Problem Statement

Create an app from scratch which takes in an Insurance Policy document image and extracts important information from it in the form of text and auto-fills a form using it. Users should have a Camera Scan or Upload option to provide Insurance document images. Some common entities we are aiming to extract are Name, Email, Mobile No, Policy No, Vehicle Registration No, Engine No, Chassis No etc.

Challenges

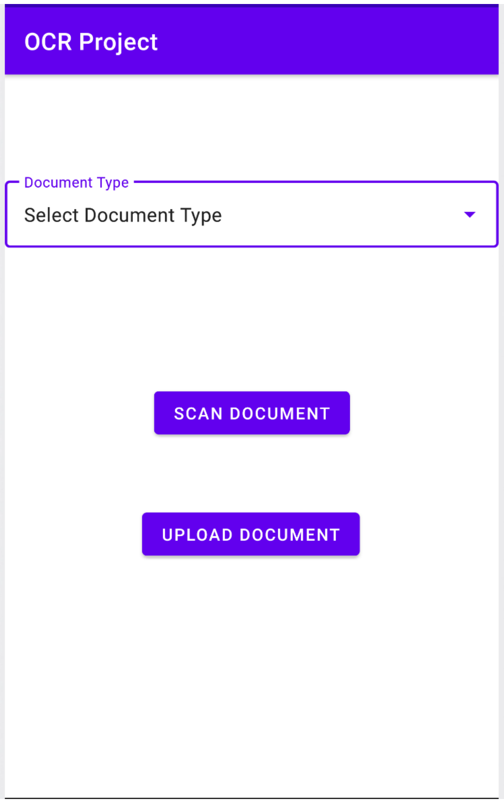
There are multiple challenges in this project. Different images require different image processing. So a camera scanned image requires adaptive thresholding as its lighting conditions vary across the image while an image which is a screenshot of a Policy Document will require thresholding techniques which calculates a global threshold value for the whole image. So the app distinguishes between camera scanned and uploaded image and applies image processing according to it, but if a user provides camera scanned image via upload method then this might lead to wrong results. Among the camera scanned images itself image processing will greatly vary like brightness and contrast adjustment, removing noise, removing blur to make text clearer etc. If a type of image processing is applied (like applying Noise filters to an image which is completely noise free) to the image then it will ruin the image and significantly impact the output results.

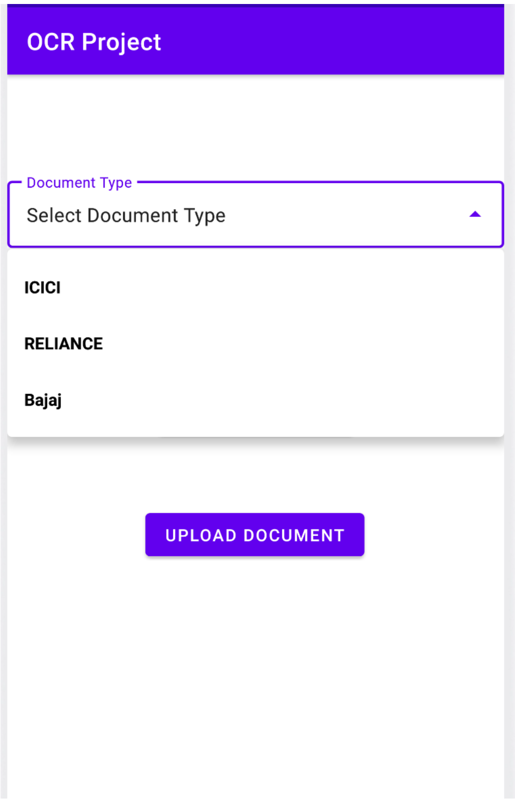
The OCR model used in the project should give output in a specific fixed format so that information extraction from it is reliable and accurate. OCR models like ML kit was able to detect different text boxes in the policy document but the output was some string with shuffled words inside it making it difficult to exactly locate the relevant information inside it. There are no reliable models which can exactly cater to the use case of extracting text from insurance policy documents which have a fixed structure. The OCR model used in this project is a generic OCR but the output it produced was something that could be utilized to extract relevant information.

This is an open ended project as documents greatly vary in their structure and information present in it. So each document type is handled separately according to its output structure. To add support for a new document its own static class has to be written.

App Workflow

The app is built on MVVM architecture so UI and its logic are separated in the code. The Main Page opens when the app is started. It has a drop down box and two buttons (Camera Scan and Upload). You have to select document type from the drop down present on the main screen. It is a required field without selecting which you cannot move further. After Selecting the document type press Scan or Upload button to move further.





After selecting the document type when Upload Document is pressed, file manager opens where relevant document images can be selected.
On pressing the Scan Document button, the device camera preview will open where the document image can be captured. CameraX apis are used to control the camera hardware in the device.

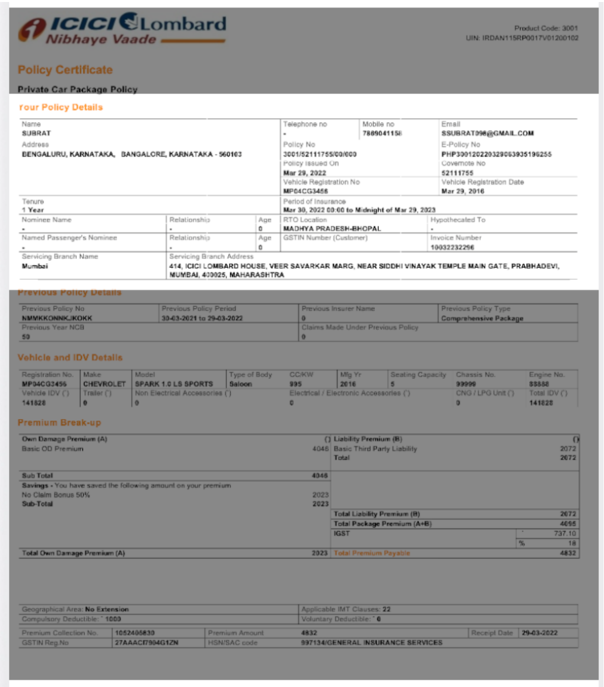
After selecting the image file an image cropper appears on the screen where the image can be cropped and the orientation of the image can be fixed. This step in the app is very important and it will have a significant impact on the output. After this the cropped image is pre processed and displayed on the screen before OCR model is run on it. If the ‘OK’ button is pressed then the OCR model is run and information is extracted from it. After that a final output page opens which has a form in which contents are auto filled.

Image Cropping

After the image is selected an image cropper appears on the screen. This step is very important as it will have a significant impact on the final output. If the image is cropped in the right way, it will reduce the processing time of the OCR model and improve its accuracy as it removes unwanted text and re-scales the image increasing its font size.

Some flexibility is provided when cropping the image but you are not completely free to crop in any way you want. The reason being that there are plethora of ways to crop an image, each cropped image resulting in a different output structure but some specific structure is needed to locate relevant information in it. Different structures from differently cropped images can be handled explicitly in the code but there will always be some case or the other where output of the OCR will give structure which would not be recognised in the code leading to wrong results.

The correct way of cropping the document image is to extend the cropping window from the left margin to the right margin. There is flexibility in the height of the cropping window but make sure that the value you want is present inside the window along with its key.
Some correct ways of cropping is shown below:



ICICI Lombard
Nibhaye Vaade

Product Code: 3001
UIN: IRDAN115RP0017V01200102

Policy Certificate

Private Car Package Policy

Your Policy Details

Name SUBRAT	Telephone no -	Mobile no 7869041158	Email SSUBRAT098@GMAIL.COM
Address BENGALURU, KARNATAKA, BANGALORE, KARNATAKA - 560103	Policy No 3001/52111755/00/000 Policy Issued On Mar 29, 2022		E-Policy No PHP300120220329053935196255 Covernote No 52111755
	Vehicle Registration No MP04CG3456		Vehicle Registration Date Mar 29, 2016
Tenure 1 Year	Period of Insurance Mar 30, 2022 00:00 to Midnight of Mar 29, 2023		

Nominee Name -	Relationship -	Age 0	RTO Location MADHYA PRADESH-BHOPAL	Hypothecated To -
Named Passenger's Nominee -	Relationship -	Age 0	GSTIN Number (Customer)	Invoice Number 10032232296
Servicing Branch Name Mumbai	Servicing Branch Address 414, ICICI LOMBARD HOUSE, VEER SAVARKAR MARG, NEAR SIDDHI VINAYAK TEMPLE MAIN GATE, PRABHADEVI, MUMBAI, 400025, MAHARASHTRA			

Previous Policy Details

Previous Policy No NMMKKONNKJKOKK	Previous Policy Period 30-03-2021 to 29-03-2022	Previous Insurer Name 0	Previous Policy Type Comprehensive Package
Previous Year NCB 50		Claims Made Under Previous Policy 0	

Vehicle and IDV Details

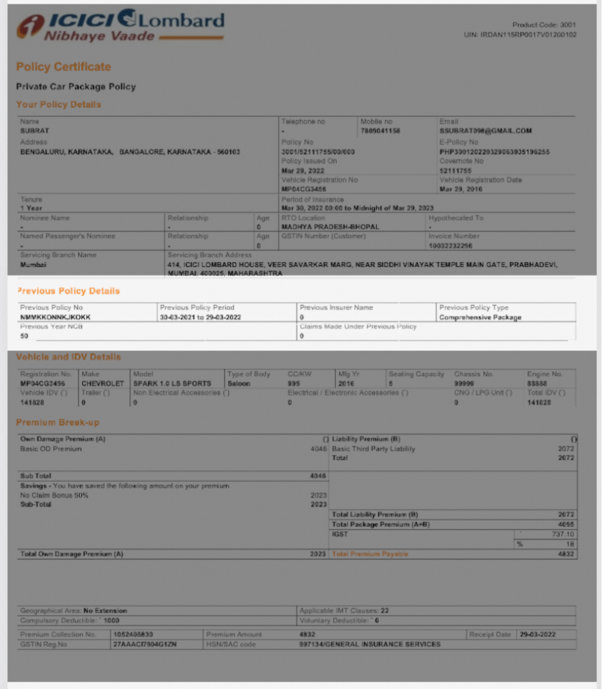
Registration No. MP04CG3456	Make CHEVROLET	Model SPARK 1.0 LS SPORTS	Type of Body Saloon	CC/KW 995	Mfg Yr 2016	Seating Capacity 5	Chassis No. 99999	Engine No. 88888
Vehicle IDV (') 141828	Trailer (') 0	Non Electrical Accessories (') 0		Electrical / Electronic Accessories (') 0			CNG / LPG Unit (') 0	Total IDV (') 141828

Premium Break-up

Own Damage Premium (A)	()	**Liability Premium (B)**	()
Basic OD Premium	4046	Basic Third Party Liability	2072
		Total	**2072**
Sub Total	**4046**		
Savings - You have saved the following amount on your premium			
No Claim Bonus 50%	2023		
Sub-Total	**2023**		
		Total Liability Premium (B)	2072
		Total Package Premium (A+B)	4095
		IGST	' 737.10
			% 18
Total Own Damage Premium (A)	2023	Total Premium Payable	4832

Geographical Area: **No Extension**	Applicable IMT Clauses: **22**
Compulsory Deductible: ' **1000**	Voluntary Deductible: ' **0**

Premium Collection No.	1052405830	Premium Amount	4832	Receipt Date	29-03-2022
GSTIN Reg.No	27AAACI7904G1ZN	HSN/SAC code	997134/GENERAL INSURANCE SERVICES		



ICICI Lombard
Nibhaye Vaade

Product Code: 3001
UIN: IRDAN115RP0017V01200102

Policy Certificate

Private Car Package Policy

Your Policy Details

Name SUBRAT | Telephone no - | Mobile no 7869041158 | Email SSUBRAT098@GMAIL.COM
Address BENGALURU, KARNATAKA,   BANGALORE, KARNATAKA - 560103 | Policy No 3001/52111755/00/000 | Policy Issued On Mar 29, 2022 | E-Policy No PHP300120220329053935196255 | Covernote No 52111755
Vehicle Registration No MP04CG3456 | Vehicle Registration Date Mar 29, 2016
Tenure 1 Year | Period of Insurance Mar 30, 2022 00:00 to Midnight of Mar 29, 2023
Nominee Name - | Relationship - | Age 0 | RTO Location MADHYA PRADESH-BHOPAL | Hypothecated To -
Named Passenger's Nominee - | Relationship - | Age 0 | GSTIN Number (Customer) | Invoice Number 10032232296
Servicing Branch Name Mumbai | Servicing Branch Address 414, ICICI LOMBARD HOUSE, VEER SAVARKAR MARG, NEAR SIDDHI VINAYAK TEMPLE MAIN GATE, PRABHADEVI, MUMBAI, 400025, MAHARASHTRA

Previous Policy Details

Previous Policy No NMMKKONNKJKOKK | Previous Policy Period 30-03-2021 to 29-03-2022 | Previous Insurer Name 0 | Previous Policy Type Comprehensive Package
Previous Year NCB 50 | Claims Made Under Previous Policy 0

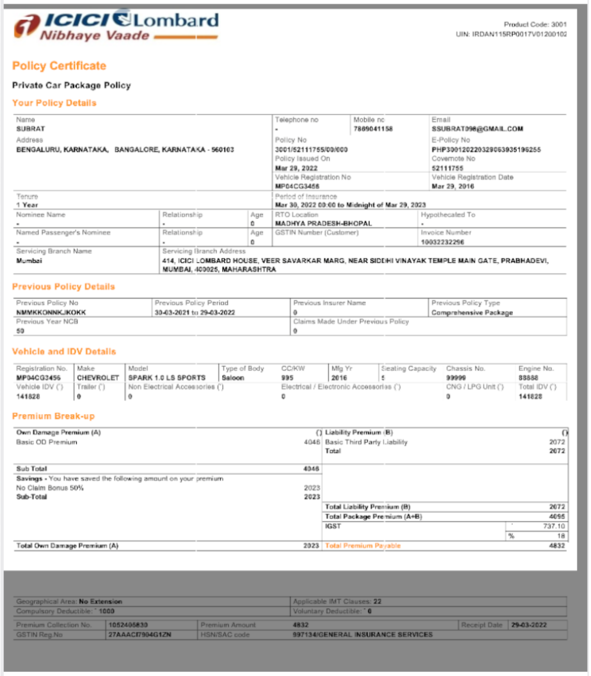
Vehicle and IDV Details

Registration No. MP04CG3456 | Make CHEVROLET | Model SPARK 1.0 LS SPORTS | Type of Body Saloon | CC/KW 995 | Mfg Yr 2016 | Seating Capacity 5 | Chassis No. 99999 | Engine No. 88888
Vehicle IDV (') 141828 | Trailer (') 0 | Non Electrical Accessories (') 0 | Electrical / Electronic Accessories (') 0 | CNG / LPG Unit (') 0 | Total IDV (') 141828

Premium Break-up

Own Damage Premium (A) () | Liability Premium (B) ()
Basic OD Premium 4046 | Basic Third Party Liability 2072
Total 2072
Sub Total 4046
Savings - You have saved the following amount on your premium
No Claim Bonus 50% 2023
Sub-Total 2023
Total Liability Premium (B) 2072
Total Package Premium (A+B) 4095
IGST ' 737.10 % 18
Total Own Damage Premium (A) 2023 | Total Premium Payable 4832

Geographical Area: No Extension | Applicable IMT Clauses: 22
Compulsory Deductible: ' 1000 | Voluntary Deductible: ' 0
Premium Collection No. 1052405830 | Premium Amount 4832 | Receipt Date 29-03-2022
GSTIN Reg.No 27AAACI7904G1ZN | HSN/SAC code 997134/GENERAL INSURANCE SERVICES



ICICI Lombard
Nibhaye Vaade

Product Code: 3001
UIN: IRDAN115RP0017V01200102

Policy Certificate

Private Car Package Policy

Your Policy Details

Name SUBRAT | Telephone no - | Mobile no 7869041158 | Email SSUBRAT098@GMAIL.COM
Address BENGALURU, KARNATAKA,   BANGALORE, KARNATAKA - 560103 | Policy No 3001/52111755/00/000 | Policy Issued On Mar 29, 2022 | E-Policy No PHP300120220329053935196255 | Covernote No 52111755
Vehicle Registration No MP04CG3456 | Vehicle Registration Date Mar 29, 2016
Tenure 1 Year | Period of Insurance Mar 30, 2022 00:00 to Midnight of Mar 29, 2023
Nominee Name - | Relationship - | Age 0 | RTO Location MADHYA PRADESH-BHOPAL | Hypothecated To -
Named Passenger's Nominee - | Relationship - | Age 0 | GSTIN Number (Customer) | Invoice Number 10032232296
Servicing Branch Name Mumbai | Servicing Branch Address 414, ICICI LOMBARD HOUSE, VEER SAVARKAR MARG, NEAR SIDDHI VINAYAK TEMPLE MAIN GATE, PRABHADEVI, MUMBAI, 400025, MAHARASHTRA

Previous Policy Details

Previous Policy No NMMKKONNKJKOKK | Previous Policy Period 30-03-2021 to 29-03-2022 | Previous Insurer Name 0 | Previous Policy Type Comprehensive Package
Previous Year NCB 50 | Claims Made Under Previous Policy 0

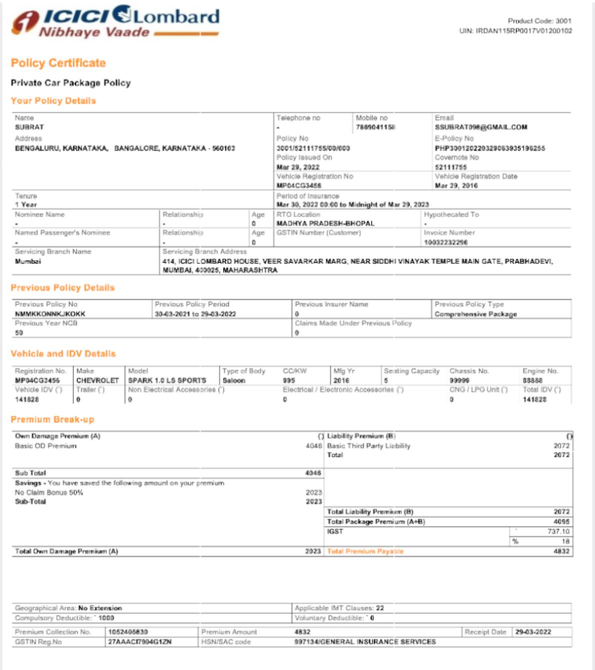
Vehicle and IDV Details

Registration No. MP04CG3456 | Make CHEVROLET | Model SPARK 1.0 LS SPORTS | Type of Body Saloon | CC/KW 995 | Mfg Yr 2016 | Seating Capacity 5 | Chassis No. 99999 | Engine No. 88888
Vehicle IDV (') 141828 | Trailer (') 0 | Non Electrical Accessories (') 0 | Electrical / Electronic Accessories (') 0 | CNG / LPG Unit (') 0 | Total IDV (') 141828

Premium Break-up

Own Damage Premium (A) () | Liability Premium (B) ()
Basic OD Premium 4046 | Basic Third Party Liability 2072
Total 2072
Sub Total 4046
Savings - You have saved the following amount on your premium
No Claim Bonus 50% 2023
Sub-Total 2023
Total Liability Premium (B) 2072
Total Package Premium (A+B) 4095
IGST ' 737.10 % 18
Total Own Damage Premium (A) 2023 | Total Premium Payable 4832

Geographical Area: No Extension | Applicable IMT Clauses: 22
Compulsory Deductible: ' 1000 | Voluntary Deductible: ' 0
Premium Collection No. 1052405830 | Premium Amount 4832 | Receipt Date 29-03-2022
GSTIN Reg.No 27AAACI7904G1ZN | HSN/SAC code 997134/GENERAL INSURANCE SERVICES



ICICI Lombard
Nibhaye Vaade

Product Code: 3001
UIN: IRDAN115RP0017V01200102

Policy Certificate

Private Car Package Policy

Your Policy Details

Name SUBRAT | Telephone no - | Mobile no 7869041158 | Email SSUBRAT098@GMAIL.COM
Address BENGALURU, KARNATAKA,   BANGALORE, KARNATAKA - 560103 | Policy No 3001/52111755/00/000 | Policy Issued On Mar 29, 2022 | E-Policy No PHP300120220329053935196255 | Covernote No 52111755
Vehicle Registration No MP04CG3456 | Vehicle Registration Date Mar 29, 2016
Tenure 1 Year | Period of Insurance Mar 30, 2022 00:00 to Midnight of Mar 29, 2023
Nominee Name - | Relationship - | Age 0 | RTO Location MADHYA PRADESH-BHOPAL | Hypothecated To -
Named Passenger's Nominee - | Relationship - | Age 0 | GSTIN Number (Customer) | Invoice Number 10032232296
Servicing Branch Name Mumbai | Servicing Branch Address 414, ICICI LOMBARD HOUSE, VEER SAVARKAR MARG, NEAR SIDDHI VINAYAK TEMPLE MAIN GATE, PRABHADEVI, MUMBAI, 400025, MAHARASHTRA

Previous Policy Details

Previous Policy No NMMKKONNKJKOKK | Previous Policy Period 30-03-2021 to 29-03-2022 | Previous Insurer Name 0 | Previous Policy Type Comprehensive Package
Previous Year NCB 50 | Claims Made Under Previous Policy 0

Vehicle and IDV Details

Registration No. MP04CG3456 | Make CHEVROLET | Model SPARK 1.0 LS SPORTS | Type of Body Saloon | CC/KW 995 | Mfg Yr 2016 | Seating Capacity 5 | Chassis No. 99999 | Engine No. 88888
Vehicle IDV (') 141828 | Trailer (') 0 | Non Electrical Accessories (') 0 | Electrical / Electronic Accessories (') 0 | CNG / LPG Unit (') 0 | Total IDV (') 141828

Premium Break-up

Own Damage Premium (A) () | Liability Premium (B) ()
Basic OD Premium 4046 | Basic Third Party Liability 2072
Total 2072
Sub Total 4046
Savings - You have saved the following amount on your premium
No Claim Bonus 50% 2023
Sub-Total 2023
Total Liability Premium (B) 2072
Total Package Premium (A+B) 4095
IGST ' 737.10 % 18
Total Own Damage Premium (A) 2023 | Total Premium Payable 4832

Geographical Area: No Extension | Applicable IMT Clauses: 22
Compulsory Deductible: ' 1000 | Voluntary Deductible: ' 0
Premium Collection No. 1052405830 | Premium Amount 4832 | Receipt Date 29-03-2022
GSTIN Reg.No 27AAACI7904G1ZN | HSN/SAC code 997134/GENERAL INSURANCE SERVICES

Image Pre-Processing

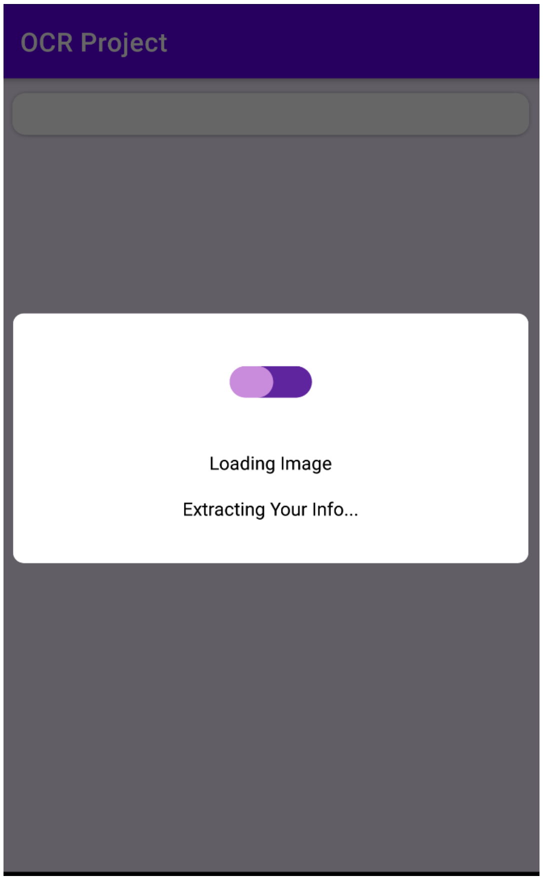
After cropping the image some preprocessing is done on it using OpenCV library. This step is important because it makes the image more clearer and readable for the OCR. In this step Gray scaling of the image is done. For a camera scanned image, first its contrast and brightness is adjusted a bit and after that Adaptive Thresholding is done on the image in which localized threshold value is generated according to the lighting conditions in that area of the image. For the uploaded image which is a screenshot of the Policy Document, results were better and more accurate without any pre-processing. Internal image processing of the OCR model is sufficient in that case.

OCR Model

As mentioned before a specific output format is needed to locate the data in the output text. The OCR model used in this project is Tesseract OCR which is one of the most popular models for OCR.
Training model used for Tesseract OCR is ‘tessdata_best’ which is the most accurate trained LSTM model. After pre-processing of the image, it is used for text extraction using Tesseract.
The main reason for using this model was that it does OCR and provides an output string which has line by line text; as it is present in the document there is no shuffling of lines or words.
Tesseract is set to recognise only the characters present inside the below string:

```
"-%*&#!@/,.:'\"0123456789ABCDEFGHIJKLMNOPQRSTUVWXYZabcdefghijklmnopqrs
tuvwxyz "
```

It is done to make sure that OCR doesn’t recognise some unnecessary characters.
The model runs on the background thread while a loading spinner runs on the main thread so that Tesseract doesn’t block the main thread ensuring smooth UI experience.



Loading Image

Extracting Your Info...

Logic Used for Information Extraction

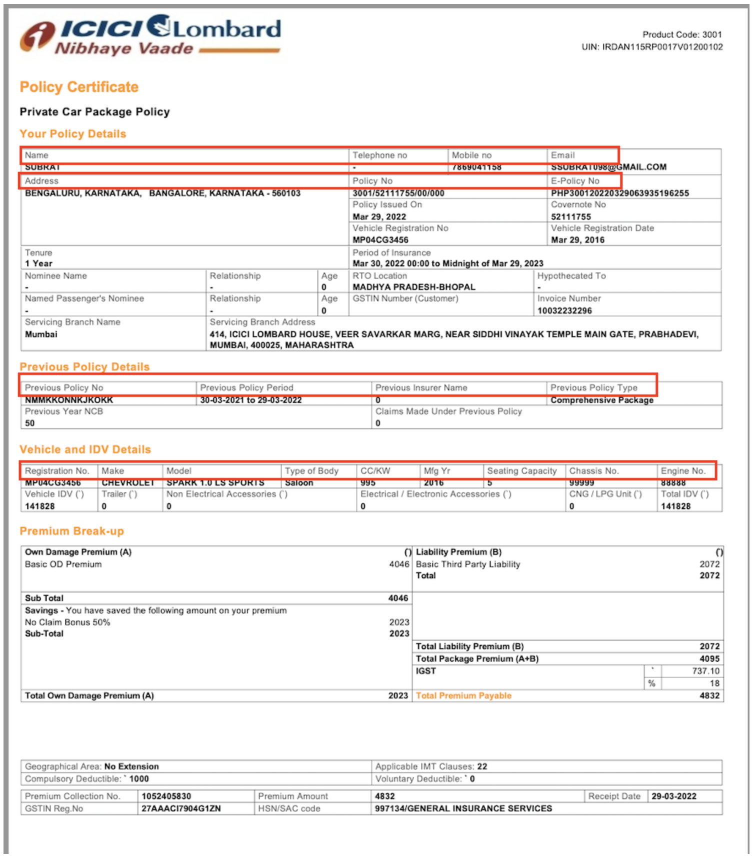
Logic for different types of documents are different specific to their structure. The underlying concept is the same but the way of implementing it is different. There are different static classes for different policy types. As mentioned before, Tesseract provides an output string that has line by line result according to the document. So there are some specific lines, phrases or words that are being detected in the output. These lines, phrases or words are common for all the documents of that particular type, for e.g., The word “Name” will be present in all the documents but its value may differ or a line which is like a heading or a key which will be present in all the documents of that type.

So these specific lines/phrases/words that we are trying to detect will be matched with the output string of the OCR line by line and a matching percentage is calculated.
For calculating matching percentage, Longest Common Subsequence (LCS) is calculated divided by the max of total characters in both the lines.

Note:: Matching could also have been calculated by detecting the occurrence of the line in the output of the OCR with the line we were trying to detect as a substring using KMP algorithm but the reason it was not done was to handle the inaccuracies in the output of OCR result. If OCR misreads some character (like spelling mistakes) or omits some char or adds some character then substring matching will not work. So LCS matching made sure that matching is flexible and code doesn't fail due to some small errors from the OCR.

Information Extraction Techniques Used

ICICI Document :



ICICI Lombard
Nibhaye Vaade

Product Code: 3001
UIN: IRDAN115RP0017V01200102

Policy Certificate

Private Car Package Policy

Your Policy Details

Name SUBRAT	Telephone no -	Mobile no 7869041158	Email SSUBRAT098@GMAIL.COM
Address BENGALURU, KARNATAKA, BANGALORE, KARNATAKA - 560103	Policy No 3001/52111755/00/000		E-Policy No PHP300120220329063935196255
	Policy Issued On Mar 29, 2022		Covernote No 52111755
	Vehicle Registration No MP04CG3456		Vehicle Registration Date Mar 29, 2016
Tenure 1 Year	Period of Insurance Mar 30, 2022 00:00 to Midnight of Mar 29, 2023		

Nominee Name -	Relationship -	Age 0	RTO Location MADHYA PRADESH-BHOPAL	Hypothecated To -
Named Passenger's Nominee -	Relationship -	Age 0	GSTIN Number (Customer)	Invoice Number 10032232296
Servicing Branch Name Mumbai	Servicing Branch Address 414, ICICI LOMBARD HOUSE, VEER SAVARKAR MARG, NEAR SIDDHI VINAYAK TEMPLE MAIN GATE, PRABHADEVI, MUMBAI, 400025, MAHARASHTRA			

Previous Policy Details

Previous Policy No	Previous Policy Period	Previous Insurer Name	Previous Policy Type
NMMKKONNKJKOKK	30-03-2021 to 29-03-2022	0	Comprehensive Package
Previous Year NCB 50		Claims Made Under Previous Policy 0	

Vehicle and IDV Details

Registration No.	Make	Model	Type of Body	CC/KW	Mfg Yr	Seating Capacity	Chassis No.	Engine No.
MP04CG3456	CHEVROLET	SPARK 1.0 LS SPORTS	Saloon	995	2016	5	99999	88888
Vehicle IDV (`) 141828	Trailer (`) 0	Non Electrical Accessories (`) 0		Electrical / Electronic Accessories (`) 0			CNG / LPG Unit (`) 0	Total IDV (`) 141828

Premium Break-up

Own Damage Premium (A)	(`)	Liability Premium (B)		(`)
Basic OD Premium	4046	Basic Third Party Liability		2072
		Total		2072
Sub Total	4046			
Savings - You have saved the following amount on your premium				
No Claim Bonus 50%	2023			
Sub-Total	2023			
		Total Liability Premium (B)		2072
		Total Package Premium (A+B)		4095
		IGST	`	737.10
			%	18
Total Own Damage Premium (A)	2023	Total Premium Payable		4832

Geographical Area: No Extension	Applicable IMT Clauses: 22
Compulsory Deductible: ` 1000	Voluntary Deductible: ` 0

Premium Collection No.	1052405830	Premium Amount	4832	Receipt Date	29-03-2022
GSTIN Reg.No	27AAACI7904G1ZN	HSN/SAC code	997134/GENERAL INSURANCE SERVICES		

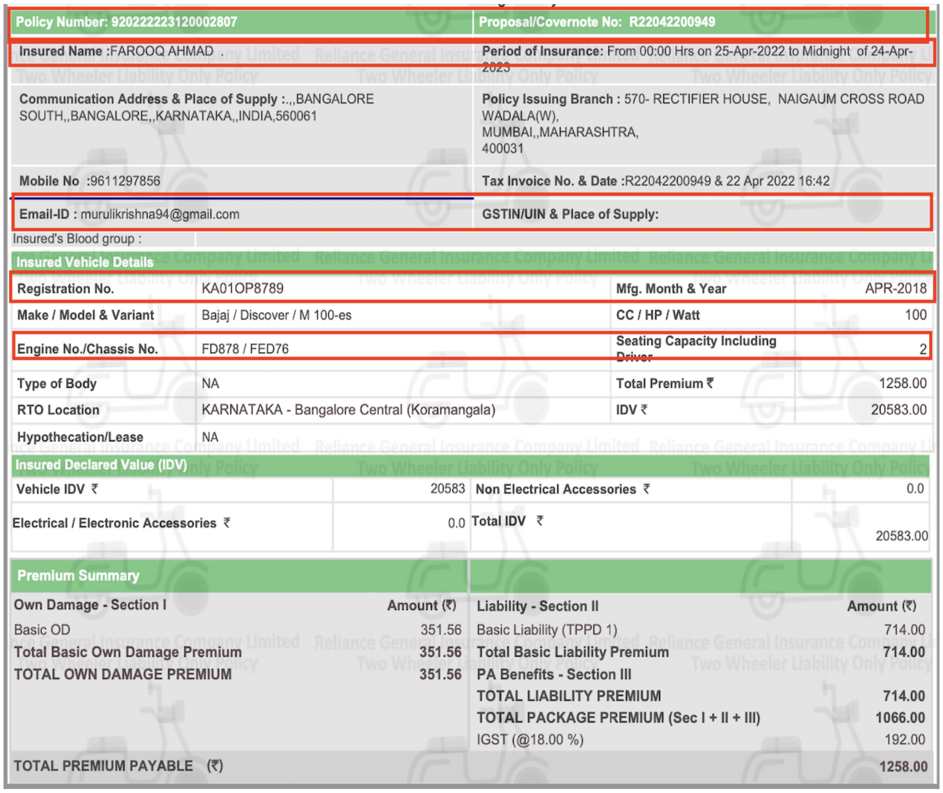
In the above document you can see that there are some red rectangular boxes present. In the output of the OCR, text inside these boxes will be read as a line. So in the code these lines are being detected.

For eg., the string "Name Telephone No Mobile No Email" is being detected in the output lines of the OCR. To detect it, this string is matched with every line in the output and a matching percentage is calculated (as described above). Among all the matching percentages, the line which has the highest matching percentage with the string is considered. If the best matching percentage is more than 50%, then we know by the structure pattern of Tesseract output that the values of that string will be present in the next line of that output (just like how it is present in the document).

Similarly, other strings like "Address Policy No E-Policy No" , "Registration No Make Model Type of Body CC/KW Mfg Yr Seating Capacity Chassis No. Engine No." etc is being detected using the same method and its value is then extracted from the next line of that best matched output line.

Value of the Model is sandwiched between the values of Make and Type of Body. So to extract it the first word from the left is taken as value of Vehicle Registration No, after that a dictionary is being used to detect Make from the left and Type of Body from the right. After detecting values of Make and Type of Body all the remaining words in between them are concatenated and assigned as the value of Model.

Reliance Document :



Policy Number: 920222223120002807	**Proposal/Covernote No: R22042200949**
Insured Name :FAROOQ AHMAD .	**Period of Insurance:** From 00:00 Hrs on 25-Apr-2022 to Midnight of 24-Apr-2023
Communication Address & Place of Supply :.,,BANGALORE SOUTH,,BANGALORE,,KARNATAKA,,INDIA,560061	**Policy Issuing Branch :** 570- RECTIFIER HOUSE, NAIGAUM CROSS ROAD WADALA(W), MUMBAI,,MAHARASHTRA, 400031
Mobile No :9611297856	**Tax Invoice No. & Date :**R22042200949 & 22 Apr 2022 16:42
Email-ID : murulikrishna94@gmail.com	**GSTIN/UIN & Place of Supply:**
Insured's Blood group :	

Insured Vehicle Details

Registration No.	KA01OP8789	**Mfg. Month & Year**	APR-2018
Make / Model & Variant	Bajaj / Discover / M 100-es	**CC / HP / Watt**	100
Engine No./Chassis No.	FD878 / FED76	**Seating Capacity Including Driver**	2
Type of Body	NA	**Total Premium ₹**	1258.00
RTO Location	KARNATAKA - Bangalore Central (Koramangala)	**IDV ₹**	20583.00
Hypothecation/Lease	NA		

Insured Declared Value (IDV)

Vehicle IDV ₹	20583	**Non Electrical Accessories ₹**	0.0
Electrical / Electronic Accessories ₹	0.0	**Total IDV ₹**	20583.00

Premium Summary

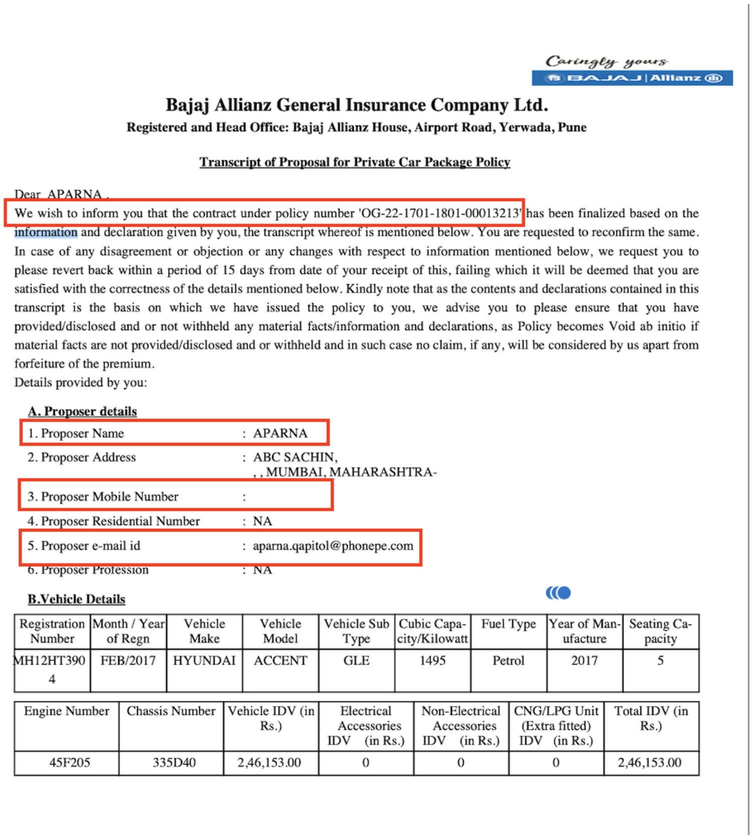
Own Damage - Section I	**Amount (₹)**	**Liability - Section II**	**Amount (₹)**
Basic OD	351.56	Basic Liability (TPPD 1)	714.00
Total Basic Own Damage Premium	**351.56**	**Total Basic Liability Premium**	**714.00**
TOTAL OWN DAMAGE PREMIUM	**351.56**	**PA Benefits - Section III**	
		TOTAL LIABILITY PREMIUM	**714.00**
		TOTAL PACKAGE PREMIUM (Sec I + II + III)	**1066.00**
		IGST (@18.00 %)	192.00
TOTAL PREMIUM PAYABLE (₹)			**1258.00**

In this document you can see lines that are being detected inside the red coloured boxes. For this document there is an observation that each line has exactly 2 keys that can be detected. Like in line 2 in above image, “Insured Name :FAROOQ AHMED Period of Insurance: From 00:00 Hrs on 25-Apr-2022 to Midnight of 24-Apr-“, there are 2 keys that can be detected here ‘Insured Name’ and ‘Period of Insurance’.

Similarly, for the 4th red box in the above image, the line “Registration No. KA01OP8789 Mfg. Month & Year APR-2018”, there are exactly 2 keys that can be detected here, ‘Registration No.’ and ‘Mfg. Month & Year’.

So to detect these keys each line of the output is broken in 4 parts about the spaces. 1st & 3rd part being the keys and 2nd & 4th part being its values respectively. So 1st and 3rd are being matched with the keys using the above mentioned algorithm, keys with best matching percentage (more than 50% or 70% in case of small words as small words can easily cross 50% threshold) is considered and the 2nd and 4th parts of the line are assigned to its corresponding keys.

Bajaj Document :



Bajaj Allianz General Insurance Company Ltd.

Registered and Head Office: Bajaj Allianz House, Airport Road, Yerwada, Pune

Transcript of Proposal for Private Car Package Policy

Dear APARNA ,
We wish to inform you that the contract under policy number 'OG-22-1701-1801-00013213' has been finalized based on the information and declaration given by you, the transcript whereof is mentioned below. You are requested to reconfirm the same. In case of any disagreement or objection or any changes with respect to information mentioned below, we request you to please revert back within a period of 15 days from date of your receipt of this, failing which it will be deemed that you are satisfied with the correctness of the details mentioned below. Kindly note that as the contents and declarations contained in this transcript is the basis on which we have issued the policy to you, we advise you to please ensure that you have provided/disclosed and or not withheld any material facts/information and declarations, as Policy becomes Void ab initio if material facts are not provided/disclosed and or withheld and in such case no claim, if any, will be considered by us apart from forfeiture of the premium.
Details provided by you:

A. Proposer details

1. Proposer Name : APARNA
2. Proposer Address : ABC SACHIN,
   , , MUMBAI, MAHARASHTRA-
3. Proposer Mobile Number :
4. Proposer Residential Number : NA
5. Proposer e-mail id : aparna.qapitol@phonepe.com
6. Proposer Profession : NA

B.Vehicle Details

Registration Number	Month / Year of Regn	Vehicle Make	Vehicle Model	Vehicle Sub Type	Cubic Capa-city/Kilowatt	Fuel Type	Year of Man-ufacture	Seating Ca-pacity
MH12HT390 4	FEB/2017	HYUNDAI	ACCENT	GLE	1495	Petrol	2017	5

Engine Number	Chassis Number	Vehicle IDV (in Rs.)	Electrical Accessories IDV (in Rs.)	Non-Electrical Accessories IDV (in Rs.)	CNG/LPG Unit (Extra fitted) IDV (in Rs.)	Total IDV (in Rs.)
45F205	335D40	2,46,153.00	0	0	0	2,46,153.00

Due to the limitations of the OCR model inside the black boxes, information (like vehicle registration number) were not extracted. For the remaining information like Policy Number, Proposer Name, Proposer Mobile Number and Proposer e-mail id which are present inside the red boxes in the above image, they were extracted by using the simple key matching. So each line inside the red box was split in two, about space (considered all the ways we can split). So the first part was matched with the keys mentioned like ‘Proposer Name’ and the best matching percentage was calculated. So the best matching split is used to assign the value (second half of the split) to its corresponding key.

For finding the Policy Number, an assumption was made that the actual policy number will come after the string “We wish to inform you that the contract under policy number”. So this string is detected in the output using string matching. After finding the best match, policy number value is taken as the next word which occurs after this string.

Limitations

Some limitations of this project are:

1. Currently image pre-processing is not smart so if the user uploads a camera scanned image then this might lead to undesired results. Also currently no option to remove noise or blur are added because document images that actually need it cannot be identified and we cannot do these preprocessing on all types of images as it will ruin a good image.
2. Information extraction from the output string of the OCR is not smart so separate static classes as per the structure of the document are needed to extract information from the documents.
3. Cropping needs to be done as mentioned above otherwise it might lead to wrong results.
4. Tesseract OCR is slower as we are aiming for best accuracy.
5. Tesseract OCR gives inaccurate results in the case when the text is enclosed inside a solid box (like the box present in Bajaj Policy).

Improvements

Some improvements that can done are:

1. Smart pre-processing catering to the image type.
2. Training customized entity extraction models for our use case, which will also eliminate the need for the rules while cropping the image and cropping will become very flexible.
3. Some settings for pre-processing can be provided so that users can customize the image according to the need.

The Entity Extraction model currently present open source was giving very inaccurate results in our use case and also it only supported extraction of email and address which were not very useful in this project.