# Tool to Extract and Summarize Methodologies of Research Articles for Visually Impaired Researchers

***Abstract*—With the development of technology, not only the university students but also visually impaired students have also started engaging in higher studies and research. They have started using electronic documents comprehensively for their studies. However, some components of electronic documents are also in a difficulty to be referred as per their complexity. This paper has introduced an application to address such a component and it is to summarize the research methodology section of Information Technology related research papers including the flow charts and graphs. The requirements of visually impaired students in research were identified and the paper has introduced a system to help these students by summarizing methodologies of a given paper within a very short period of time accurately in the document form and in voice. The methods such as machine learning, Information extraction and natural language processing were mainly applied to bring the solution out. Currently available applications were thoroughly observed and introduced this concept with several features that are lacking in the existing solutions. The results were observed via visually impaired researches with their consent on the support provided by the output.**

***Keywords—Visually impaired, Information extraction, Natural Language Processing***

## I. INTRODUCTION

Over the past years, more and more visually handicapped students have attempted higher studies. There are about 45 million blind people and 135 million visually impaired people worldwide. These students face different kinds of problems when dealing with educational materials. One of the problems is, the need to study text documents in electronic devices, in depth and in a reasonable time. Disability of text reading visually has a huge impact of day today life and the quality of life for visually impaired people.

With the development in the technology, all the communities including visually impaired students have got interest in referring different types of electronic documents as they have started educating themselves and have started working with the people in different areas. These visually impaired students have been doing their studies with a lot of difficulties. With the development of technology, these students too wanted taking advantage of the new technologies introduced. Earlier they used the method 'braille' to read documents. And they could only read paper work with that methods. These students face different kinds of problems when dealing with educational materials.

There are some software to navigate through documents, but these software do not support blind people to extract only the exact section of the e-book. It only navigates to the beginning of a specific section. The research which have been carried out so far have invented summarization tools. But the accuracy of these tools are low and they do not have the feature of diagram extraction and graph extraction.

Therefore, in this research we have a solution to extract and summarize methodologies of research papers related to information technology and get the main idea in point form. This application will extract texts, flow chart diagrams and graphs in the methodologies and summarize the texts, explains the flow chart diagrams and graphs. These extracted and summarized data will be transformed to visually impaired students through voice.

The main aim of this proposed system is to help visually impaired students to get the summary of a methodology which includes texts, graphs and diagrams of research papers related to Information Technology accurately and in a quality manner. The final output will be presented as a document and in voice.

The tool we implement will extract and summarize text, flow chart diagrams and graphs in the methodology of the research paper. From this tool,

- The main idea of the text will be meaningfully summarized
- The flow chart diagrams will be extracted and describe what is it about
- The graphs will be extracted and describe what the graph is drawn, and the impacts and results shown in the graph.

The summarized details of text, diagrams and graphs will be entered to a document in text and converted to voice. This summarizing tool will be finally delivered to the visually impaired person as a web-based application. This is developed with standards specified for visually impaired people and will also give suggestions on similar research papers.

Since, so many visually impaired students have started doing their higher studies with the development of technology, this product will help them in order to do their higher studies efficiently and effectively. The expected outcome of the system is to get a valuable service to the visually impaired students as well as to all the students who engage in higher studies.

There are various summarizing tools developed using different technologies for common use of people, but not specifically for visually impaired people. Therefore, reading research papers by visually impaired students become a time-consuming activity to get a clear summary of a research paper. The accuracy of the summary obtained by the visually impaired student can also be low.

This tool will summarize the methodology of a research paper within a short period of time accurately and efficiently. The quality of the summary will be a main factor we consider in the tool.

## II. LITERATURE REVIEW

Several investigations on summarization of documents and image processing were done separately throughout years. But a limited amount of research were done to in order to help visually impaired students.

### *A. Summarization tools*

For text summarization a huge number of researches were done successfully in the past. Among them, the research done by SRA International and Department of Defense in the United States of America [1], they describe a trainable and scalable summarization system which utilizes features derived from information retrieval, information extraction, and NLP techniques and on-line resources.

Eduard Hovy and Chin-Yew Lin of Information Sciences Institute of the University of Southern California [2], implemented a summarization tool named "SUMMARIST" to create summaries of arbitrary text in English and selected other languages. It was included language-specific techniques of parsing and semantic analysis, and was combined robust Natural Language Processing with symbolic world knowledge, derived from WordNet

The survey done by Vishal Gupta of Computer Science & Engineering department, in Panjab University Chandigarh, India [3], presents a number of techniques used for text summarization and extraction. Also, he has pointed out the main functionalities of some successive summarizers, such as, Newsblaster, Automatic text summarization system [4], The trainable document summarizer [5] and the ANES [6] text summarization system.

### *B. Diagram Extraction*

So many researches were carried out for image extraction and description, but very few were done aiming visually impaired students. And some researches were conducted to identify and describe flow charts.

On-line handwritten flowchart recognition, beautification, and editing system [7] was conducted in order to identify handwritten flow charts. Flow chart symbol identification is done using loops. Shapes are identified and inside details are extracted. As this research is done for handwritten flow charts, they have modified the flow charts using beautification tool.

The research, A sign reading system for the visually impaired [8], done by American University of Sharjah, that aims at helping the visually impaired by locating indoor signs and reading their content out loud, thus guiding them toward their destination. The process is to capture the signs in public places using a camera and to extract the text and to process the image and to read it aloud in order to give directions.

Text/Image Region Separation for Document Layout Detection of Old Document Images using Non-linear Diffusion and Level Set, research [9] discusses how to identify images out of a document and to extract the features of the image. Profiling or morphological operations are used to separate diagrams from text. They have introduced a nonlinear diffusion method known as edge enhancement diffusion (EED). The result was that any diagram was separated from a document successfully.

The research, Textline detection in degraded historical images [10], which was conducted by EURASIP Journal on Image and Video Processing has been carried out to extract text from degraded images. Textline detection and binarization were the main techniques used.

### *C. Graph information extraction tool based on the image*

According to the past literature reviews related to the graph information extraction, got to know that some components are not that much implemented.

"Graph-based representations and techniques for image processing and image analysis" [11] is one of the research that is based on graph extraction, they have mentioned more theoretical techniques than applying practically. They explained the graph partitioning greedy algorithm for colour image segmentation. Image segmentation is the process of partitioning an image into a set of non-intersecting regions such that each region is homogeneous and the transition from one region to another is sharp. And also it has described a novel fusion of color-based segmentation and depth from stereo that yields a graph representing every object in the scene. Like that all the theoretical parts have mentioned in this research. But the way of extracting the graph information and storing the information has not been mentioned practically.

"Graphs for image processing, analysis and pattern recognition" [12] also another thesis that describe the theory behind the graph information extraction. It has described single graph methods like segmentation or labeling graph-cuts, graphs pattern recognition. Under graph matching it has described the graph or sub graph isomorphism, error tolerant graph matching and more. But it has not been mentioned the practical situation.

"Extraction of graph information based on image contents and the use of ontology" [13] is one of the research based on the graph information extraction. It describes the way of analyzing X, Y axis and the more attention to extracting information from the graph.

*D. AI Chatbot*

AI Chatbot will try to understand the query and provide a definitive answer. There will be four main units to the system working together to understand the question and return an appropriate answer: Generic question construction - capable of taking a natural language question and making it more generic. Generic answer construction [14] - capable of taking a generic question template and providing a generic answer template. Generic answer population - capable of taking a generic answer template and populating it with information from the database to form an answer [15]. Information extraction - capable of finding information through structured or unstructured websites, and storing that information in a database

## III. METHODOLOGY

Proposed solution addresses a tool to extract and summarize the research methodology for the visually impaired students. This project has very important research areas like, Natural Language Processing, Machine learning, Image Processing, Text to Speech conversion and web application development. This research is conducted on the above-mentioned research areas to achieve the project objectives.

*1.Text Summarization*

In a computer, a text is represented as a string of characters. The first step in processing natural language is to break up this character string into units that will provide the basis for all our Natural Language Processing by Tokenizing. Grammatical words such as prepositions and determinants do not make much difference for the meaning in the text. These are also called stop words and are usually removed in tasks such as information retrieval.

Secondly, we may want to consider two or more words as being one unit since they have the same root or stem. For example, 'descendant' and 'descending' can be assimilated to 'descend'. Combining words having the same root into one stem is called 'Stemming'. The remaining words after removing the stop words are stemmed using the well-known and widely used Porter Stemmer. At the end of stemming the output will be gained as a sentence matrix.

Finally, the tool generates a summary based on the rules derived from supervised machine learning algorithms which are, Sci-Kit, Pagerank and Textrank algorithms. The text summarization approach also has a polynomial time complexity which returns the output at a reasonable time.

*2. Flow chart extraction and Describing tool*

As the input of the system is a pdf document, in order to extract images out of it, we have used pdf2image library files to convert pdf to image format. Image processing techniques and library files have been used to identify and extract flow charts out the document. The content of the image is extracted in the order of the flow chart using pytesseract library files. Recurrent Neural Network (RNN) algorithms are used for text extraction and identification. The details of the flow chart will be presented to the visually impaired student as document with the correct order of the flow chart.

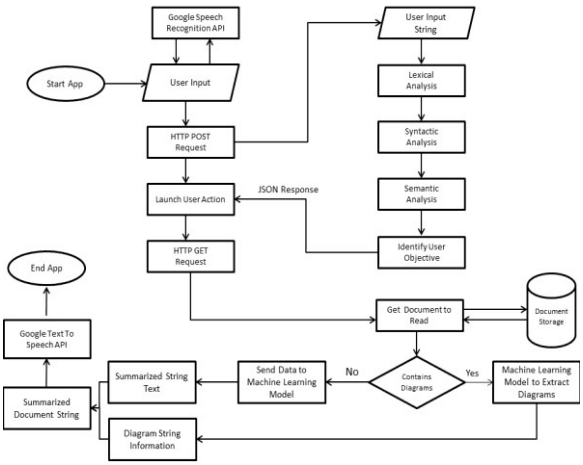


*Fig. 1. Design of the system*

## IV. RESULTS AND DISCUSSION

*A.Text summarization tool*

The design of a tool of the text summarizer is of great importance in the current world which is so filled with data. It would reduce the pain visually impaired people suffer when using e-documents in their day to day life. The text summarization part was developed based on tokenization, stop-word removal, sentence segmentation, stemming and vectorization. It generates a summary based on the rules derived from a supervised machine learning algorithm, The Sci-Kit algorithm. The text summarization approach also has a polynomial time complexity which returns the output at a reasonable time.

In order to receive a precise summary, the Textrank algorithm is mentioned when developing the text summarizing tool. Here, we have used some special features and gained results, such as, removal of punctuation and special characters, create a word vector each with size 100 of each sentence, check similarity between sentences and making a word graph by applying Pagerank algorithm.

Table I shows the accuracy, precision and recall of the text summarization at different approaches in percentage.

| ID | TEXT SUMMARIZING | | |
|---|---|---|---|
| | *Accuracy* | *Precision* | *Recall* |
| 1 | 72.0 | 44.4 | 48.0 |
| 2 | 57.6 | 40.0 | 28.6 |
| 3 | 72.9 | 45.5 | 31.3 |
| 4 | 76.6 | 47.6 | 50.0 |
| 5 | 84.2 | 60.0 | 60.0 |
| 6 | 87.8 | 28.9 | 64.7 |
| 7 | 74.5 | 33.3 | 40.0 |
| 8 | 100 | 100 | 100 |
| 9 | 77.0 | 54.5 | 54.5 |
| 10 | 86.8 | 52.5 | 58.9 |

*TABLE I. Accuracy, Precision and Recall of the Text Summary*

Figure 2. shows the snapshot of the returned piece of text as the summary of the methodology of the research paper. This gives the main idea of what the methodology is discussed and the procedures it contains.

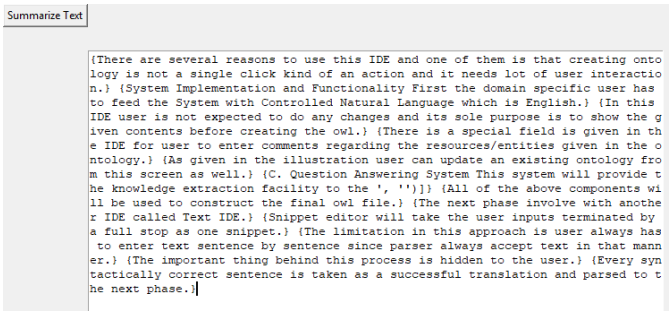


*Fig. 2. Snapshot of the Text Summarization output of the Methodology*

*B. Flow Chart Extraction and Describing Tool*

This section of the research produces an application to identify flow charts out of a research paper related to Information Technology and to extract the content of the flow chart. Basically, image processing techniques are used in order to accomplish this task. First as a pdf document is given as the input to the system, this pdf is broken down as pages and these pages are converted into images using pdf2image library files. As the second step, these images are processed, and flow chart diagrams are identified.

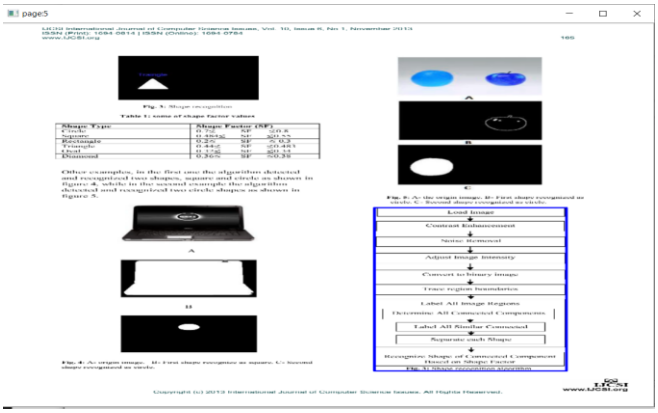

*Fig. 3. Identification of flow charts*

Next, these identified images are extracted and processed to get the content of them.

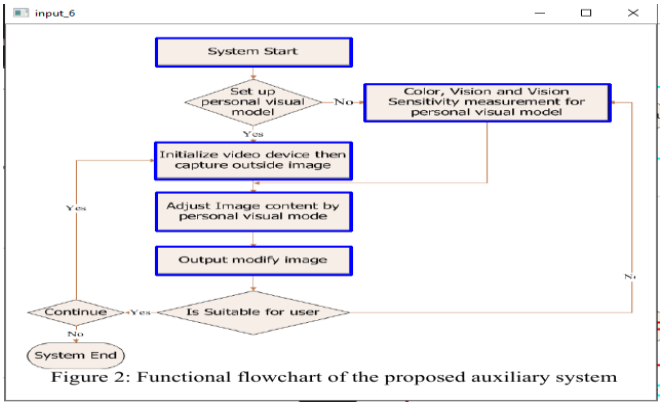


*Fig. 4. Identification of internal details of the flow chart*

The content of the image is extracted in the order of the flow chart using pytesseract library files. Recurrent Neural Network(RNN) algorithms are used for text extraction and identification. Dependent and independent nodes are identified and extracted in the order of the flow chart. Decision nodes are identified separately and mentioned in the output document. The output will be a pdf document with the information of the flow chart.

By this section of the system, a visually impaired student will be able to get to know whether there are flow charts in the research paper. If there are any flow charts they will be extracted, and the content will be given as that the visually impaired student will be able to imagine the flow chart through a pdf document. The student will be able to imagine the flow chart as the details are given in the exact order of the flow chart with the decision nodes.

Even if the visually impaired students get the chance to read a research paper using other tools, the student will not be able to identify flow charts and will not be able to get to know about what is written in the flow chart. Therefore, this section of the system will be a great help for the visually impaired students to do their studies.

*C. AI Chatbot*

The design of an AI Chatbot to interact with the user to simulate human conversation via text or audio messages to communicate with the application. A Generative based Conversational AI model will be used for the purpose of question answering and navigation throughout the web application. The model will take the user input and understand the intentions of the user's message , determine what type of response message is required, and follow correct grammatical and lexical rules while forming the response.

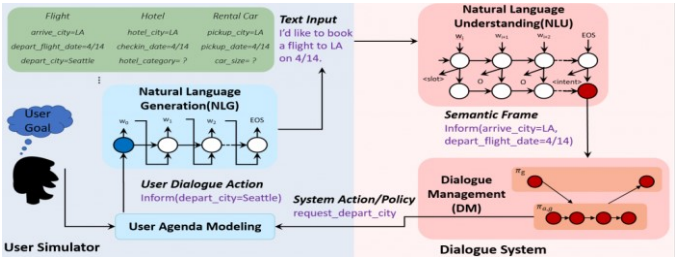


*Fig. 6. The figure shows how Deep Learning based chatbot work internally.*

*D. Graph information extraction tool based on the image*

A graph is an effective form of data representation used to summarize complex information. Understanding the graph details means an important task. When we speak about visually handicapped students, graph information extraction become a dream. Because without seeing the graph image, it is very difficult to identify the details behind the graph. In this section represent fully descriptive information related to the "extraction of the graph information" based on image contents.

Initially system will identify the curve graph images in the given pdf by using graph-based image segmentation algorithm. After that it will automatically convert the graph image into binary image by using thresholding .It will extract all the X, Y coordinates (use OCR and the mask creator to identify the coordinates) related to each curve graphs and listed in a separate excel sheet by using "xlswrite ()" function. It will process all the extracted information and finally give a complete explanation about the behavior of the graph with important details (ex: - minimum, maximum, slope, intercept, etc.)

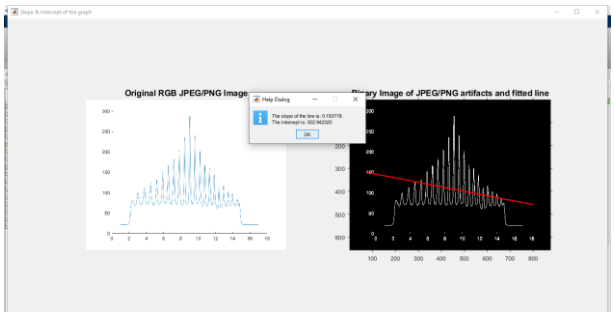


*Fig. 7. Original image to binary image conversion and identify the slope and the intercept.*

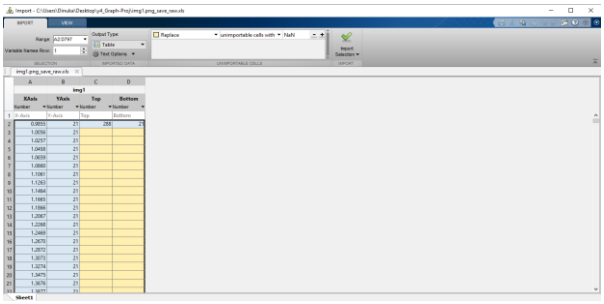

*Fig. 8. Extracted graph information.*

An identifying graph information is also become a huge task for not blind people. So final output become voice converted as well as the summarized documentation with including all the summarized details about the graphs. Therefore, people can easily get the idea about each and every graphs in detailed manner.

## V. CONCLUSION AND FUTURE DIMENSIONS

*A.Conclusion*

Outcome of this research produce an application to summarize research papers was built mainly focusing on visually impaired students. The whole solution as a system provides an application to select a research paper and get the summarization of it with a description about flow charts and graphs if present. This task will be done within a few minutes and with a high accuracy.

The summarization will be presented in voice and as a document which contains the summarized information in point form. Visually impaired students will be able to use this as a mobile application and a desktop application.

In addition, this application can be used by ordinary students to get a summary of a research paper very easily within few minutes without reading the paper.

*B.Future Dimensions*

Currently system is developed to summarize the methodology, but the authors hope to develop the application to summarize the whole research paper.

This application summarizes only the flow charts, it can be developed to identify and summarize all other types of diagrams related to the field of IT.

Currently system efficiently summarizes only research articles related to Information Technology. We expect to expand the range of this tool to summarize research articles related to any field.

## ACKNOWLEDGMENT

The authors would gratefully acknowledge Prof. Samantha Thelijjagoda, who lead us by giving suggestions, encouraging, and helping in coordinating the research activities. Furthermore, we thank Mr. Upul Weerasinghe (Assistant Director of Central Bank, Colombo) and Mr. Sanka who provided us with information on the needs of visually impaired students by giving us necessary ideas. We would appreciate the guidance given by the panel of lecturers who gave us their valuable ideas and time.

## REFERENCES

[1] Chinatsu Aone, Mary Ellen Okurowski, James Gorlinsky, ”Trainable, scalable summarization using robust NLP and machine learning”

[2] ]Eduard Hovy , Chin-Yew Lin, Automated text summarization in SUMMARIST .

[3] Vishal Gupta, A survey of text summarization extractive techniques, 2010.

[4] H. P. Edmundson.,” New methods in automatic extracting”, Journal of the ACM, 16(2):264-285, April 1969.

[5] J. Kupiec, J. Pedersen, and F. Chen, “A trainable document summarizer”, In Proceedings of the 18th ACMSIGIR Conference, pages 68-73, 1995

[6] Ronald Brandow, Karl Mitze, and Lisa F. Rau. “Automatic condensation of electronic publications by sentence selection. Information Processing and Management”, 31(5):675-685,1995.

[7] Hidetoshi Miyao and Rei Maruyama Faculty of Engineering, Shinshu University, Japan, ”On-line handwritten flowchart recognition, beautification, and editing system”.

[8] Mahmoud Younes, Basma AlMoshaigih, Michel Pasquier, Ghassan Qadah, “A sign reading system for the visually impaired”, International Conference on Frontiers in Handwriting Recognition,2012.

[9] Sachin Kumar Sa *, Parvathy Rajendrana , Prabaharan Pb , K P Somana, Text/Image Region Separation for Document Layout Detection of Old Document Images using Non-linear Diffusion and Level Set [online] Available:https://ac.els-cdn.com/S1877050916314776/1-s2.0-S1877050916314776-main.pdf?_tid=fb1ca9ae-6f82-4172-8241-b9197b8cd4d5&acdnat=1551705164_01be6405467f516266fbe755c6dce536.[Accessed:4-March-2019].

[10] EURASIP Journal on Image and Video Processing2017**2017**:82, “Textline detection in degraded historical document images” [online] Available:https://jivp-eurasipjournals.springeropen.com/articles/10.1186/s13640-017-0229-7. [Accessed: 4-March-2019].

[11] Graph-based representations and techniques for image processing and image analysis. A. Sanfeliua, R. Alquézarb, J. Andradea, J. Climentc, F. Serratosad and J.Vergésa. [Online] Available: https://pdfs.semanticscholar.org/2a15/d3f6127b8ff8d0a301b371641728e3178967.pdf [Accessed: 17- Feb- 2019].

[12] Graphs for image processing, analysis and pattern recognition. FlorenceTupinFlorence.

[13] Extraction of graph information based on image contents and the use of ontology. Sarunya Kanjanawattana1 and Masaomi Kimura2 1Graduate School 2 Information Science and Engineering Shibaura Institute of Technology, 3-5-7 Koto-ku Toyosu, Tokyo 135-8548, Japan[Online] Available: https://files.eric.ed.gov/fulltext/ED571596.pdf . [Accessed: 17-Feb-2019].

[14] Bhagwat, Vyas Ajay, "Deep Learning for Chatbots" (2018).Master's Master's Projects. 630

[15] Ravi, R. (2018). Intelligent Chatbot for Easy Web-Analytics Insights. 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI). [Online] Available: https://ieeexplore.ieee.org/document/4462971 [Accessed: 30 Apr-2019].