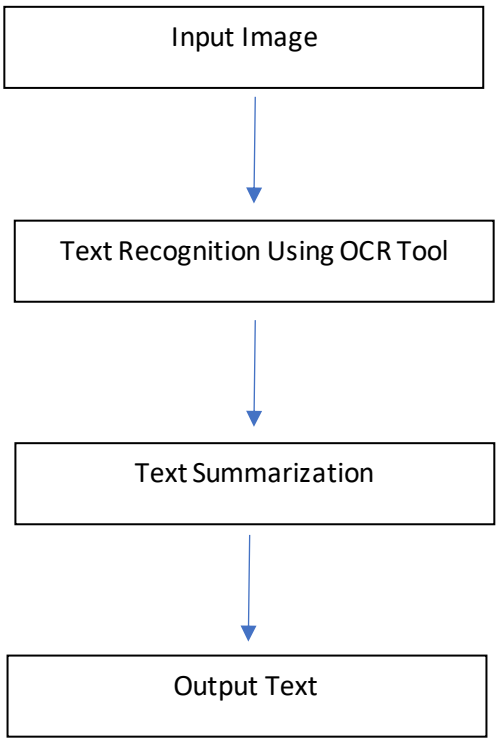
HINDI NEWS OCR-PROJECT REPORT

Project Plan:

The aim is to build an Optical Character Recognition (OCR) for Hindi or Devanagari (Indian Language). The user will input an image consisting of a piece of text from a newspaper or a magazine; this image will be pre-processed and passed onto the OCR tool - pytesseract. The OCR tool will read and recognize the text in the image. This text will be loaded into the summarization model that will give the final summary of the text.

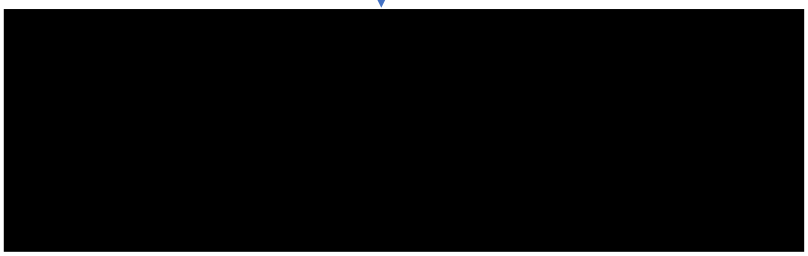
The processes involved are as follows:



An example of how the model will work:

इंटरनेट मीडिया में विज्ञापन देकर बेच रहे थे रेमडेसिविर इंजेक्शन, दो गिरफ्तार

Input Image



Text Recognition

इंटरनेट मीडिया में विज्ञापन देकर बेच रहे थे रेमडेसिविर इंजेक्शन, दो गिरफ्तार जासं, आगरा : कोरोना संक्रमण मेंरेमडेसिविर इंजेक्शन के सबसे कारगर होने की अफवाह इतनी ज्यादा फैली है कि युवक ने इंटरनेट मीडिया में विज्ञापन दे दिया। फोन करने पर जरूरतमंदों से 40 हजार रुपये कीमत मांगी। शुक्रवार शाम ग्राहक बनकर पहुंची आगरा पुलिस ने दोयुवकों को दबोच लिया। उनसे पूछताछ की जा रही है।आगरा के एसपी सिटी बोत्रे रोहन प्रमोद ने बताया कि शुक्रवार शाम कुछ लोगों ने उन्हें इंटरनेट मीडिया में दिए गए विज्ञापनके बारे में जानकारी दी। बताया कि विज्ञापनदाता युवक रेमडेसिविर इंजेक्शन के बदले में उनसे 40 से 60 हजार रुपये मांग रहा है। इस पर उन्होंने खुद ग्राहक बनकर उससे बात की। पहले तो उसने मना कर दिया, मगर कुछ देर बाद फोन कर 40 हजार रुपये में रेमडेसिविर इंजेक्शन लेने के लिए सिकंदरा हाईवे स्थित एक अस्पताल पर बुलाया। पुलिस सादे कपड़ों में वहां पहुंच गई। इंस्पेक्टर सदर अजय कौशल ने युवक से इंजेक्शन मांगा। युवक का साथी कुछ दूर इंजेक्शन लेकर खड़ा था। पुलिस ने दोनों को दबोच लिया। उनसे एक रेमडेसिविर इंजेक्शन बरामद हुआ है।उस पर लिखा मूल्य मिटा दिया गया था।एक आरोपित दयाल बाग का रहने वाला है और स्टेशनरी की दुकान चलाता है। दूसरा सिकंदरा के कारगिल शहीद पेट्रोल पंप के पास का रहने वाला है।आरोपित ने पूछताछ में बताया कि उसकी मां कोरोना संक्रमित हैं। यह इंजेक्शन उनके लिए खरीदे थे।

Text Extracted

Appling Summarization Model

कोरोना संक्रमण मेंरेमडेसिविर इंजेक्शन के सबसे कारगर होने की अफवाह इतनी ज्यादा फैली है कि युवक ने इंटरनेट मीडिया में विज्ञापन दे दिया। फोन करने पर जरूरतमंदों से 40 हजार रुपये कीमत मांगी। आगरा के एसपी सिटी बोत्रे रोहन प्रमोद ने बताया कि शुक्रवार शाम कुछ लोगों ने उन्हें इंटरनेट मीडिया में दिए गए विज्ञापनके बारे में जानकारी दी। बताया कि विज्ञापनदाता युवक रेमडेसिविर इंजेक्शन के बदले में उनसे 40 से 60 हजार रुपये मांग रहा है। इस पर उन्होंने खुद ग्राहक बनकर उससे बात की। पहले तो उसने मना कर दिया, मगर कुछ देर बाद फोन कर 40 हजार रुपये में रेमडेसिविर इंजेक्शन लेने के लिए सिकंदरा हाईवे स्थित एक अस्पताल पर बुलाया। पुलिस सादे कपड़ों में वहां पहुंच गई। उनसे एक रेमडेसिविर इंजेक्शन बरामद हुआ है। शुक्रवार शाम ग्राहक बनकर पहुंची आगरा पुलिस ने दोयुवकों को दबोच लिया।

Summary

Algorithms and Tools Used:

Text Recognition-

Python-tesseract is an optical character recognition (OCR) tool for python. That is, it will recognize and “read” the text embedded in images.

Text Summarization –

Using the NLTK library we built a simple summarization model that takes English text as input and returns its summary. To work with this model, we had to translate the Hindi text received from OCR to English, then run the summarization model. Lastly we translated the output from summarizer back to Hindi text. For translation we used the Google Trans API.

Dataset Used:

The OCR used the dataset provided by Tesseract which can be downloaded during installation or can be manually downloaded from: https://github.com/tesseract-ocr/langdata

Future Modifications:

- Enhance Text Detection - Though the OCR has good accuracy we can increase it further by adding some layers of processing using Open-CV to the input image to better recognize the text.
- Build a reliable Hindi Summarization Model using NLP to increase the accuracy of the summary.
- Additional Features – Using Open CV we can identify the Heading of the News so that the heading is displayed in the summary too. Currently the summarization model is skipping the heading as unwanted text.

Contributions:

Yash Bhanushali –

Used Pytesseract to build and implement the Hindi OCR.

Built an English summarization model and then implemented it using the translation approach using the googletrans API.

Documented the Project Report and README File.

Dhananjay Meena –

Tried building a Hindi summarization model using Huggingface Transformer.

Documented the README File.

Shubham Pal –

Could not contribute much due to some health emergencies.

Timeline:

12/04/2021-18/04/2021: Learning how Convolutional Neural Network (CNN) works via: https://www.youtube.com/watch?v=vT1JzLTH4G4&list=PL3FW7Lu3i5JvHM8ljYj-zLfQRF3EO8sYv

19/04/2021-23/04/2021: Working on classification of images using CNN on the CIFAR-10 dataset. https://pytorch.org/tutorials/beginner/blitz/cifar10_tutorial.html

24/04/2021-25/04/2021: Drafting Project Proposal.

26/04/2021-02/05/2021: Working on Pytesseract for text recognition.

02/04/2021-07/05/2021: Adding Text Summarizer and Additional features to the OCR.

Manav Saraf (Mentor)
Dhruv Grover (Mentor)
Yash Bhanushali
Shubham Pal
Dhananjay Meena