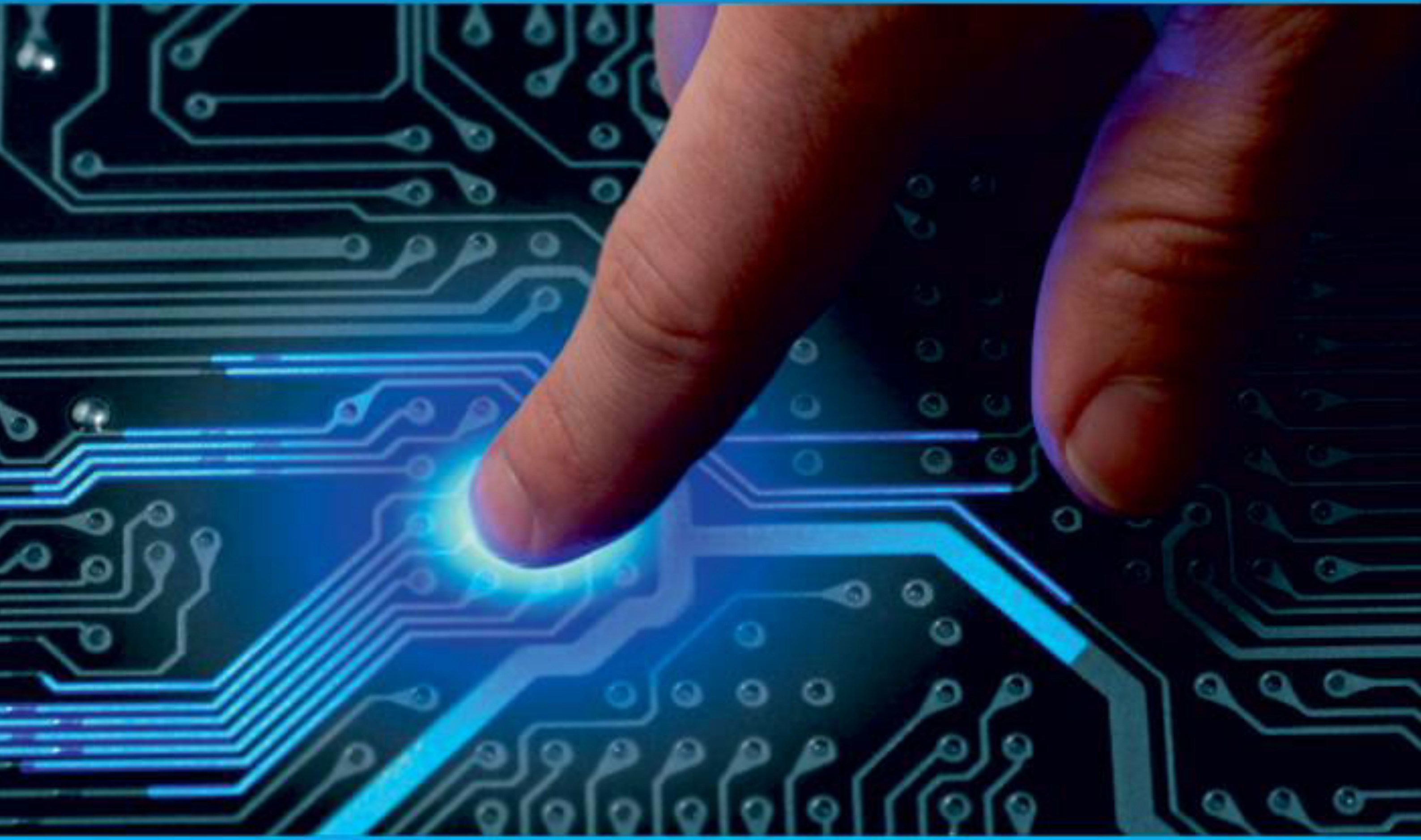
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 4, April 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

Automatic Legal Document Summarization Using NLP

Komal Gaikwad[1], Rashmi Ranmale[1], Shubhangi Dhanavate[1], Suresh Waghmode[1], Mr. S.R Warhade[2]

UG Student, Dept. of I.T., SCTR'S Pune Institute of Computer Technology, Pune, Maharashtra, India [1]

Assistant Professor, Dept. of I.T., SCTR'S Pune Institute of Computer Technology, Pune, Maharashtra, India [2]

ABSTRACT: Summary of text is one of the most important issues in today's world, and it should be addressed in the light of growing data. Our project can create automated summaries that can play a major role in this type of extraction problem. Manual abbreviations are thought of as a way of selecting information using incorrect vocabulary that does not have the correct statements. Returned sentences can be used as a basis for summaries with the help of excavation tools. The role of the project in extracting information from legal documents is largely based on the legal analysis of legal documents.

KEYWORDS: Summarization; extraction of information; semantic processing; TF-IDF

I. INTRODUCTION

The purpose of the summary is to provide the reader with an accurate and complete reading of the source content. In this paper, we often refer to an issue known as an official text summary. As large numbers of documents are provided electronically, the interest in automated summaries has continued to grow in recent years. In this paper, we tend to provide our method of producing short text from long legal text that ends up in our database. Our approach investigates the removal of the most important units based on the identification of text structures and the determination of the important role of the text units within the judgment.

II. RELATED WORK

Jiawen Jiang, Haiyang Zhang, Chenxu Dai, Qingjuan Zhao, Hao Feng, Zhanlin Ji, and Ivan Ganchev [1] introduced four ATS novel models in a Sequence-to-Sequence format (Seq2Seq), using Long-based bidirectional Long Short- Time Memory (LSTM), with additional additions to increase the interaction between the generated text summary and source text and to solve unregistered word problems, overcome duplicate words and avoid an increase in compound errors in created text abbreviations. In this paper Experiments conducted on two public data sets have confirmed that the ATS models introduced achieve better performance than the basics and some of the technologies considered.

RavaliBoorugu, Dr. G. Ramesh [2] has described in detail some of the remarkable works in the field of textual summary. Summary has always been important for many years as there is a lot of information that is published online every day. This paper describes all the important methods of summarizing and the outstanding activities performed in each process. There have been improvements in the past process that have improved accuracy such as a single document summary with greater accuracy compared to multiple documents summary and domain-specific access achieves greater accuracy compared to a process without prior domain knowledge.

Virender Dehru, Prideep Kumar Tiwari, Gaurav Aggarwal, Bhavya Joshi and Pawan Kartik. [3] describe different techniques for summarizing the text. For.e.g., extractive-based, and abstractive-based text Summarization. They also explain the pros and cons of using automatic text summarization.

Shengli Song, Haitao Huang, TongxiaoRuan [4], described the latest ATSDL-based LSTM-CNN model overcoming a number of key issues in the TS sector. Current types of ETS are concerned with syntactic formations, while current forms of ATS are concerned with semantics. The model uses the power of both summarizing models. The new ATSDL model first uses a phrase-extraction method called MOSP to extract key phrases from the original text and learns compound phrases.After training, the model will produce a sequence of phrases that meet the requirements of the synthetic structure. In addition, they use the location information of the phrases to solve the problem of unfamiliar words that almost all ATS models may encounter. Finally, they did extensive research on two different databases and the result shows that our model exceeds high-level approaches in both semantics and syntactic structure.

Subash Voleti, Chaitan Raju, Teja Rani, Mugada Swetha [5], focused on creating a system that finds summaries of articles or a large collection of texts. Completing this could mean that data collection will be easier and will save a lot of time. This project will reduce the time required to read and provide a summary saving lot of time. To get efficient output, unsupervised text algorithm was implemented. Access time for searching the information will be upgraded. Converting short text into audio-file is used in various real-time environments. API functionality is also displayed.

Dr. Majharul Haque, Suraiya Pervin, ZerinaBegum [6], suggested the concepts of a single document text summarization that differentiates the various approaches in this reviewed form. The text review emphasizes the process of summarizing text.

Rahul C Kore, Prach Ray, Priyanka Lade, Amit Nerurkar [7], introduced a system that focuses only on extractive summarization only. The system presented focuses on producing a summary from a given document using a variety of language processing methods such as word embedding, similarity measures and ranking algorithm.

Varun Pandya[8], presented a method that works well against previous approaches. The summary generated closely simulates the original text as generated by the model and can possibly be used in the court of law after further improvements. The summary generated by the presented system after further changes and structuring can possibly be used for real-time cases in the future, after carrying out real-world tests with the model. Since it is an unsupervised approach, including clustering using k-means and extracting top-ranked sentences from each cluster, and is computationally favorable, it provides a promising start towards developing a fully functional Legal Case Summarizer.

Ranjitha and Kallimani in [9] explained the findings and results of a summary of several texts. A multi-document summary contains an automated summary from multiple documents that shows us about the same topic or event. There are three main ways to summarize many documents based on compression, based on extraction, based on Extraction. The output-based approach is less effective as it creates a summary by selecting sentences as they appear in the input but often leads to unemployment. Some of the errors detected in the subtraction-based approach are solved by a supportive design method by making a needless summary by deleting phrases or words but failing to integrate sentences from different sources. The most adequate method is an ambiguous method that creates sentences that are not in the original text or documents. Two important approaches to summarizing multiple texts are Phrase Salience Calculation. The first collection and compilation of sentences will be polished by checking nouns with clear letters and action sentences. They also say that Clustered Semantic Graph can be made. Obtained by content monitoring and editing. The value of this is high in terms of efficiency. The maximum energy benefit is provided by the reduction of layoffs in combination. This paper also declares the findings of a well-developed standard grammar. Sentence selection and strategy integration create new sentences from texts by exploring a combination of syntactic units such as Noun Phrase / Verb Phrase.

III. PROPOSED ALGORITHM

1. At first, we wanted to split the sentences and store it in a list. We did that by using Spacy for splitting the lines but sinceit was noisy, we created a custom sentence splitter and split the sentences of a doc accordingly.
2. Next, we wanted to use the tf-idf vectorizer over the list of documents, so that we can use that to obtain the tf-idf of a sentence. The tf-idf vector of a sentence will be of a particular dimension where the number of rows = 1, as there is only one sentence, and the number of columns will be = the number of unique words in the whole corpus. Thus, each sentence of a document can be obtained as a vector.
3. We then pre-computed the similarities between each pair of sentences in the document and stored it in a 2-D array by computing the dot product of two vectors. For each sentence the maximum similarity was thus found by comparing the values for that sentence with other sentences by using the precomputed similarity values.
4. After that we applied the Maximum Marginal Relevance(MMR) algorithm to find the summary. This algorithm was used to increase the relevancy of the sentences that are being added to the summary as well as reduce the redundancy (i.e., preventing the same sentence from getting added to the summary).
5. We iterated through all the sentences in the document and found the maximum marginal relevance score of all the sentences and updated the summary by adding the sentence with the maximum marginal relevance score.

6. We kept on adding sentences to the summary unless it crossed a particular threshold which was passed as a parameter
7. We read the files and generated the 2 summaries based on the reference summaries A1 and A2 for 3 values of λ: 0.3, 0.5, 0.7
8. We calculate the rogue scores for all the generated summaries with respect to the reference summaries and store the rogue-1,rogue-2,rogue-l scores in 'rogue.txt'. Rouge helps us to assess the adequacy of the summary by simply counting how many n-grams in your generated summary matches n-grams in our reference summary.
9. We calculate and display the average of rogue scores of 50 documents at the end.

IV. SYSTEM ARCHITECTURE

1. Divide the case document into sentences, each sentence being a separate document.
2. Initialize summary to be empty.
3. While the word count is below the desired word count, we add to summary the sentence (From a given 'List ofsentences' = Case document in our case) with 'maximal MMR (Maximal Marginal Relevance) score'.
4. MMR score = F(Sentence_to_be_added,Case_document)*λ - F(Sentence_to_be_added,Summary)*(1-λ).
5. Here F(Sentence_to_be_added,some_document) = { i>= 0 and i < len(Some_document) }max(Similarity(Sentence_to_be_added,Some_document[i]))

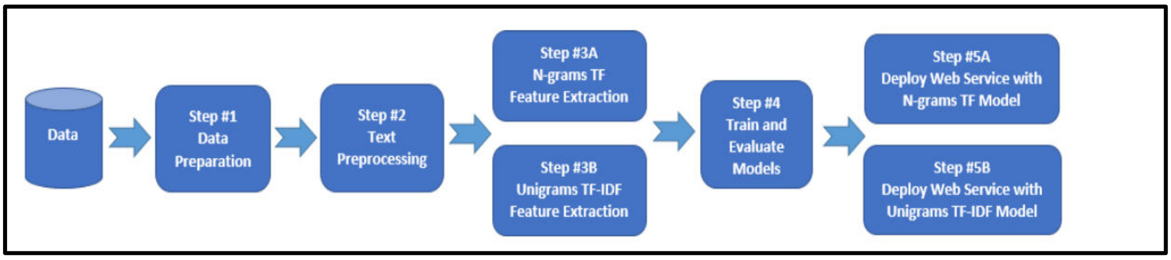
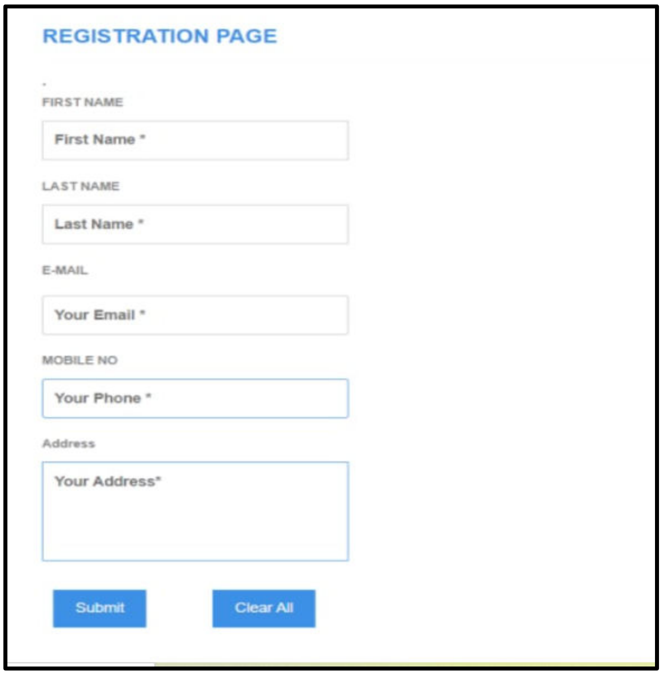


Fig.1. System Architecture

V. RESULT

Registration Module: User must register to access the portal. Password will be sent to email, to verify the email id is legitimate.



REGISTRATION PAGE

FIRST NAME

First Name *

LAST NAME

Last Name *

E-MAIL

Your Email *

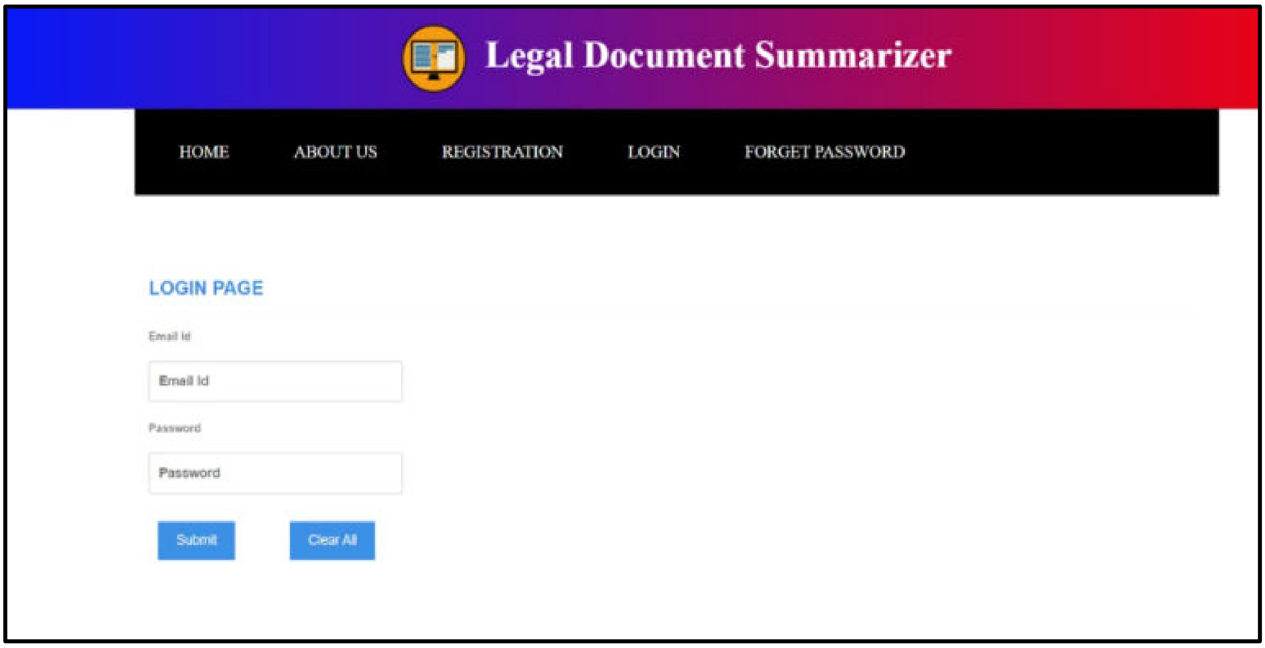
MOBILE NO

Your Phone *

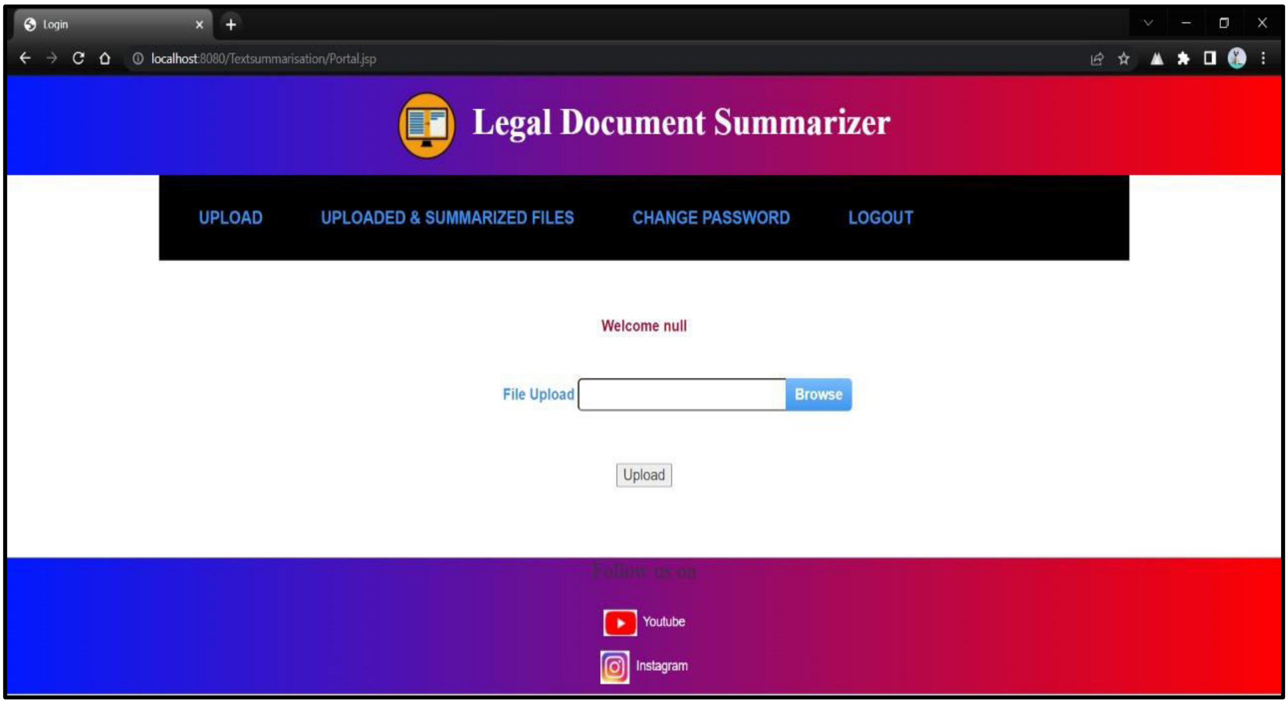
Address

Your Address*

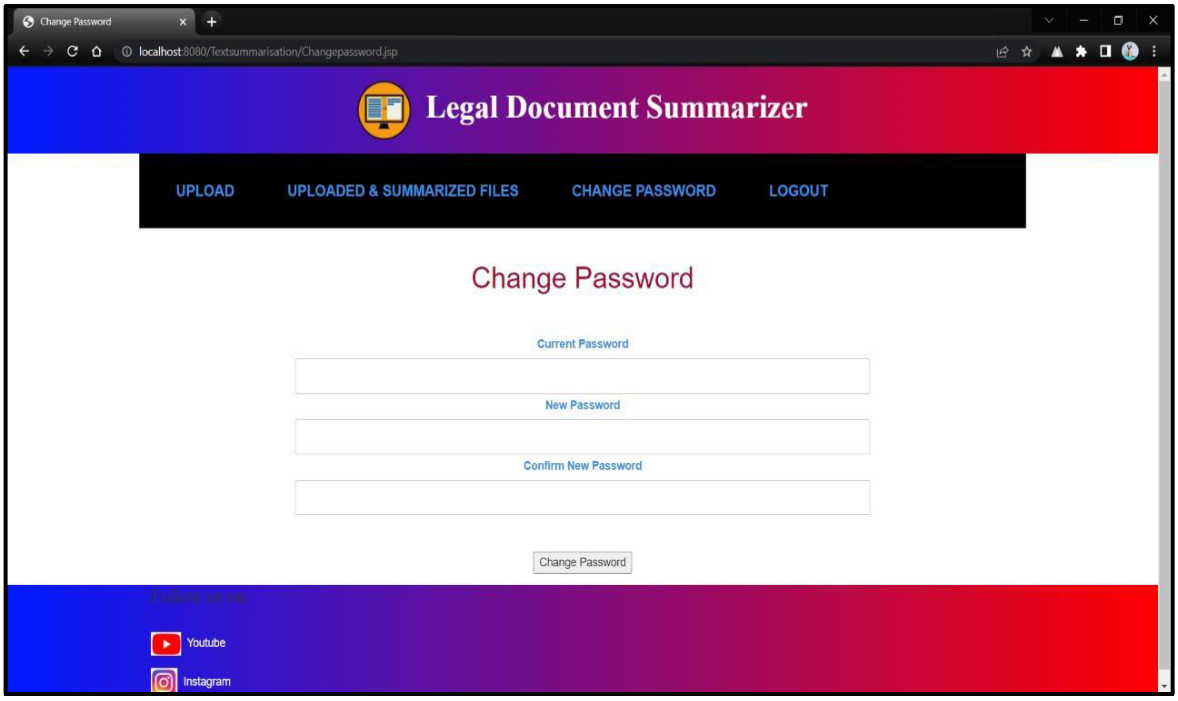
Login Module: Once user is register, he/she can login with password send on email and email id as username.



File Upload Page-


Login
localhost:8080/Textsummarisation/Portal.jsp
Legal Document Summarizer
UPLOAD
UPLOADED & SUMMARIZED FILES
CHANGE PASSWORD
LOGOUT
Welcome null
File Upload
Browse
Upload
Follow us on
Youtube
Instagram

Change Password Page-



VI. CONCLUSION AND FUTURE WORK

In this paper, we've given our approach for handling automatic summarization techniques. This work refers to the matter of process an enormous volume of electronic documents within the legal field that becomes additional and harder to access.

In the future, there are many points that deserve our attention.

- The multiple-legal document summarization.
- Different file formats for summarization.
- Summarization of legal documents containing Indian languages.

REFERENCES

[1] Jiawen Jiang, Haiyang Zhang, Chenxu Dai, Qingjuan Zhao, Hao Feng, Zhanlin Ji, and Ivan Ganchev., “Enhancements of Attention-Based Bidirectional LSTM for Hybrid Automatic Text Summarization.”, IEEE 2021.

[2] Ravali Boorugu, Dr. G. Ramesh, “A Survey on NLP based Text Summarization for Summarizing Product Reviews.”, IEEE 2020.

[3] Virender Dehru, Pradeep Kumar Tiwari, Gaurav Aggarwal, Bhavya Joshi, and Pawan Kartik, “Text Summarization techniques and applications”, IOP publishing 2020.

[4] Shengli Song, Haitao Huang, Tongxiao Ruan, "Abstractive text summarization using LSTM-CNN based deep learning”, Springer Science-2 February 2018.

[5] Subash Voleti, Chaitan Raju, Teja Rani, Mugada Swetha,”, Text Summarization using Natural Language processing and google text to speech API", IRJET-05 May 2020.

[6] Md. Majharul Haque, Suraiya Pervin, Zerina Begum, "Literature Review of Automatic Single Document Text Summarization Using NLP”, International Journal of Innovation and Applied Studies-3 July 2013.

[7] Rahul C Kore, Prachi Ray, Priyanka Lade, Amit Nerurkar, "Legal Document Summarization Using Nlp and Ml Techniques",IJESC 05 May 2020.

[8] Varun Pandya, "Automatic Text Summarization of Legal Cases: a Hybrid Approach".

[9] N.S. Ranjitha, Jagadish S Kallimani. Abstractive multi-document summarization, 2017 International Conference on Advances in Computing, Communications, and informatics (ICACCI), 2017.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.165

doi cross ref

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

Scan to save the contact details

www.ijircce.com