

**Application Integrating Optical Character
Recognition(OCR) and Text Summarization to
summarize lengthy articles**

Submitted in partial fulfillment of the requirements for the degree of

Bachelor of Technology
in
**Computer Science and Engineering with
Specialization in Bioinformatics**

by

ALLEN SALDANHA

18BCB0088

Under the guidance of

Dr. Sunil Kumar PV

School of Computer Science & Engineering

VIT, Vellore.



June, 2022

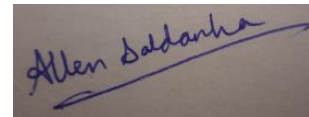
DECLARATION

I hereby declare that the thesis entitled “Application Integrating Optical Character Recognition(OCR) and Text Summarization to summarize lengthy articles” submitted by me, for the award of the degree of Bachelor of Technology in Computer Science and Engineering with specialization in Bioinformatics to VIT is a record of bonafide work carried out by me under the supervision of Dr. Sunil Kumar PV.

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Place : Vellore

Date : 03/06/2022

A handwritten signature in blue ink, reading "Allen Saldanha", with a long horizontal stroke extending to the right.

Signature of the Candidate

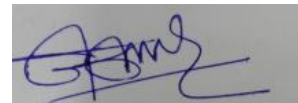
CERTIFICATE

This is to certify that the thesis entitled “Application Integrating Optical Character Recognition(OCR) and Text Summarization to summarize lengthy articles” submitted by Allen Saldanha, 18BCB0088, SCOPE, VIT University, for the award of the degree of Bachelor of Technology Computer Science and Engineering with specialization in Bioinformatics, is a record of bonafide work carried out by him under my supervision during the period, 01. 01. 2022 to 30.05.2022, as per the VIT code of academic and research ethics.

The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university. The thesis fulfills the requirements and regulations of the University and in my opinion meets the necessary standards for submission.

Place : Vellore

Date : 03.06.2022



Signature of the Guide

Internal Examiner

External Examiner

Dr. Priya G

B.Tech CSE with spl in BioInformatics And B.Tech CSE &Business
Systems

ACKNOWLEDGEMENTS

I would like to thank my guide, Dr. Sunil Kumar PV, SCOPE, Vellore Institute of technology, Vellore, for helping me in every step of the journey taken in developing this project titled “Application Integrating Optical Character Recognition(OCR) and Text Summarization to summarize lengthy articles”. He has provided me with the opportunity to learn and explore a lot more whilst working on this project.

I would also like to thank Dr. Priya G, my HOD, along with all the teaching staff and working members for their support in the pursuit of my degree. I express my gratitude to my Dr. K Ramesh Babu, Dean SCOPE for giving me this wonderful and nurturing experience.

I would also like to thank VIT university for giving me the chance to grow and challenge myself and giving a very fruitful educational experience. I express my gratitude to Dr. G. Vishwanathan and Mr. G. V. Selvam for providing me with a learning atmosphere to work in and their influence and inspiration during the tenure of the course.

Finally, I would like to thank my family and friends who have supported me throughout this journey for their love and blessings.

Allen Saldanha

Executive Summary

Nowadays people want the information to be crisp and to the point. To provide this there have been several applications that have come out. An example of this is an app called Inshorts, this is a news application that provides news in 60 words or less. The app is currently valued at anywhere between \$450-470 million.

With respect to the above context, the motivation behind this project is to be able to reduce the time taken to read articles. This is achieved by summarizing the articles for the user. The concept of the application works by taking a picture of the article that one would want to summarize and the application would recognize the text in the image and summarize the text. The point is to be able to reduce the time taken on reading lengthy articles which can be time-consuming and can sometimes distract the reader from what he/she is reading.

To build this particular application, flutter would be used which will be implemented with a Flask backend. The backend would implement OCR (Optical Character Recognition) to recognize the text and then a text summarizer to summarize the text that is recognized.

TABLE OF CONTENTS

Contents	Pg no.
Acknowledgement	i
Executive summary	ii
Table of contents	iii
List of figures	iv
List of abbreviations	v
1 Introduction	1
1.1 Objective	1
1.2 Motivation	1
1.3 Background	2
2 Project description and goals	3
3 Technical specifications	4
4 Design approach and details	5
4.1 Design approach/materials and methods	5
4.2 Codes and standards	8
4.3 Constraints, alternatives and tradeoffs	9
5 Schedule, tasks and milestones	10
6 Project demonstration	11
7 Result and discussion	18
8 Summary	19
References	20

List of Figures

Figure No.	Title	Page No.
1	Architecture model	5
2	Milestones flowchart	10
3	Homepage	11
4	Uploaded image being displayed	12
5	Recognized Text	13
6	Summarized Text	14
7	Pasting text page	15
8	Text pasted	16
9	Pasted text summarised	17

List of Abbreviations

NLP	Natural Language Processing
AI	Artificial Intelligence
RNN	Recurrent Neural Network
CNN	Convolutional Neural Network
OCR	Optical Character Recognition
GB	Gigabyte
TTS	Text-to-speech
OS	Operating System
API	Application Programming Interface

1. INTRODUCTION

1.1. OBJECTIVE

The objective of the project is to be able to reduce the time taken on reading lengthy articles which can be time-consuming and can sometimes distract the reader from what he/she is reading. Having an application that can give the reader the summary of the entire article with a simple click of a button would be a time saver. Having a mobile app in place for this kind of application would turn out to be very useful as the user can have a summarizer in their pockets and can be easily accessed when compared to a web app which could be tedious for certain situations. To build this particular application, flutter would be used which will be implemented with a Flask backend. The backend would implement OCR (Optical Character Recognition) to recognize the text and then a text summarizer to summarize the text that is recognized.

1.2 MOTIVATION

The expression time is money is very valid. It applies very efficiently in all aspects of life. Thus to reduce the time used and thereby save money there have been various developments that help achieve the same. Nowadays people want the information to be crisp and to the point. To provide this there have been several applications that have come out. An example of this is an app called Inshorts, this is a news application that provides news in 60 words or less. This saves people a lot of time and helps them grasp information in a short amount of time. The usage of this application has been of great importance which has brought the valuation of Inshorts to be valued anywhere between \$450-470 million. With respect to the above context, the motivation behind this project is to be able to reduce the time taken to read articles. Be it from journals or posts or even news articles. This is achieved by summarizing the articles for the user. The concept of the application works by taking a picture of the article that one would

want to summarize and the application would recognize the text in the image and summarize the text. To not restrict the scope of the application to just journals or posts or news articles, the end result would be a mobile application that can be used to summarize any long text. Apart from the above application of the project, the project has another application which would benefit the visually challenged people, where in the system would recognize/summarize the text and would allow users to convert the text to speech inorder to listen to the text that is on the article. This helps out as an option to be able to understand the text in articles for users that are visually challenged.

1.3 BACKGROUND

Optical Character Recognition (OCR) is an image processing application that is used to recognize characters from documents. The main advantage of this application is to be able to convert the text in a physical hardcopy form to a softcopy form, this helps in saving space. Text summarization is an NLP (Natural Language Processing) application that is used to convert long texts into a short text. It is used to condense content.

2. PROJECT DESCRIPTION AND GOALS

In this application, we will be using flutter to build a mobile application. This app would take a picture of an article and send it to the flask app for processing. In the process of processing the OCR would be deployed to recognize the text in the picture. And once the text is recognized, the text will be sent to another application where a text summarizer is deployed, and this text summarizer would summarize the recognized text and send it back to the flutter application to display to the user.

In addition to the above main application. The app also has a working text to speech function that reads out the text. This option is available on both the screens, that are the screen where the recognized text is displayed and the screen where the summarised text is displayed. There is also a copy text option which copies the text onto the clipboard.

Apart from being able to use an image and recognize the text, there is also an option to paste text and get that text summarized.

The goals that were meant to be achieved have been achieved, which include:

- Working routes in the backend
- Tesseract - OCR
- Summarizer
- An easy to use and simplistic frontend
- Text to speech
- Copy text to clipboard
- Pasted text summarizer
- Recognized content correction

3. TECHNICAL SPECIFICATIONS

As of now the app is fully ready to be deployed onto google play store after a few bugs checks that are to be undertaken. Flask was used for backend, Tesseract OCR was used for OCR, hugging face transformers was used for text summarization, flutter was used for the mobile app.

Hardware:

Dell 5370 - 8GB ram, integrated graphics card - This device runs the flask app on it and it connects the frontend that is the mobile app to the backend. The backend is running on the laptop

Mobile camera - 20 megapixels for capturing the content with good quality.

Software:

Python 3.7.9

Flutter 2.2.3

Dart 2.13.4

Tesseract 5.0.1.20220118

Allowed versions:

Any version of iOS after 9.0

Any versions of macOS and Safari above El Capitan

Any version of Android after API 19 (Kitkat)

Any version greater than windows 7

Chrome, Firefox and Edge should be above 84, 72.0 and 1.2.0 respectively

Version of Linux Debian that's above 10

4. DESIGN APPROACH AND DETAILS

4.1. DESIGN APPROACH

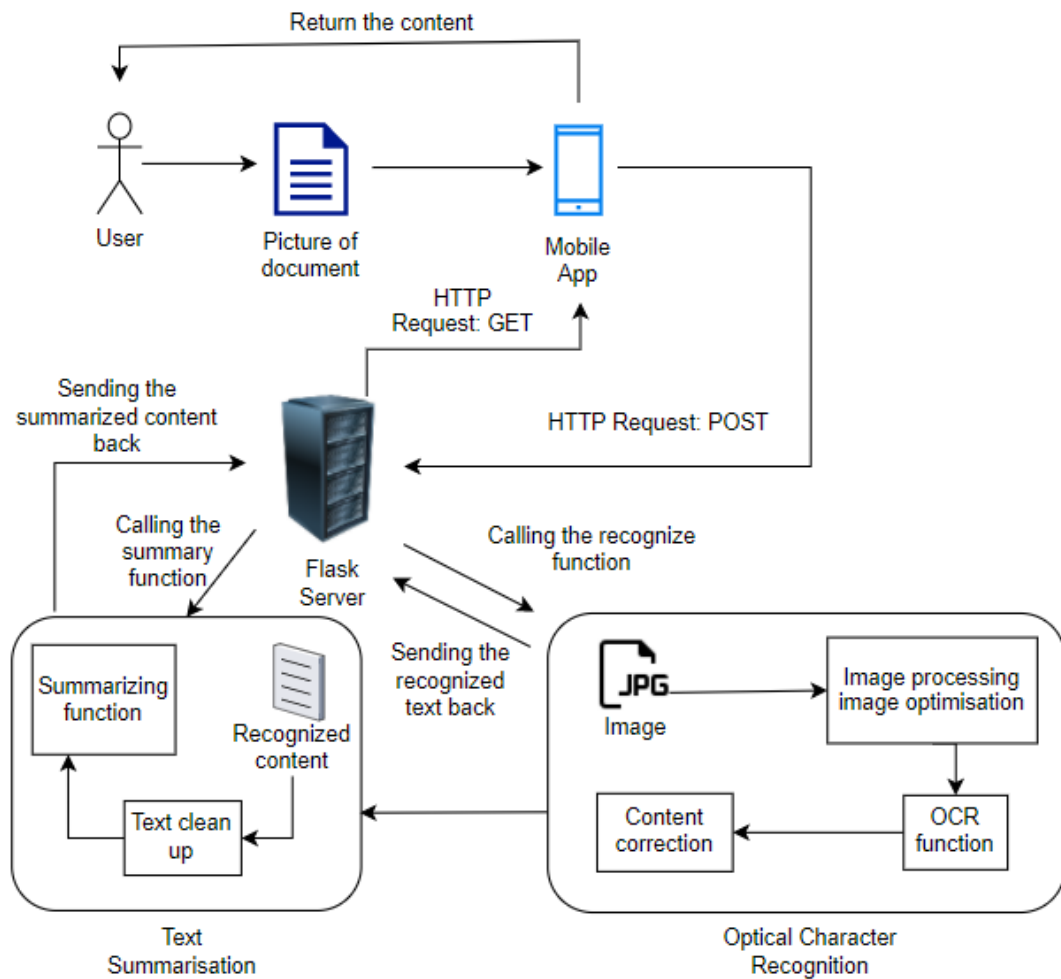


Figure 1:Architecture Model

Optical Character Recognition:

Optical Character Recognition or OCR as what it is abbreviated as, is an application of image processing, where the image processing techniques are used to be able to detect the characters in an image/live document. Image processing is the process of working on an image with the help of certain techniques and methods to get some information

from the image or alter an image or process the image in a way the user wants to. The application of OCR ranges from detecting text from government documents to help faster processing, to detecting the registration number of vehicles violating road rules. OCR has a 3 step process when processing images to recognize texts. The first step is a pre-processing step where the image is cleaned in order to make the processing easier for the machine and thereby increase accuracy. The next step is the actual algorithmic part where there are 2 methods to do the recognition. One of the methods tries to recognize the text, by comparing the characters with what is stored in its memory, while the other method recognizes lines and curves in the text. Tesseract runs the input through 2 passes where the algorithms checks the input twice to better recognize. Tesseract also uses neural networks to help understand words and sentences in a whole rather than individual characters. The next and final step is the post-processing step where the text that is recognized is checked whether it grammatically makes sense. And also checks if the word exists in its inbuilt dictionary upto a certain threshold to not auto-correct names that might be unique. Also for certain applications there are application specific optimizations to be able to help the recognition and ensure it can recognize the text with the available context. We use a python library called pytesseract to be able to access the tesseract library which has been open sourced and modified to suit the recognition needs of this particular application.

Text Summarization:

Text summarization is an application of natural language processing. Natural language processing is basically the computer system being able to understand natural language which humans communicate with. It is based on the linguistics section of Artificial intelligence and machine learning. Natural language processing or NLP as it is abbreviated is of great use in situations where processing of texts take place. This can be to understand the opinions in the text or understand the sentiment with which the text is used, or in the case of this application being able to summarize the text. Text summarization is the method of being able to condense information into smaller bits of information in order to process it much more easily. There are two types of text summarization, namely abstractive and extractive. Extractive text summarization is a form of summarization where there are no changes in words of any text from the

original text to the final text, but its more on the lines of finding the most common or sentence from the list that can act as the summary of the entire text. Abstractive summarization on the other hand generates a context for the text and creates a summary based on the same along with paraphrasing the content. The more older techniques of abstractive summarization include the use of RNN, but now transformers have started showing much better results. This is the reason why transformers are being used in this particular application. The python library huggingface's transformer is used to summarize texts.

Flutter:

Flutter is an open-source kit that has been developed by Google. It is a UI software kit that helps in building applications that spans various platforms. This ranges from mobile systems such as Android and iOS to the desktop systems such as Windows and macOS. In this application Flutter is used to provide a very user friendly interface.

Flask:

Flask is a python framework that helps create web applications with ease. Flask helps setup the backend for applications by providing multiple features and flexibility to implement the use case. Web applications can be easily built with this framework within a single python file. And since its using python, its opens up a lot of possibilities of applications thanks to python's extensive library. Flask is very customizable. In this particular application the entire backend runs on flask.

4.2. CODES AND STANDARDS

Tesseract - OCR pseudocode

- 1) Input the image
- 2) pre-process the image
- 3) Gathering outlines of characters of the text and storing them. Checking and relating child and grandchild outlines.
- 4) Storing the collected outlines as blobs
- 5) These blobs are arranged into lines and these lines are analyzed and are further split up into words based on the spacing
- 6) Next is a 2 pass recognizer, where the second pass tries to improve the recognition from the first pass
- 7) The last step checks alternate possibilities and also verifies with the inbuilt dictionary

Transformers:

A Transformer is a model architecture that doesn't use recurrence and instead draws relationships between the input and the output by using the attention method. It was released in 2017 in a paper titled "Attention is all you need". Prior to the release of Transformers, the most common sequence transduction models included complicated RNN or CNN with an encoder and a decoder. The Transformer has an encoder and decoder as well, but by eliminating recurrence in favour of attention mechanisms, it achieves substantially greater parallelization than RNNs and CNNs.

4.3 CONSTRAINTS, ALTERNATIVES, AND TRADEOFFS

Initially backend wise, the system to recognise images and summarise texts consumes a lot of physical space, so to put it in an app to be used directly wouldnt be feasible as the size of the app would then exceed 1 GB because of the backend alone. So the alternative to this would be to host the backend online and access it. The OCR engine in itself has an accuracy of 90.2%, but this is in the case when the image taken is clear, even after pre-processing. In the cases where the image is not clear at all or if its written in a different font which is differs by a great margin than the usual default font, then there is a good chance of error in the OCR systems results. In the process of choosing an algorithm that is efficient enough for text summarization, several algorithms were tried out, mostly extractive text summarization algorithms and the results in these didnt provide the required summary. The alternative to this was using an abstractive based text summarization which provide the desired results.

5. SCHEDULE, TASKS, AND MILESTONES

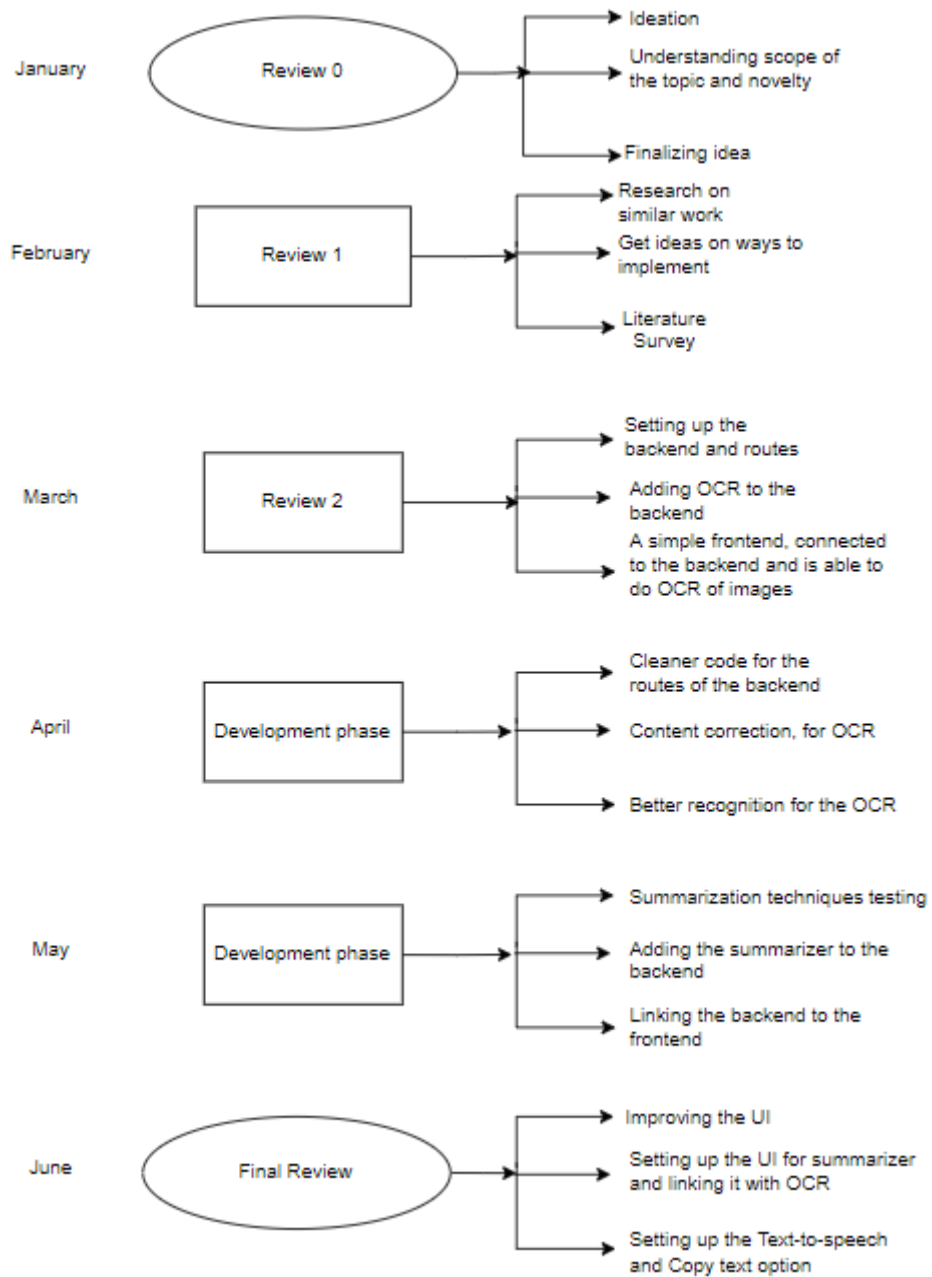


Figure 2: Milestones flowchart

6. PROJECT DEMONSTRATION

a) Home Page

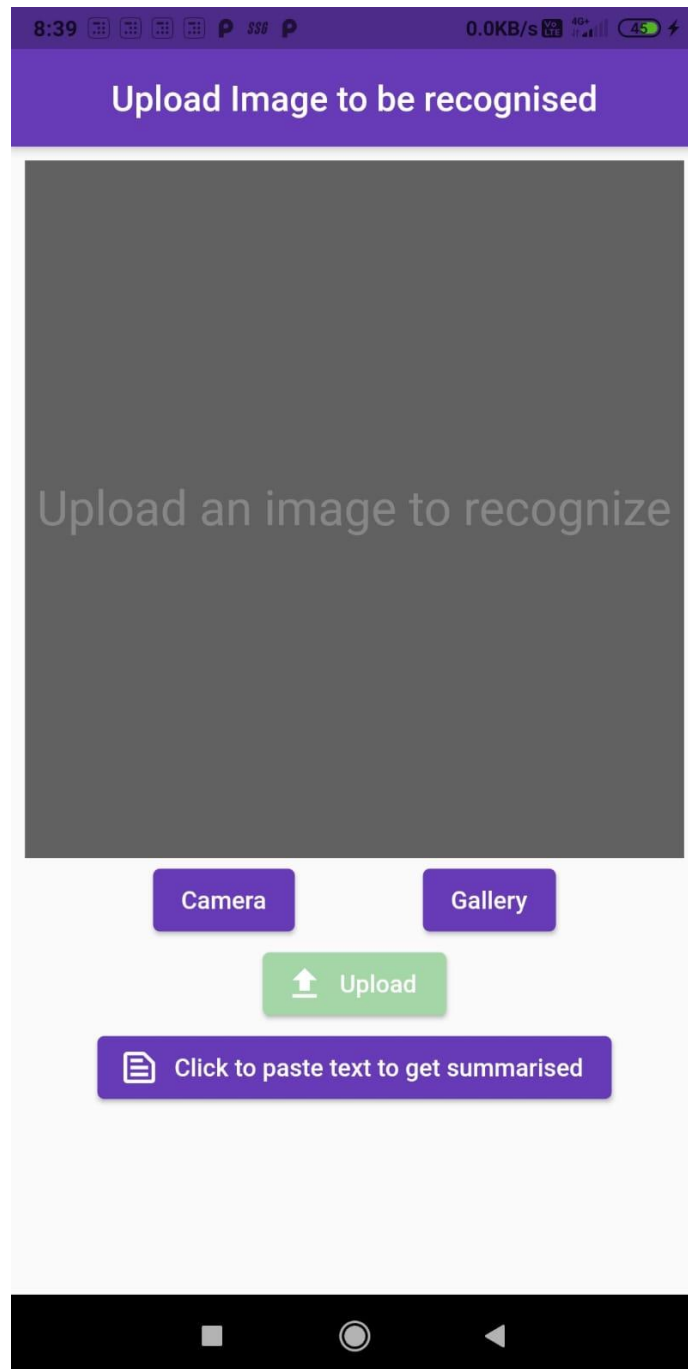


Figure 3: Homepage

b) Uploading first image taken from gallery



Figure 4: Uploaded image being displayed

c) Text recognized from the image

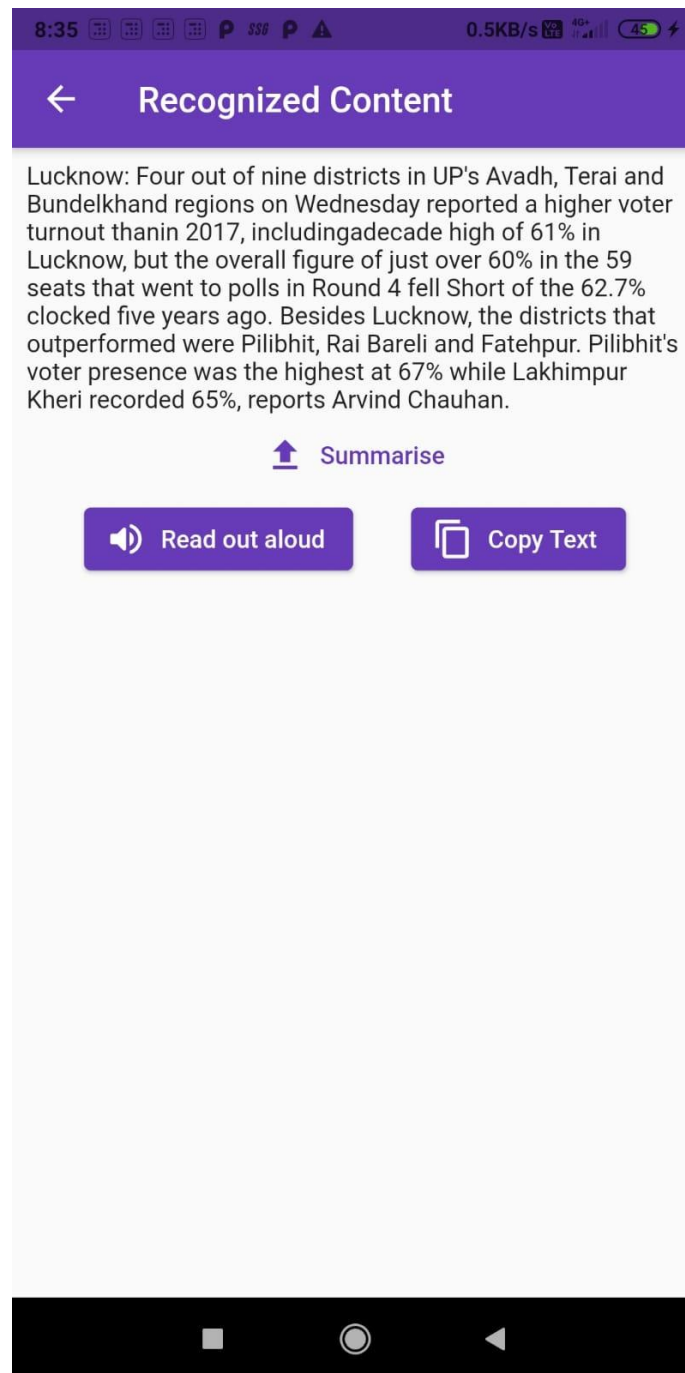


Figure 5: Recognized text

d) Text summarized from the image

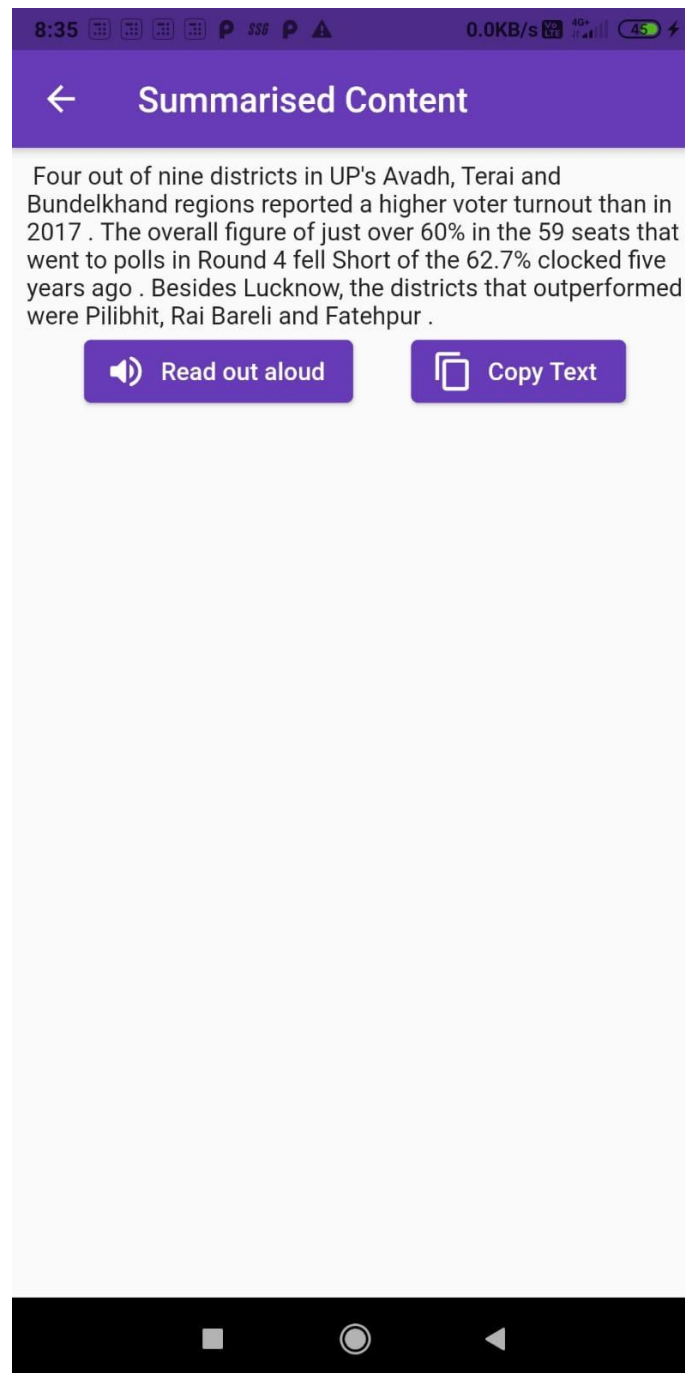


Figure 6: Summarized Text

e) Pasting the text screen

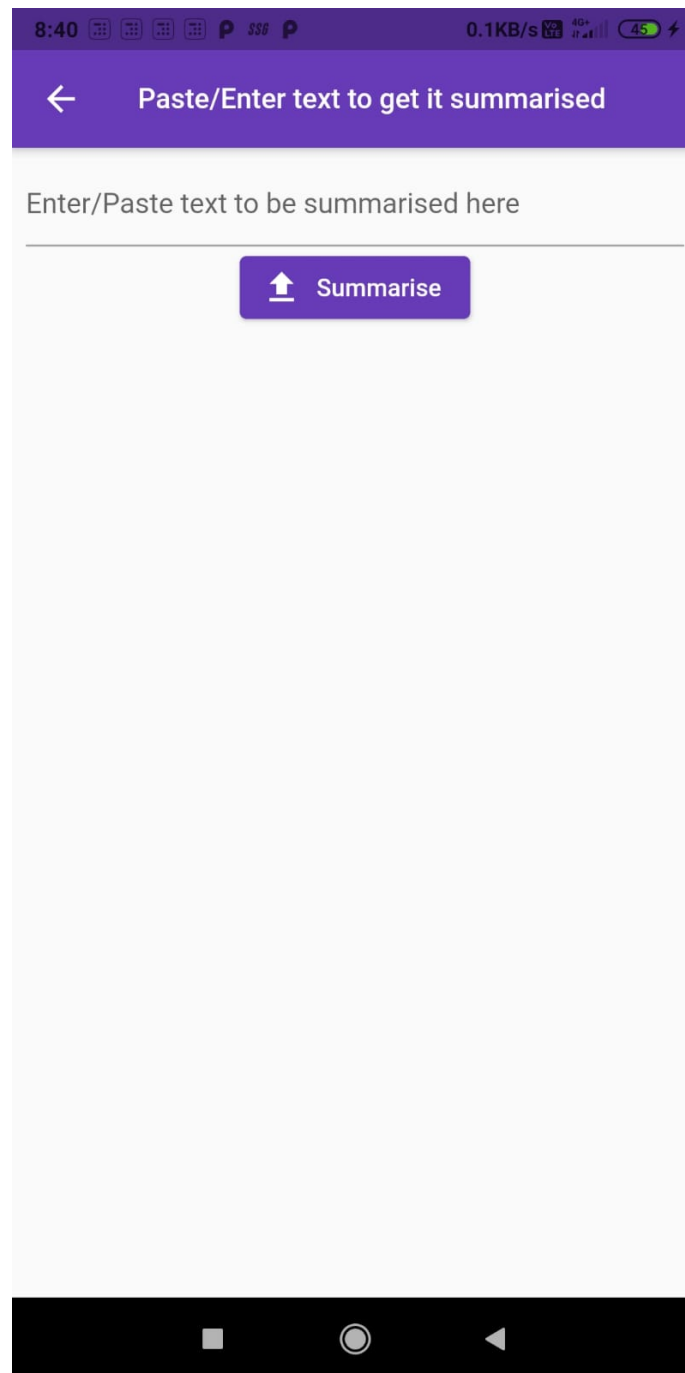


Figure 7: Pasting text page

f) Text copy and pasted from the abstract of a research article

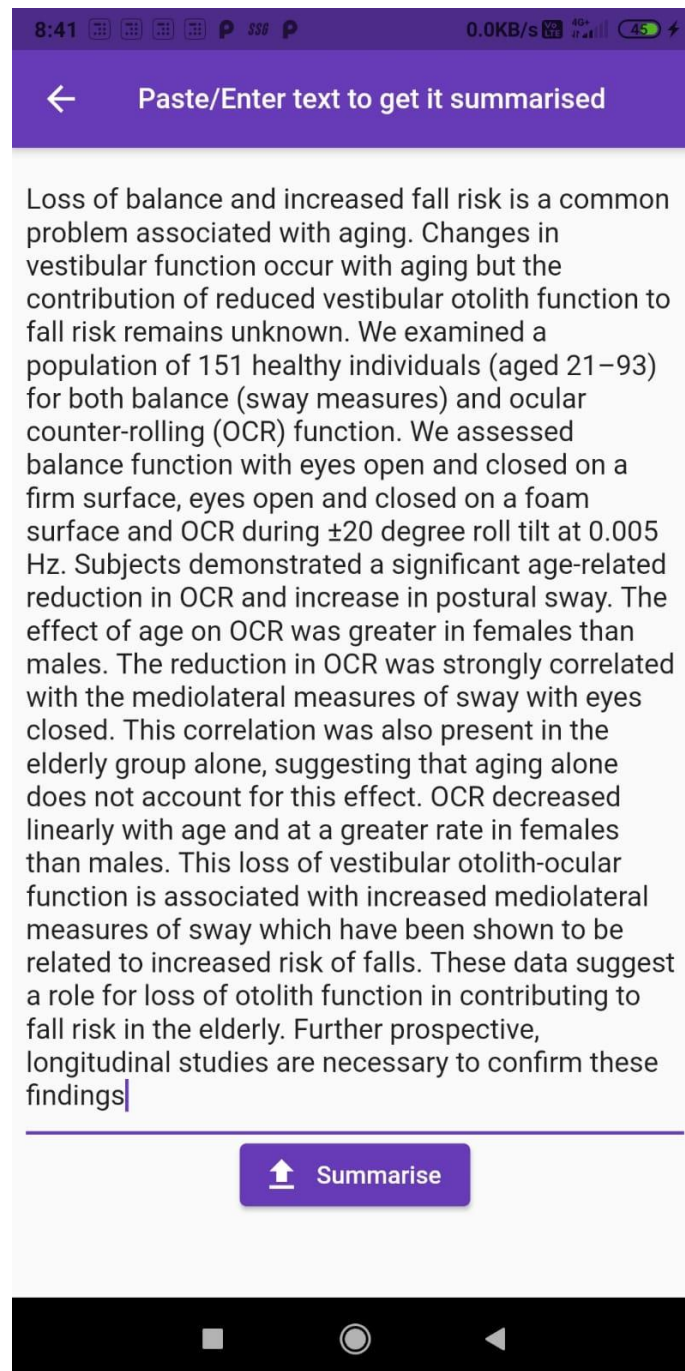


Figure 8: Text pasted

g) Pasted text summarized

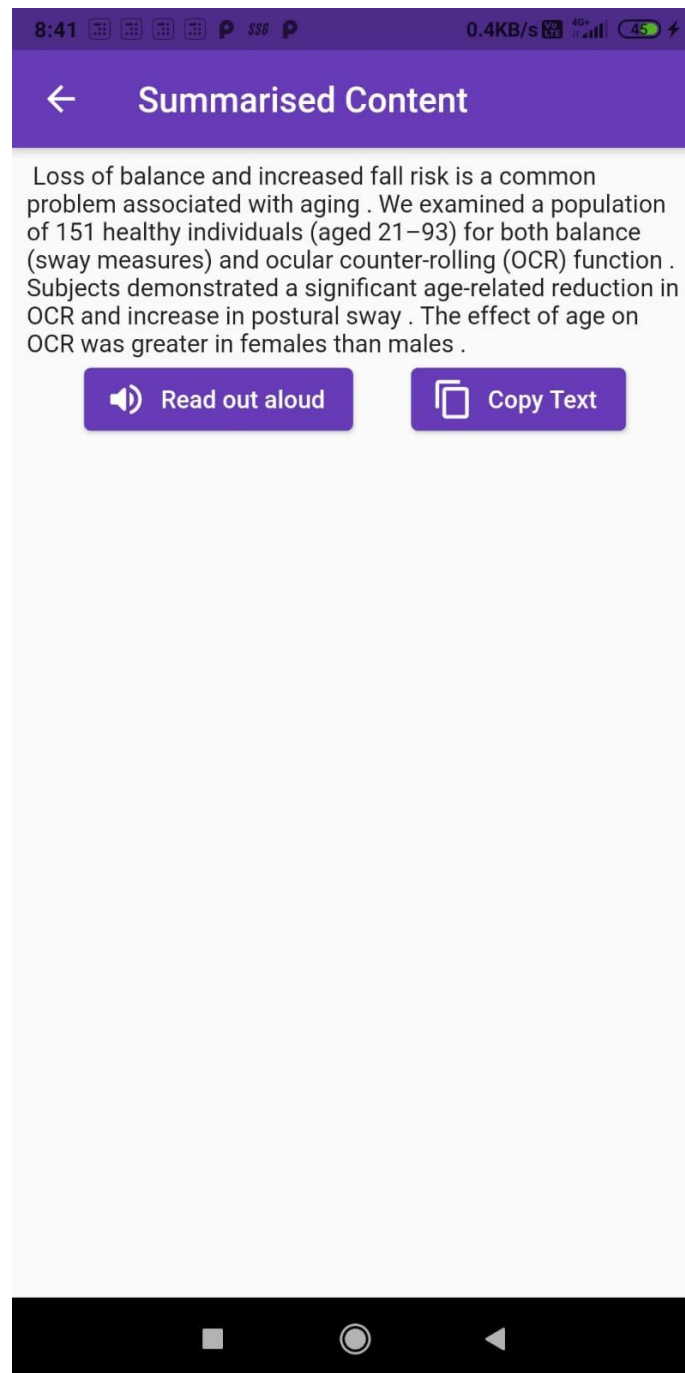


Figure 9: Pasted text summarised

7. RESULT AND DISCUSSION

The aim that was set out to develop the project has been achieved. The application is a fully working result of the proposed model. It is able to recognize the text and summarize the text as and when the user requests it.

From the test inputs that were tried out on the model, both the options worked as per the specifications. The recognized text had a few issues. But the rest were working well. The text-to-speech worked well and the summarizer worked too.

There is a wide range of audience that this application would benefit. After a thorough research into the customers that would be using this app, the following conclusion was drawn upon.

- Students/Researchers – That would want to get a hold of the physical text in a digital form or would want to get a particular text summarized
- Visually Challenged individuals – The app would have a text to speech feature which would read out the text
- General group of people that would want a gist of a long text

8. SUMMARY

Whilst working on this particular application, the understanding of the scope of the application increased. The application on its own as of now can recognize words with a 90.2% accuracy. The application is able to summarize the text that is recognized. And to not limiting the application to just recognition and summarization, a text to speech function is in place to read out the recognized/summarised text which works perfectly. This particular add-on is of great use to people that are visually challenged. In addition to this incase people have text that they would like to get summarised, the option for the same is available as well. Future works for this application would be to release it for public use into the app store. Future additions to whatever exists would be to integrate translation into the application. Also an option to add multiple images to be able to concatenate all the texts from all the images will also be something that would be added soon. The scope of this application can be very beneficial to a wide group individuals, especially students as I have received very good feedback from a lot of students on the idea of the topic.

References

- [1] Hubert, P. Phoenix, R. Sudaryono, D. Suhartono, "Classifying Promotion Images Using Optical Character Recognition and Naïve Bayes Classifier," *Procedia Computer Science*, 179, 498-506. 10.1016/j.procs.2021.01.033, Jan 2021
- [2] R.D. Suvaris, S. Sathyanarayana, "Broken Character Recognition using Connected Components and Convolutional Neural Network," *International Journal of Recent Technology and Engineering (IJRTE)*, Jan 2020
- [3] S. Drobac, K. Lindén, "Optical character recognition with neural networks and post-correction with finite state methods," *International Journal on Document Analysis and Recognition (IJDAR)*, 23, 279–295, Aug 2020
- [4] Z. Zhao, M. Jiang, S. Guo, Z. Wang, F. Chao and K. C. Tan, "Improving Deep Learning based Optical Character Recognition via Neural Architecture Search," *IEEE Congress on Evolutionary Computation (CEC)*, pp. 1-7, July 2020
- [5] J. Park, E. Lee, Y. Kim, I. Kang, H. I. Koo and N. I. Cho, "Multi-Lingual Optical Character Recognition System Using the Reinforcement Learning of Character Segmenter," in *IEEE Access*, vol. 8, pp. 174437-174448, Sep 2020
- [6] M. Li, T. Lv, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, F. Wei, "TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models," *arXiv:2109.10282v3*, Sep 2021
- [7] R. Matteo, N. Sven, and R. Bruce, "Optical Character Recognition of 19th Century Classical Commentaries: the Current State of Affairs. In The 6th International Workshop on Historical Document Imaging and Processing)," *Association for Computing Machinery*, New York, NY, USA, 1–6, Oct 2021
- [8] P. Divya, M. Varma, U. R. Mouli, Srinivas, Garima, Nikhil, Vishistha, "Web based optical character recognition application using flask and tesseract," *Materials Today: Proceedings*, ISSN 2214-7853, Jan 2021
- [9] R. Parthiban, R. Ezhilarasi and D. Saravanan, "Optical Character Recognition for English Handwritten Text Using Recurrent Neural Network," *International Conference on System, Computation, Automation and Networking (ICSCAN)*, pp. 1-5, July 2020
- [10] M. Ahmed, and A.I. Abidi, "REVIEW ON OPTICAL CHARACTER RECOGNITION," *International Research Journal of Engineering and Technology (IRJET)*, July 2019