# **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.

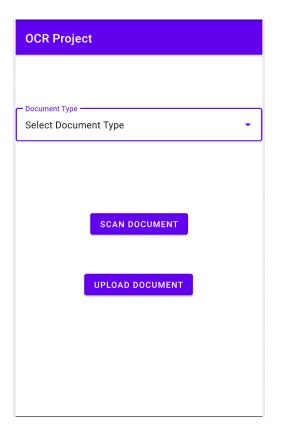
The OCR model used in the project should give output in a specific fixed format so that information extraction from it is reliable and accurate. OCR models like ML kit was able to detect different text boxes in the policy document but the output was some string with shuffled words inside it making it 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.

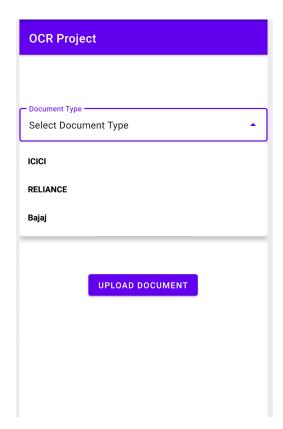
This is an open ended project as documents greatly vary in their structure and information present in it. So each document type is handled separately according to its

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:

Nibl	haye Va	ade	mbaro	1						UIN: IRDAN115R	reduct Code: 3001 P0017V0120010
olicy Cer	tificate										
rivate Car P	ackage Polic	cy									
our Policy D	Details										
Name					_	Telephone no		Mobile no	En	ral	
SUBRAT						-		7869041158		UBRAT098@GMAIL.C	юм
iddress	ADNATAKA DA	NGALOS	E. KARNATAKA - :	60103		Policy No 3001/52111755/	naiona			Policy No IP3001202203290638	15195255
ENGALONO, R	ARONA I ANDA, DA	MUNLON	E, IONGRA IARA	NO 103		Policy Issued O			Co	vernote No	30190200
						Mar 29, 2022 Vehicle Registra	nion No			111755 Nicle Registration Dat	
						MP04CG3456	KION NO			nicie rregistration Dat er 29, 2016	
enure						Period of Insurar					
Year Iominee Name			Relationship	Relationship Age		RTO Location		dnight of Mar 29		hecated To	
					0	MADHYA PRAD					
Named Passeng	ers Nominee		Relationship		Age	GSTIN Number (	Custome	1)		e Number 232296	
Servicing Branch	Name		Servicing Branc	h Address							
Mumbai			414, ICICI LOMI MUMBAJ, 40002	BARD HOUS 5. MAHARA	E, VE	ER SAVARKAR N IA	targ, n	EAR SIDDHI VIN	AYAK TER	APLE MAIN GATE, PR	ABHADEVI,
			,								
revious Pol											
Previous Policy	No		Previous Policy P. 30-63-2021 to 29-6			Previous Ir		ame		evious Policy Type Imprehensive Package	
	CB		3043-2021 10 294	312022		Claims Ma	de Unde	r Previous Policy		impremensive rackage	
50						0					
ehicle and I	DV Details										
		Model	1.0 LS SPORTS	Type of B Saloon	ody	CC/KW	Mfg Yr	Seating 5	Capacity	Chassis No.	Engine No.
								Accessories ()		CNG / LPG Unit ()	Total IDV (')
MP04CG3456	CHEVROLET Trailer (1)		ctrical Accessories								141828
/P34CG3456 /ehicle IDV ()			ctrical Accessories			0				0	141020
AP34CG3456 /ehicle IDV ()  41828	Trailer (*)	Non Ele	ctrical Accessories							0	141020
AP34CG3456 Pehicle IDV () 41828 remium Bre	Trailer () 0	Non Ele	ctrical Accessories			0				0	
APS4CG3456 Vehicle IDV () 41828 remium Bre	Trailor () 0 eak-up emium (A)	Non Ele	ectrical Accessories			0 () 1 4048 I	Liability I	Premium (B)		0	2072
APS4CG3456 Vehicle IDV () 41828 remium Bre	Trailor () 0 eak-up emium (A)	Non Ele	ectrical Accessories			0 () 1 4048 I	Liability I	Premium (B)		9	2072
VP04CG3456 Vehicle IDV () 141828  remium Bre Dwn Damage Pre Basic OD Premis	Trailer (') 0 eak-up emium (A) un	Non Ele				0 () 1 4048 I	Liability I	Premium (B)			2072
WP34CG3456 Vehicle IDV () 441628  remium Bre Dwn Damage Pro Basic OD Premii Sub Total Savings - You hu	Trailer (') 0 eak-up emium (A) um	Non Ele	ount on your premi			0 (7 I 4048 I 4946	Liability I	Premium (B)			2072
WP34CG3455 Vehicle IDV () 141628  remium Bre Dwn Damage Pre Basic OD Premi Sub Total Savengs • You Ni vo Claim Bonus	Trailer (') 0 eak-up emium (A) um	Non Ele				0 () 1 4048 1	Liability I	Premium (B)			2072
WP34CG3455 Vehicle IDV () 141628  remium Bre Dwn Damage Pre Basic OD Premi Sub Total Savengs • You Ni vo Claim Bonus	Trailer (') 0 eak-up emium (A) um	Non Ele				0 4048 4946 2023 2023	Liability I Basic Th Fotal	Premium (B) ird Party Liability bility Premium (B	0		2072 2072 2072
WP34CG3455 Vehicle IDV () 141628  remium Bre Dwn Damage Pre Basic OD Premi Sub Total Savengs • You Ni vo Claim Bonus	Trailer (') 0 eak-up emium (A) um	Non Ele				0 (1) 4048 4348 2023 2023	Liability I Basic Th Fotal	Premium (B) ird Party Liability	0	•	2072 2072 2072 4095
MP94CG3455 //ehicle IDV (*) //41626 remium Bre Dwn Darnage Pri assic OD Premis Sub Total Saviengs - You Ni vo Citalin Bonus Sub-Total	Trailer ()  ak-up  emium (A)  un  ave saved the foll	Non Ele				() () (4048   4348   2023   2023   1	Liability I Basic Th Fotal Cotal Lia Fotal Pac GST	Premium (B) ird Parly Liability bility Premium (d kage Premium (	0		2072 2072 2072 2072 4095 737.10
PPACG3455  (*dhide IDV (*)  41828  remium Bre  Dwn Damage Pri Islatic CD Premis  Islatic CD Premis  Islatic CD Income  Income Total  Islatic Communication  Income Total  Islatic Communication  Income Total  Islatic Communication  Income Total  Islatic Communication  Income Total  I	Trailer ()  ak-up  emium (A)  un  ave saved the foll	Non Ele				() () (4048   4348   2023   2023   1	Liability I Basic Th Fotal Cotal Lia Fotal Pac GST	Premium (B) ird Party Liability bility Premium (B	0		2072 2072 2072 2072 4095 737.10
PPACG3455  (*dhide IDV (*)  41828  remium Bre  Dwn Damage Pri Islatic CD Premis  Islatic CD Premis  Islatic CD Income  Income Total  Islatic Communication  Income Total  Islatic Communication  Income Total  Islatic Communication  Income Total  Islatic Communication  Income Total  I	Trailer ()  ak-up  emium (A)  un  ave saved the foll	Non Ele				() () (4048   4348   2023   2023   1	Liability I Basic Th Fotal Cotal Lia Fotal Pac GST	Premium (B) ird Parly Liability bility Premium (d kage Premium (	0		2072 2072 2072 2072 4095
MP94CG3455 //ehicle IDV (*) //41626 remium Bre Dwn Darnage Pri assic OD Premis Sub Total Saviengs - You Ni vo Citalin Bonus Sub-Total	Trailer ()  ak-up  emium (A)  un  ave saved the foll	Non Ele				() () (4048   4348   2023   2023   1	Liability I Basic Th Fotal Cotal Lia Fotal Pac GST	Premium (B) ird Parly Liability bility Premium (d kage Premium (	0		2072 2072 2072 2072 4095 737.10
WPs4C02459  Vehicle (DV ()  (41820  remium Bre  Own Damaga Pri  Basic OD Premiu  Bub Total  Basic OD Premi  Bub Total  Basicaps - You ha  Vo Claira Bonus  Bub-Total	Trailer () 0 ask-up emium (A) um are saved the foli 50% ge Premium (A)	Non Els				0 4046 4046 2023 2023 2023	Ciability I Basic Th Fotal Fotal Pac GST Fotal Pre	Premium (B) and Party Liability billity Premium (B klage Premium (	0		2072 2072 2072 2072 4095 737.10
WP94C0459  Wholes (DV ()  141028  remium Bre  Den Dennigs Pr  Salic Total  Salic To	Trailer ()  Ben-up  Be	Non Els				() () (4048   4348   2023   2023   1	Liability I Basic Th Fotal Lia Fotal Pac GST Fotal Pro	Premium (R) and Party Liability bildly Premium (R kage Premium (R mium Payable	0		2072 2072 2072 2072 4095 737.10
No Claim Bonus Sub-Total Total Own Danna	Trader ()  8X-UP  menum (A)  m  ge Premium (A)  ge Premium (A)  ca. No Extension  butble: 1600	Non Els	sound on your premi			0 () () () () () () () () () () () () () (	Fotal Lia Fotal Lia Fotal Pac GST Fotal Pre	Premium (R) and Party Liability bildly Premium (R kage Premium (R mium Payable	7) A+B)	ŀ	2072 2072 2072 2072 4095 737.10

	aye Va		nbar	-							UIN: IRDAN115R	reduct Code: 300 P0017V0120010	
Policy Certi													
Private Car Pa		У											
Notro Bubbat Addonio Bendaluru, Karnataka, Bandalore, Karnataka - 960183						Telephone no Mobile no 7865041158 Policy No 3001152111755/08/009 Policy Issued On Mar 29, 2022 Vetricie Registration No MPpac(CLIA58					Email SSUBRAT996@GMAIL.COM  E-Policy No  PHP300120220329053935195255  Covernole No  32111755  Vehicle Registration Date  Mar 29, 2016		
Tenure 1 Year						Period of Insur Mar 30, 2022 0		Midwinds.	- May 20 202				
Nominee Name			Relationship		Age						ecated To		
					Age	GSTIN Number							
					0						232296		
Servicing Branch N Mumbai	ane		Servicing Bran 414, ICICI LON MUMBAL 4000	BARD HO	USE. VE	ER SAVARKAR	MARG	NEAR SI	DHI VINAYA	KTE	PLE MAIN GATE, PR	ABHADEVI.	
Previous Year NGE 50		33	63-2021 to 29-	03-2022		Ctairns h	lade Uni	der Previo	us Policy	Co	mprehensive Packag		
Vehicle and ID	V Detalls												
MP04CG3456 Vehicle IDV (1)	Trailor (*)		LS SPORTS	Type of Saloon is ()		CC/KW 995 Electrical / E 0	Mfg 1 2016 Jectronia		Seating Cap 5 ries (*)	acity	Chassis No. 99999 CNG / LPG Unit () 0	Engine No. 83888 Total IDV (*) 141828	
Own Damage Premium Basic OD Premium						() Liability Premium (B) 4048 Basic Third Party Liability Total				207: 207:			
Sub Total						4046							
Savings - You have No Claim Bonus 50	saved the follo	wing amou	n on your pren	auth		2023 2023			omium (B) remium (A+B)			207 409 737.1	
Sub-Total												483	
Sub-Total  Total Own Damage	Premium (A)					2023							
Sub-Total  Total Own Damage  Geographical Area:	No Extension						le IMT C	Nauses: 2				***************************************	
Sub-Total  Total Own Damage	No Extension	101630					le IMT C	Nauses: 2					

	haye Va		mbarc	1							UIN: IRDAN115R	reduct Code: 30 P0017V01200	
Policy Cer	tificate												
Private Car P	ackage Poli	зу											
four Policy I	Details												
Norre SUBRRAT Address BENGALURU, KARNATAKA, BANGALORE, KARNATAKA - 560103						Policy Issued Mar 29, 2022	7889041158 Polity No 1001/52111755/00/009 Polity Issued On face 29, 2022 Polity Issued No				Email SSUBRATONE@GMAE.COM E-Policy No PHP390120229329063935196255 Covernote No S21117765 Vehicle Registration Date Mar 23, 2016		
Tenure						Period of Insur							
1 Year Nominee Name			Relationship		Age	Mar 30, 2022 0	10:00 to 5	Eidnight	of Mar 29, 20		thecated To		
- Named Passeno				0		MADHYA PRA							
samed Passeng	ers Nominee		Relationship .		Age	GSTIN Number	r (Custom	ier)			ce Number 2232296		
Bervicing Branct Numbai	iervicing Branch Name Servicing Branch Address				USE, VE	ER SAVARKAR	MARG,	NEAR S	IDEHI VINAY			ABHADEVI,	
Previous Policy NMMKKONNKJI Previous Year N 59	No KOKK		Previous Policy Pr 30-03-2021 to 29-0			Previous 0 Claims N			ous Policy		revious Policy Type omprehensive Packag		
Registration No. MP54CG3456 Vehicle IDV (*)	Make CHEVROLET		1.0 LS SPORTS	Type of Saloon	Body	CCIKW 995 Electrical / E	Mig Y 2016		Seating Ca	pacity	Chassis No. 99999 CNG / LPG Unit (*)	Engine No. 88888 Total IDV (1)	
141828 remium Bre	0	0				0					0	141828	
Dwn Damage Pr							Liability	. Desmin	m (B)				
Basic OD Premi									ty Liability			20	
Sub Total						4046							
No Claim Bonus	ave saved the fol 50%	lowing am	ount on your premi	ium		2023 2023			renium (8) Premium (AH	1)		20	
Sub-Total							IGS1				-		
	ge Premium (A)					2023	Total P	remium	Payable			737. % 48	
Total Own Dama	na: No Extension					Applicab	Total P	lauses:			-	%	
Sub-Total  Total Own Dama  Geographical An  Compulsory Dec	ea: No Extension	1405830		mium Arno			Total P	lauses:			Receist Date	48	

	laye va	ade											P0017V01200102	
olicy Cer	tificate													
rivate Car P	ackage Poli	су												
our Policy [	Details													
Name SUBRAT						Telep	no enod		Mobile 786904		Em	ail UBRAT096@GMAIL.C	ом.	
Iddress						Polic	y No		700304			folicy No		
BENGALURU, K	ARNATAKA, BA	ANGALORE	, KARNATAKA - 5	50103		3001	52111755				PH	P30012022032905393	35196255	
							y Issued (	0n				vernote No 111755		
							r <b>s, zozz</b> de Registr	ation No				111755 Nicle Registration Date		
						MPO	CG3456					r 29, 2016		
Tenure 1 <b>Year</b>						Perio Mar 3	d of Insure	ince 0:00 to M	lidnight	of Mar 29, 202	3			
Nominee Name			Relationship		Age	RTO Location						ecated To		
<ul> <li>Named Passeng</li> </ul>	er's Nominee		Relationship		<b>O</b> Age	MADHYA PRADESH-BHOPAL GSTIN Number (Customer)					hvoice	Number		
					0							232296		
Servicing Branch Mumbai	h Name		Servicing Branch						E 4 D 01	DESIL MINISTER		PLE MAIN GATE, PR	ADUADON	
	lass Parkalla		MUMBAI, 400026	, MAHARA	ASHTR	EA.								
revious Pol											_			
Previous Policy			Previous Policy Pe 30-03-2021 to 29-03				Previous 0	Insurer N	lame			Previous Policy Type Comprehensive Package		
Previous Year N		-	10-03-2021 (0 25-01	-2022				ade Und	er Previ	ous Policy	100	inpronounce in accept		
50							0							
ehicle and l	DV Detelle													
	DV Details													
	Make	Model		Type of E	Body		ĸw	Mig Y		Seating Cap	acity	Chassis No.	Engine No.	
MP04CG3456	Make CHEVROLET	SPARK	1.0 LS SPORTS	Saloon	Body	995	5	2016		5	acity	99999	88888	
MP04CG3456 Vehicle IDV (*)	Make	SPARK	1.0 LS SPORTS trical Accessories	Saloon	Body	995		2016		5	acity			
MP04CG3456 Vehicle IDV (*) 141828	Make CHEVROLET Trailer (*) 6	SPARK 1		Saloon	Body	995 Ele	5	2016		5	acity	99999 CNG / LPG Unit (*)	\$8888 Total IDV (1)	
MP94CG3456 Vehicle IDV () 141828 Premium Bre Own Damage Pre	Make CHEVROLET Trailer (*) 0 eak-up emium (A)	SPARK 1		Saloon	Body	995 Ele	otrical / E	2016 lectronic	Access	s ories (')	acity	99999 CNG / LPG Unit (*)	38388 Total IDV () 141828	
Registration No. MP04CQ3456 Vehicle IDV () 141828 Premium Bre Own Damage Pe Basic OD Premi	Make CHEVROLET Trailer (*) 0 eak-up emium (A)	SPARK 1		Saloon	Body	995 Ele	otrical / E	2016 lectronic	Access	s ories (')	acity	99999 CNG / LPG Unit (*)	\$8888 Total IDV (1)	
MP04CQ3456 Vehicle IDV (*) 141828 Premium Bre Own Damage Pri Basic OD Premii	Make CHEVROLET Trailer (*) 0 eak-up emium (A)	SPARK Non Elect	trical Accessories	Saloon (*)	Body	995 Ele	otrical / E	2016 lectronic	Access	s ories (')	acity	99999 CNG / LPG Unit (*)	\$8888 Total IDV () 141828	
MP94CG3456 Vehicle IDV (*) 141828 Premium Bre Own Damage Pri Basic OD Premii Sub Total Savings - You hi	Make CHEVROLET Trailer () 0 pak-up emium (A) um	SPARK Non Elect		Saloon (*)	Body	995 Ele	() 4045 4046	2016 lectronic	Access	s ories (')	acity	99999 CNG / LPG Unit (*)	\$8888 Total IDV () 141828	
MP04CG3456 Vehicle IDV () 141628  Termium Bre Dwn Damage Pr Basic OD Premi Sub Total Savings - You h No Claim Bonus	Make CHEVROLET Trailer () 0 pak-up emium (A) um	SPARK Non Elect	trical Accessories	Saloon (*)	Body	995 Ele	() 4046 4046 2023	2016 lectronic	Access	s ories (')	acity	99999 CNG / LPG Unit (*)	\$8888 Total IDV () 141828	
MP04CG3456 Vehicle IDV () 141628  Termium Bre Dwn Damage Pr Basic OD Premi Sub Total Savings - You h No Claim Bonus	Make CHEVROLET Train (1) 0 pak-up emium (A) um	SPARK Non Elect	trical Accessories	Saloon (*)	Body	995 Ele	() 4045 4046	2016 lectronic Liability Basic TI Total	Access Premiu	s ories (')	acity	99999 CNG / LPG Unit (*)	\$8888 Total IDV () 141828	
MP04CG3456 Vehicle IDV () 141628  Termium Bre Dwn Damage Pr Basic OD Premi Sub Total Savings - You h No Claim Bonus	Make CHEVROLET Train (1) 0 pak-up emium (A) um	SPARK Non Elect	trical Accessories	Saloon (*)	Body	995 Ele	() 4046 4046 2023	2016 lectronic Liability Basic TI Total Lia Total Pa	Access Premiu	s ories (') m (B) ly Liability	acity	99999 CNG / LPG Unit (*)	8888 Total IDV () 141828 ( 2072 2072 2072 4055	
MP04CG3456 Vehicle IDV () 141628  Termium Bre Dwn Damage Pr Basic OD Premi Sub Total Savings - You h No Claim Bonus	Make CHEVROLET Train (1) 0 pak-up emium (A) um	SPARK Non Elect	trical Accessories	Saloon (*)	Body	995 Ele	() 4046 4046 2023	2016 lectronic Liability Basic TI Total	Access Premiu	s ories ()  m (B) ly Lizbilly	acity	99998 CNG / LPG Unit () 9	\$5888 Total IDV () 141828 ( ( 2072 2072 2072 2072 4695	
MP04CQ3456 /ehicle IDV () /41526 remium Bre Own Damage Ph Basic OD Premi Sub Total Savings - You In vio Claim Bonus Sub-Total	Make CHEVROLET Trailer (*) 0 Dak-up emium (A) um ave saved the foil 50%	SPARK Non Elect	trical Accessories	Saloon (*)	Body	995 Ele	(1) 4045 4046 2023 2023	2016 lectronic Liability Basic TI Total Total Lia Total Pa	Premiu hird Pan shility Pr ckage F	m (B) by Liability remium (A+B)	acity	99999 CNG / LPG Unit (*)	\$5888 Total IDV () 141828 ( 2072 2072 2072 2072 - 737.10 5 18	
MP04CQ3456 Vehicle IDV (*) 141828 Premium Bre Own Damage Pri Basic OD Premii	Make CHEVROLET Trailer (*) 0 Dak-up emium (A) um ave saved the foil 50%	SPARK Non Elect	trical Accessories	Saloon (*)	Body	995 Ele	(1) 4045 4046 2023 2023	2016 lectronic Liability Basic TI Total Lia Total Pa	Premiu hird Pan shility Pr ckage F	m (B) by Liability remium (A+B)	acity	99998 CNG / LPG Unit () 9	\$5888 Total IDV () 141828 ( 2072 2072 2072 2072 - 737.10 5 18	
MP04CG3455 Vehicle IDV () 141828 remium Bre Own Damaga Pn Basic OD Premi Sub Total Savings - You In No Claim Bonus Sub-Total	Make CHEVROLET Trailer (*) 0 Dak-up emium (A) um ave saved the foil 50%	SPARK Non Elect	trical Accessories	Saloon (*)	Body	995 Ele	(1) 4045 4046 2023 2023	2016 lectronic Liability Basic TI Total Total Lia Total Pa	Premiu hird Pan shility Pr ckage F	m (B) by Liability remium (A+B)	acity	99998 CNG / LPG Unit () 9	\$5888 Total IDV () 141828 ( 2072 2072 2072 2072 - 737.10 5 18	
MP94C03459  Weshold EIDV () 141628  Fremium Bre Own Damage Pe Basic Dor Premi Basic Do Premi Basic Do Premi Basic Do Dereni	Make CHEVROLET Trailer (*)  Bak-up emisser (A) under the following save saved the following Premiser (A) under the following Premise	SPARK Non Elect	trical Accessories	Saloon (*)	Body	995 Ele	() 4046 4046 2023 2023 2023	2016 lectronic Liability Basic Ti Total Lia Total Pa IGST Total Pe	Premium  Pre	s (B) by Liebilly viewince (B) reminer (B) reminer (B) Payable	acity	99998 CNG / LPG Unit () 9	\$5888 Total IDV () 141828 ( ( 2072 2072 2072 2072 4695	
MP94C03459  Wholice IDV ()  141628  Premium Bre Own Damage Pn Basic OD Premi Sub Total Savings - You h No Claim Bonus Sub-Total  Total Own Dama	Make CHEVROLET Trailer (*)  Bak-up eenikum (A)  um  ave saved the foll 500%  ge Premium (A)  un  No Extersion  butüble: 1000	SPARK Non Elect	trical Accessories	Saloon (*)		995 Ele	(1) 4048 4048 2023 2023 2023	2016 lectronic Liability Basic Ti Total Lia Total Pa IGST Total Pe	Premium  Pre	s (B) by Liebilly viewince (B) reminer (B) reminer (B) Payable	acity	99998 CNG / LPG Unit () 9	98458 ( 1014 IDV ( ) 141828 ( 2072 2672 2672 2672 2672 4655 5 165 4852	

# 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**

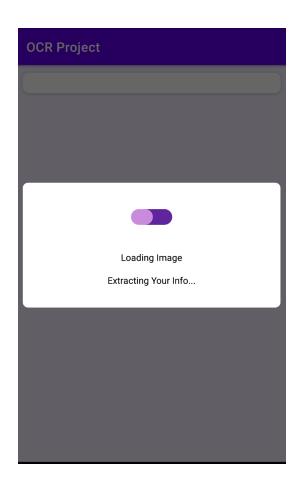
As mentioned before a specific output format is needed to locate the data in the output text. The OCR model used in this project is Tesseract OCR which is one of the most popular models for OCR.

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.



# 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:

Policy Certif									OIN. IND	MNIIDRE	20017V0120010
	ficate										
	ckage Polic	;y									
our Policy De	tails										
Name					Telephone no		obile no		nail		
SUBKAI						71	659041158		ORKY10886	MAIL.C	MC
Address					Policy No			E-I	Policy No IP3001202203		
SENGALURU, KAR	NATAKA, BA	NGALORE, KARNA	TAKA - 560103		3001/5211175 Policy Issued				vernote No	2906393	5196255
					Mar 29, 2022				111755		
					Vehicle Regist	ration No			hicle Registra	tion Date	
Tenure		MP04CG3456 Period of Insur			Ma	ır 29, 2016					
1 Year							night of Mar 29,	2023			
Nominee Name		Relation	nship	Age	RTO Location		-		hecated To		
Named Passenger's	e Meminee	Relation	achie	Age	MADHYA PRA GSTIN Number		PAL	leur'-	e Number		
vamed Passenger's	s ryominee	- Relation	ianip	Age 0	GOTIN Number	(Customer)			e Number 232296		
Servicing Branch N Mumbai	lame	414, ICI	ig Branch Address CI LOMBARD HO N, 400025, MAHA	use, ve		MARG, NE.	AR SIDDHI VINA			ATE, PRA	BHADEVI,
revious Policy	y Details	,	.,								,
Previous Policy No			Policy Period			Insurer Nar	10		evious Policy		
MMKKONNKJKOP Previous Year NCB		30-03-202	1 to 29-03-2022		O Claima N	Indo Hador	Previous Policy	Co	mprehensive	Package	
	Make CHEVROLET	Model	Type o		CC/KW	Mfg Yr	Seating (	Capacity	Chassis No.		Engine No.
Vehicle IDV (')	Trailer (')	Non Electrical Acc		-			cessories (')		CNG / LPG U	Jnit (')	Total IDV (')
remium Break		,									
Own Damage Premi Basic OD Premium						Basic Thin Total	emium (B) I Party Liability				2072 <b>207</b> 2
Sub Total					4046						
Savings - You have No Claim Bonus 50 Sub-Total		owing amount on yo	ur premium		2023 2023						
Jub-10tai					2023	Total Liabi	lity Premium (B)				2072
							age Premium (A				4095
						IGST					737.10
Total Own Damage	Promium (A)				2022	Total Press	ium Payable			%	483
. o.u. Own Daniage	(A)				2023	. Jun F7011	rayawe				4034
Geographical Area:						le IMT Clau					
Compulsory Deduct	tible: ` 1000				Voluntar	Deductible	. 0				
Premium Collection GSTIN Reg.No		405830 ACI7904G1ZN	Premium Amo		4832 997134/0	ENERAL IN	ISURANCE SEF	RVICES	Receipt	Date 2	9-03-2022

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:

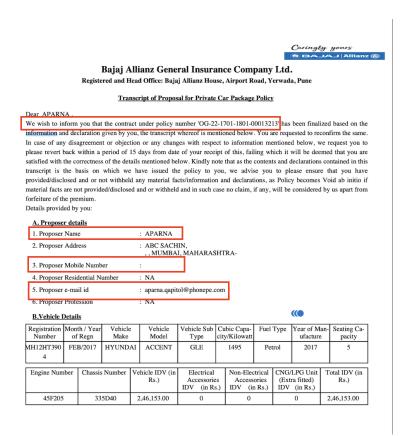
Policy Number: 9202222231	120002807		Proposal/Covernote No: R22042200949						
Insured Name :FAROOQ A	HMAD	ance General Insu	Period of Insurance: From 00:00 Hrs on 25-Apr-2022 to Midnight of 24-Apr-2023						
Communication Address & SOUTH,,BANGALORE,,KAR	RPIACE of Supply :,BANGAL RNATAKA,,INDIA,560061	ORE	Policy Issuing Branch: 570- RECTIFIER HOUSE, NAIGAUM CROSS ROAD WADALA(W), MUMBAI,,MAHARASHTRA, 400031						
Mobile No :9611297856			Tax Invoice N	o. & Date :R22042200949 & 22 Apr 2022	16:42				
Email-ID : murulikrishna94@	gmail.com	70	GSTIN/UIN &	Place of Supply:					
Insured's Blood group:									
Insured Vehicle Details	Company Limited Relia	ance General Insu	rance Compa	iny Limited Reliance General Ins	urance Company L				
Registration No.	KA010P8789	TWO WHEEler L	ability Only P	Mfg. Month & Year	APR-2018				
Make / Model & Variant	Bajaj / Discover / M 100-e	s		CC / HP / Watt	100				
Engine No./Chassis No.	FD878 / FED76			Seating Capacity Including	2				
Type of Body	NA			Total Premium ₹	1258.00				
RTO Location	KARNATAKA - Bangalo	re Central (Koraman	gala)	IDV ₹	20583.00				
Hypothecation/Lease	NA NA Instead Relia								
Insured Declared Value (ID)	Only Policy	Two Wheeler Li	ability Only P	olicy Two Wheeler	Liability Only Polic				
Vehicle IDV ₹		20583	Non Electrical	Accessories ₹	0.0				
Electrical / Electronic Acces	sories ₹	0.0	0 Total IDV ₹		20583.00				
Premium Summary									
Own Damage - Section I		Amount (₹)	Liability - Sec	Amount (₹)					
Basic OD Total Basic Own Damage TOTAL OWN DAMAGE PI		351.56 351.56 351.56	PA Benefits - TOTAL LIABI	iability Premium Section III LITY PREMIUM (AGE PREMIUM (Sec I + II + III)	714.00 714.00 714.00 1066.00 192.00				
TOTAL PREMIUM PAYABI	IF (F)				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:



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.