Smart Information Retrieval

Bhavesh Satpute[1], Bhushan Sawant[2], Sushant Yelurkar[3], Omkar Dighe[4], Rashmi Jolhe[5]
[1,2,3,4]*Student, Dept. of Information Technology, Datta Meghe College of Engineering, Maharashtra*
[5]*Asst. Professor, Dept. of Information Technology, Datta Meghe College of Engineering, Maharashtra*

Abstract—Over the years, we are studying web pages, technical papers, research papers, etc. The traditional method of retrieving a piece of information from a large collection of data or information was manual searching and it was taking huge time. But in today's fast world, we used this information retrieval concept to find data in seconds. So, it becomes a fast and less time-consuming method for retrieving information. Even though its retrieving speed has increased but the quality of information is less. So, to overcome this we introduced Smart Information Retrieval where the user will get the relevant information on the basis of the user query, and also the user will get the summary on the basis of the Abstract, we classify this abstract in labels like Background, Methodology, Objective, Result, Conclusion. The accuracy of the predicted answer based on the user query is more the 70% and the relevancy of the generating summary or classifying the Abstract is 87%. This ability to predict answers and generate a summary or classify the Abstract from the given input makes this information retrieval system different from others.

Keywords: Fast information retrieval, Relevant Information, Real-time data, Chrome Extension, Current webpage, PDF, Search on a user query.

I. INTRODUCTION

The proposed information retrieval system provides an answer to the question asked by the user. It is planned to develop a Chrome extension where the user can ask questions. In the Chrome extension, three options are provided, first is to get the answer from the current page, second is to get the answer from PDF, and third is to generate a user-specific lines summary based on input PDF or Image. each has one textbox where the user can place his question and number of lines. It has been observed that web pages show irrelevant information in spite of the required one. So, the current page option is provided, where the entire data of the current web page is extracted and sent to the server with the question asked in the textbox. On the server, the proposed model retrieves the correct answer from the extracted data and the output is sent to the user. Generally, the user has to read the PDF which can be 20 to 100 pages. So, an option called get an answer from PDF is provided where a PDF doc is uploaded, and type the question in the textbox. The user question and text from the pdf are sent to the server and the model generates the correct answer for the question asked and displays it to the user. If the PDF contains an image, then the image is also sent to the server where the text from the image is retrieved by using the OCR method (Optical Character Recognition). The text retrieved from the image is sent to the server and the model forwards the answer to the user. Whereas many times it is not sufficient to get the specific query answer, the user needs to get the summary of the PDF (technical paper) or Image which can give short important information about the whole document. So, this proposed model takes the whole information of the PDF (technical paper) as input except the Abstract and Conclusion and provides a user-specified lines summary based on the important points.

We also learned that providing a particular query response is not always adequate for the user; providing a summary in a categorized manner is also required, providing the user with a comprehensive understanding of the PDF (technical document). So, in our Chrome Extension, we've added a third option to generate a summary. There is a textbox where the user may enter the abstract of the paper. Based on that abstract, our BERT - Base - Uncase model will categorize it into five labeled sections (Background, Objective, Method, Result, and Conclusion). It is not required to display all of the labels; it will just display the labels that are compatible with the Abstract. This option may be handy for anyone who needs to read several technical papers; it relieves them that they simply have to add an abstract to the extension and the entire

information will be displayed to the user with labels. As a result, time is saved and the information retrieval procedure becomes faster.

II. CONCEPT USED

1. BERT Model

A deep learning model based on Transformers is called Bidirectional Encoder Representations from Transformers (BERT). In Transformers, each output element is connected to each input element, and the weights between them are dynamically determined depending on their relationship. The simultaneous reading in both directions capability of BERT makes it special. Bi-directionality is a quality made possible by the appearance of Transformers.

BERT in our Smart Information Retrieval helps to predict the answer based on the user-provided question. The text is gathered from the Current web page, PDF, and the Image is sent to the BERT model which is on the flask server where the BERT is pre-trained to predict the answer to a question from the large text. So, as the gathered text and the entered question are given to the model it finds the most accurate answer from that large text and is shown to the user as an output.

2. OCR

Optical character recognition refers to the process of turning a text image into a machine-readable text format (OCR). For example, if you scan a form or a receipt, your computer saves the scan as an image file. The words in the picture file cannot be modified, searched for, or counted using a text editor.

What was the need to use the OCR?

This project is proposed in such a way that I may achieve the answer predicting from the image also. Many times, it happens that the PDF may have multiple images and it may happen that our question's answer is written in one of that images but the normal text model is not able to predict that. So, for overcoming that we used this OCR (Optical character recognition) which helps this model to achieve the prediction of the answer from the Image also.

How the OCR is used?

Whenever the PDF consists of single or multiple images, we used this OCR method. When the PDF is entered by the user the whole text in that PDF and the question which is entered by the user is sent to the flask server where are BERT model is present But, if there is an Image in the PDF then first it is converted into the Base64 Format these help the transfer of the image to the server becomes fast. After receiving that Base64 format image at the server first it is converted into the normal image.

As soon as the image is received at the server the OCR method is applied to that image. The image is scanned after that text is recognized from the image then the pattern matching is done means that is the size, format, and style of the text is and compared with the previously recorded glyph then the feature extraction is done on that image where the glyph is divided into lines, closed loop, line direction, etc then the last postprocessing is done and the finally the text from the image is retrieved and sent to the model.

If the Image is directly uploaded at the place of the PDF, then also the prediction of the answer is done.

3. Base64 Format

Base64 is the encoding scheme that transforms binary data into text format so that encoded textual data may be quickly and simply sent over the network without any data loss or corruption.

We use Base64 to transform photos into text in Smart Information Retrieval so that sending many images to the server for detection would be simple. The procedure will go more quickly if the image is in Base64 format since it uses less internet and less time. Once the text-based picture has been successfully received by the server. We are once more converting the text format to an image for OCR using the base64 approach.

4. BERT-Base-Uncase Model

BERT-Based-Uncase Model is a model of Google. It is not a pre-trained model but the company name Hugging Face has pre-trained these models as PubMed for the medical summarization purpose. But this Smart Information Retrieval must be in a limited field so, we trained this model. This model takes the Abstract as input and classifies it within 5 labels (Background, Objective, Method, Result, and Conclusion).

Steps, how we trained these BERT-Based-Uncase

1) First, we load the CSV (PubMed_20K_RCT) into the Trained and Test datasets.

2) We gave the label that the abstract must be classified into Background, Objective, Method, Result, and Conclusion.
3) The Dataset cleaning is done.
4) The dataset of PubMed_20K_RCT is not pre-trained so it is divided into train and validation sets as train_texts, val_texts, train_label, and val_label.
5) The tokenizer loaded and encoded the text
   a.Tokenizer means converting the text into numerical form. The Abstract in the dataset is given to the tokenizer which converts that Abstract text into numerical form. It means encoding the Abstract text.
6) After tokenization that encoded text is given to text_encoding, val_encoding, and test_encoding
7) The target labels are set
8) a.It means setting a target that when the abstract is given to the model the target which is the classification of the Abstract into the labels may be achieved.
9) We used the library called Torch which is a Python library
10) a.The Torch another name for this is PyTorch which converts our dataset to the tensor dataset. This tensor dataset consists of Input IDs and Attention Masks. The above encoded Abstract is sent to this to get the output where the Input IDs are the inputs and the attention Mask is the output. All inputs and outputs are saved in this tensor dataset.
11) To train the data torch is given CUDA and CPU where if the device has Nvidia graphics then it runs on the CUDA or it may run on the CPU. For training the Adam Optimizer is given as lr=2e-5.
12) The dataset is trained up to 2 epochs, where the patience was 3. So, the Epochs run till the 5 but the most accuracy was received till the 2nd Epochs. So, we have taken the 1st two Epochs as the most accurate training dataset. The one Epoch included 4501 data means the 4501 Abstract, so we used 2 Epochs which means the model is trained on the 9002 Abstracts.
13) We have also tested our model on the test data which was of the 1884 Abstract and the accuracy we received of the model is 0.8681 means 86.81%.

III. LITERATURE SURVEY

1. Question Answering using Deep Learning with NLP (Natural language processing)

Due to difficulties in comprehending the question and determining the right response, answering questions is still a challenging task.

With training data, our model learns an ideal representation for the input question and response phrases as well as a matching function to relate each such pair.

2. A Novel Web Scraping Approach Using the Additional Information

Online scraping is a technique for obtaining useful and fascinating text data from websites.

There is a lot of extraneous stuff that must be removed in order to extract the information from the web pages. Hence, creating a retrieval system that can eliminate the unneeded and give users the material they need.

3. Web Extension for Text Authentication on Google Chrome

A web browser may be equipped with software known as web extensions.

Either automatically or by clicking the extension icon, the extension functions. We created a web extension for Google Chrome that pulls the text from the currently-displaying web page.

4. An automated conversation system

The exchange of text-based information is essential for human communication.

The computer first chooses the answer from a known selection of statements for that response, then chooses the response from the statement with the closest match to the input.

5. Smart Answering Chatbot

A chatbot is a computer software that uses artificial intelligence and messaging systems to have conversations with people.

The chatbot retains user input and responds each time it receives input from the user. This allows the chatbot, which initially had minimal knowledge, to develop using the responses acquired. The chatbot's accuracy improves as the number of replies rises.

6. Information Retrieval in Computing Model

Information retrieval is a method by which we gather pertinent data from a sizable number of information sources and deliver it in response to the

user's request. Depending on the user's preferences, information retrieval can be used to find papers, music, photographs, etc.

IV. PROPOSED METHODOLOGY

The proposed Smart Information Retrieval System will provide an answer to the question asked by the user and also provides a summary of the user-specified lines. We are developing a Chrome extension that will be flexible for the user to use, where the user can ask their questions and also enter a specific number of lines for the generation of a summary. In the Chrome extension, they will have three options first is the get answer from the current page, the second one will be the get answer from PDF and the third will be to generate a summary.

Get answer from the current page:
We learned that many times on web pages there is useless information that is shown and we have to search for useful information. So, we develop the current page option, where the whole data of the current web page will be extracted and sent to the server with the question asked in the textbox. On the server, our model will retrieve the correct answer from the extracted data and the output will be shown to the user.
Get answer from PDF or Image:
We also learned that to retrieve a little bit of information user has to read the whole PDF which can be 20 or 100 pages also. So, we have an option called to get an answer from PDF where we can upload our PDF doc and type the question in the textbox. The question and the whole text from the pdf are sent to the server and the model will generate the correct answer for the question asked and show it to the user. If the PDF has an image, then the image may also be sent to the server where the text from the image can also be retrieved by using the OCR method (Optical Character Recognition). The text which is retrieved from the image is also sent to the server and our model will give answers from the image as well.

Generate a summary
We also learned that getting a specific question answer is not always sufficient for the user, to get the summary in a classified way is also necessary which gives the user a whole idea of the PDF (technical paper). So, we have given a third option in our Chrome Extension which is to generate a summary. There is a textbox where the user has to enter the Abstract of the paper form that Abstract our BERT – Base – Uncase model will classify that abstract into 5 labeled parts (Background, Objective, Method, Result, and Conclusion). It is not compulsory to show all the labels, it will only show the label which is consistent with the Abstract. This may be useful for everyone who has to read the multiple tech paper, this option gave them relief that they only have to add an abstract to the extension and the whole info will be shown with the labels to the user. So, the time is decreased and the information retrieval process becomes fast.

V. OBJECTIVE

1.To make the time consumption less for retrieving information.
2.Make the reading of web pages and research papers Easier.
3.To Summarize the huge amount of information in a readable format.
4.To provide the relevant answer to the question asked by the user.
5.To provide the summarized data in the labeled form.

VI. DIAGRAM

Data Flow
Get Answers from the Current Page

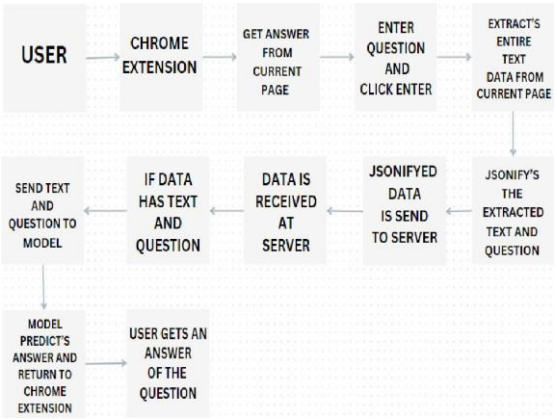

Fig.no 1

Get Answers from the PDFs

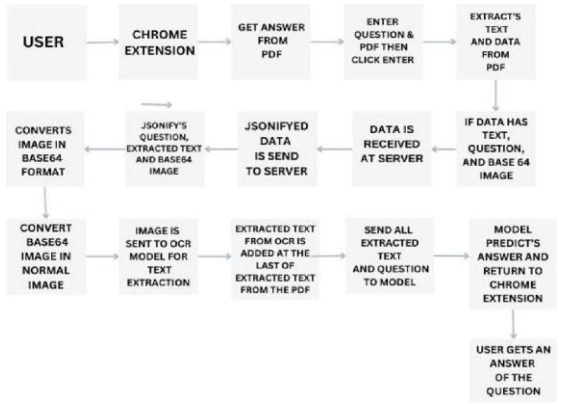

Fig.no 2

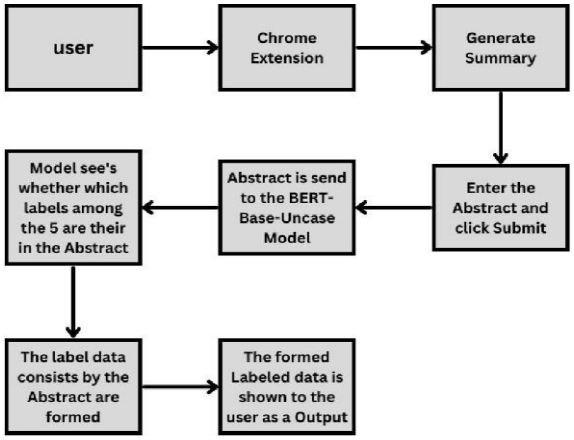

Fig.no 3

VII. SCOPE OF PROJECT

It will help to find the relevant answer in less time. No one has to read the whole pdf or the webpage to find the relevant answer. The model is designed in such a way that it will generate the answer on the basis of the information retrieved and the user's question.

So, in the future, it will take the minimum time to search for an answer based on user queries from any of the web pages that are live on the screen and from uploaded PDFs. Also, the Classification of the Abstract is done under 5 labels (Background, Objective, Method, Result, and Conclusion) which help to get the whole paper info into the summarized form.

VIII. RESULTS AND ANALYSIS

The result of Smart Information Retrieval, as above mention this project may predict the answer on the basis of the user question and the data which is provided in real-time. The accuracy of the model predicting an answer is more than 70%. As it is in real-time the data may not have good quality so the accuracy of the prediction from the current page and from the PDF or Image may differ.

The third part of the Project is the Summary generation or the classification of the Abstract with the BERT-Base-Uncase model which has an accuracy of 86.81%. The accuracy is calculated with the help of the train and the test data. We have used the 9002 Abstract to train the model and to test we used the 1884 Abstract from the dataset, the Abstracts are chosen on a random basis. The model training is done till 5 Epoch where one consists of 4501 Abstract randomly. The patience was given as 3 so till 5 it goes and takes the Epochs which are most accurate, till the second Epoch we received 0.8681 as the highest accuracy.

So, the result of the Analysis is as follows: Test accuracy: 0.8681
Test F1 score: 0.8664
Test precision: 0.8659
Test recall: 0.8681

And the confusion matrix of this results and analysis is shown below:

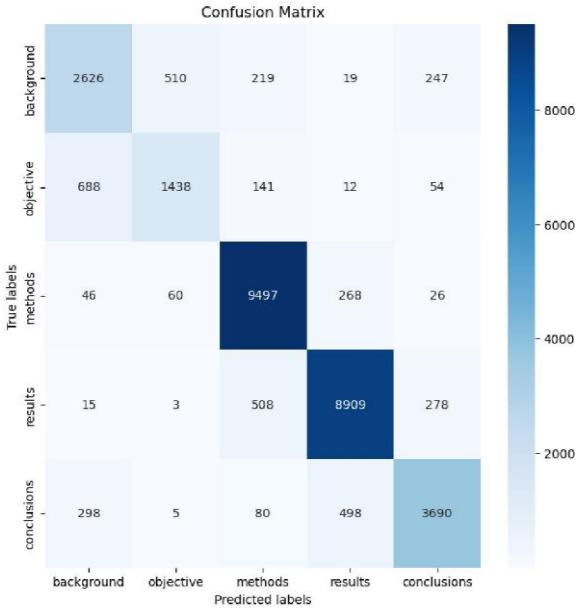

Fig.no 4

IX. COMPARISON TABLE

SR.NO	AUTHOR NAME	TITLE OF PAPER	COMPARISON BETWEEN PUBLISHED AND PROPOSED MODELS
1	Anupam Mondal, Monalisa Dey, Dipankar Das, Sachit Nagpal, Kevin Garda.	An automated conversation system	This paper says their automated conversation system answers only 1000 pairs of questions and their accuracy of prediction is 86.60% but, our proposed model may answer more than 86,000 questions based on real-time text data from PDF, current web pages, and images also and our accuracy of predicting answer for PDF and Image is more than 70%.
2	S Gholami, M Noori.	You Don't Need Labeled Data for Open-Book Question Answering	This published model states that it has the exact match percentage of 39% while the dataset used is the same as our proposed model but it is pre-trained whereas our proposed model is also pre-trained but takes input on real-time data and the answer predicting rate is more than 70% for PDF and image data. This project is one such project which predicts answers from both Image as well as PDF

X. CONCLUSION

In this research, we applied a technique that can be considered helpful and may outperform the creation of results more precisely and efficiently as compared to the original information retrieval. The key state of our Smart Information Retrieval is that it uses the OCR method to retrieve the text from the images. By this, we are able to answer from the image text also. Compared to other information retrievals this Smart Information Retrieval is dynamic because by using the BERT, OCR, Base64, and BERT-Base-Uncase model gives the ability to answer the user question from both text and image text with faster and more relevant results as well as it provides us with the summarized information in the form of labeled data.

This study is helpful for teachers, students, etc who all read lengthy papers or websites to retrieve relevant information. This proposed system is wholly based on NLP (natural language processing) which helps machines understand the user language. With all these techniques we are sure that our proposed model would be more effective and precise and may solve the various retrieval problems.

REFERENCE

[1] Affordance. (n.d.). In the Merriam-Webster.com dictionary. Retrieved 17.5.2020 from https://www.merriamwebster.com/dictionary/affordance

[2] Ariyapperuma, S., & Minhas, A. (2005). Internet security games as a pedagogic tool for teaching network security. Paper presented at the Proceedings - Frontiers in Education Conference, FIE,, 2005 S2D-1-S2D-5.

[3] Burns, T.J., Rios, S.C., Jordan, T.K., Gu, Q., & Underwood, T. (2017). Analysis and Exercises for Engaging Beginners in Online CTF Competitions for Security Education. ASE @ USENIX Security Symposium.

[4] Chapman, P., Burket, J., & Brumley, D. (2014). PicoCTF: A Game-Based Computer Security Competition for High School Students. 3GSE.

[5] Chothia, T. & Novakovic, C. (2015). An Offline Capture The Flag-Style Virtual Machine and an Assessment of Its Value for Cybersecurity Education. 3GSE.

[6] Chung, K., & Cohen, J. (2014). Learning Obstacles in the Capture the Flag Model. 3GSE.

[7] Conti, G., Babbitt, T., & Nelson, J. (2011). Hacking competitions and their untapped potential for security education. IEEE Security and Privacy, 9(3), 56-59. doi:10.1109/MSP.2011.51.

[8] Dabrowski, A., Kammerstetter, M., Thamm, E., Weippl, E.R., & Kastner, W. (2015). Leveraging Competitive Gamification for Sustainable Fun and Profit in Security Education. Eagle, C. (2013).

[9] Computer security competitions: Expanding educational outcomes. IEEE Security and Privacy, 11(4), 69-71. doi:10.1109/MSP.2013.83

[10] Gavas, E., Memon, N., & Britton, D. (2012). Winning cybersecurity one challenge at a time. IEEE Security and Privacy, 10(4), 75-79. doi:10.1109/MSP.2012.112

[11] Gholami, Sia, and Mehdi Noori. "You Don't Need Labeled Data for Open-Book Question Answering." *Applied Sciences* 12, no. 1 (2022): 111.