SuchanaSaathi: A Multimodal AI Framework for Accessible Understanding of Government Schemes Using OCR, STT, TTS, and Translation

Dr. Amol Joglekar[1], Falak Khan[2]
[1,2]*SVKM's Mithibai College of Arts, Chauhan Institute of Science & Amrutben Jivanlal College of Commerce and Economics (AUTONOMOUS), Vile Parle (W), Mumbai 400056, India*

***Abstract*—India is a country with a literacy rate of approximately 80-81%. That leaves the illiteracy rate of around 19-20%. This section of the society deals with various challenges when it comes to reading and writing. The government of India has published various schemes and programs in order to help the underprivileged people. But most of the underprivileged people lack basic education which is why people are not able to access these programs and schemes from the existing portals because language barriers. People in these areas struggle with language comprehension, digital literacy and complex documentation when they try to use the government portal for accessing these schemes. This paper is about a system called SuchanaSaathi. SuchanaSaathi is a program that helps people find out about various government schemes. It uses different tools to do this. SuchanaSaathi uses Optical Character Recognition (OCR) to read text from papers and pamphlets. It also uses Neural Machine Translation so it can understand and talk to people in regional languages. SuchanaSaathi has a feature that reads text out to people, which is called Text-to-Speech. This is really helpful for people who cannot read or who prefer to listen. SuchanaSaathi also has a feature called Speech-to-Text that lets people talk to it and ask for help. The system looks at the person's information. Then gives them advice on which government schemes they might be eligible for. SuchanaSaathi does this with a part of the system called the eligibility detection module. This module makes sure that the advice given to people is just right, for them and their situation. Designed to operate efficiently on low-resource devices, the framework enhances accessibility for citizens regardless of literacy or linguistic background. By harmonizing vision, speech, and language technologies, SuchanaSaathi promotes inclusive and citizen-centric governance through equitable digital participation.**

***Index Terms*—Multimodal AI, Optical Character Recognition (OCR), Speech-to-Text (STT), Text-to-Speech (TTS), Neural Machine Translation (NMT), Eligibility Detection, Digital Inclusion, e-Governance, Public Service Delivery**

I. INTRODUCTION

The people of India are having a lot of trouble getting the help they need from the government of India. This is a problem for the Indians. The government of India is not able to help these people because there are obstacles in the way. These people are facing technological gaps. One of the problems is that they speak a language that they cannot read or write. Hence, they need help from the government but are not able to get it because of these gaps.

The government of India has started programs to help people with things like learning, health, jobs, homes and other social services. The thing is, a lot of people in India do not know about these government programs and so they are unable to use these programs. The Indian government welfare programs and Indian government public service programs are still not reaching people. This is because people who work in the government and other offices use a lot of paper documents and the language on the India government websites is very complicated. This is a problem because it assumes that all citizens can read and write well and are good at using computers. The truth is that many people, those living in rural areas and small towns do not have these skills. Many citizens living in rural areas and small towns have a hard time with this. The government and other offices use printed documents and formal language on portals, which's a problem for many citizens. Citizens in rural and semi-

urban regions have trouble with this because they do not have the necessary literacy and digital skills to deal with printed documents and formal language, on portals.

In some areas people speak languages but cannot read it due to lack of education. So, citizens have a hard time understanding government forms and pamphlets. Government forms are usually written in English or standard Hindi. This makes it difficult for citizens to understand government forms and the rules to be eligible for any scheme. The government is using computers now. This is making it even more tough for people who cannot use computers to get the help they need from government forms and pamphlets. Citizens who are not good at using computers are having a lot of trouble with government forms and pamphlets.
The government has programs to help people. Some people have a hard time understanding these programs and using them. This paper is about an AI assistant that helps every person find out about the government programs that are right for them.

The government programs are what this AI assistant is about. It is special because it can communicate in ways like understanding what people say and talking to people. This makes the government programs easily accessible for everyone and the AI assistant is a tool that people can use to learn about these government programs. The assistant can even read what is written on papers like ID cards and ration cards and pamphlets. It does this with a tool called Optical Character Recognition. The assistant is made to help every citizen. It does not matter how well people can read or speak the language. The assistant is for everyone. We get information. Then we change it into the regional language of that area. We use Neural Machine Translation to do this. Neural Machine Translation makes it easy for people to understand the information. People can understand the information because it is in their language. Text-to-Speech does a job of changing the translated information into audio. The audio sounds like a person talking. This is very helpful for people who cannot read.

They can just ask the system for what they want by using keywords like "women's scheme". The system can do this because of Speech-to-Text. The system is really easy to talk to. Text-to-Speech and Speech-to-Text make the system feel like a person. It feels like you are having a conversation with Text-to-Speech and Speech-to-Text. Neural Machine Translation and Speech-to-Text and Text-to-Speech all work together to make this happen. Neural Machine Translation plays a role in this process. Speech-to-Text is also important because it helps to convert what people say into text. Then Text-to-Speech does the opposite: it turns the text into speech that people can hear. So Neural Machine Translation and Speech-to-Text and Text-to-Speech are all. They work together to make things easier for people.

The framework has an important part, the eligibility detection module. This eligibility detection module takes a look at the information of the user. It finds the government schemes for the user. The eligibility detection module does this by looking at a list of government schemes and then picking the government schemes that are most useful to the user.

II. LITERATURE REVIEW

Janssen added to this in 2012[3][14] by showing that loading information onto a government portal does not solve the problem if nobody can make sense of what they're reading. Availability is not the same as accessibility. That distinction matters enormously. When researchers started looking at people who struggle with reading the findings were consistent. Dense text on a screen made things worse for these people. Breaking instructions into clear steps made things better for them. Adding images and audio helped more. Medhi and colleagues spent years across studies in 2007, 2010 and 2011[5][10].

Coming to the same conclusion: what people needed was not more features but more clarity.
The government scheme is what people need to understand and human centered AI guidelines have since echoed this making the case that when technology starts making decisions that affect lives. Especially the lives of people who have been historically excluded. It has a responsibility to be transparent, simple and genuinely easy to engage with. Trust is built when people feel in control of the government scheme not when they feel overwhelmed by it. The government scheme and voice interactions go hand in hand for a lot of people. When reading is

hard and typing is harder, being able to say what you need out loud is not a small convenience. It is the difference between using the system and giving up on it entirely. Kusal and colleagues [20] noticed something in 2022: once people started using voice-based interactions they kept using them. It felt natural like talking to someone who was actually listening rather than filling out a form that might go nowhere.

Language is where things get more unequal for people who want to apply for the government scheme. Systems that work well in English or standard Hindi often fall apart for someone speaking a dialect or a less-resourced language. Simply because there was not enough data to train the model properly. Reimers and Gurevych [1] showed in 2019 that tools like Sentence-BERT, which search by meaning rather than exact words are far more forgiving and far more useful. Especially

when someone phrases a question informally or does not quite know the right terminology for the government scheme. When you add OCR to that mix suddenly even information that is stuck on a piece of paper becomes something a system can read, translate and act on. Wirtz and colleagues [9] and Savelli and colleagues [14] raised a question that sits underneath all of this: is AI actually helping everyone apply for the government scheme or is it quietly helping the people who were already doing fine? It is an important question about the government scheme. Rather than being a reason to step back it is a reason to build more carefully. What all of this research points toward is the need for one integrated system. Something that listens, reads, translates and decides.

OBJECTIVES:

We want to create a system that helps people with government schemes from start to finish. This system will guide citizens through the process. From finding government schemes to figuring out if they are eligible. The system will not leave people hanging through.

We will use a kind of search that understands what people mean. This way people do not need to know the names of government schemes or technical terms to find what they are looking for. We will use tools like Sentence-BERT to make this happen.

We will make sure that our system works well in languages. We do not want people to see translations that're not complete or interfaces that mix languages. Every person should be able to use the system in the language they understand best.

We will build a feature that checks if people are eligible, for government schemes. This feature will not just list rules. It will actually look at a person's profile and tell them if they qualify. This way people do not have to try to figure it out themselves.

III. METHODOLOGY

3.1 System Overview

SuchanaSaathi is built around an idea: no citizen should be left out of government help just because of the language they speak or the device they use. The system puts together five main parts. Optical Character Recognition, Neural Machine Translation, Text-to-Speech, Speech-to-Text and an Eligibility Detection Module. Each part is designed to remove a problem that people face when trying to get government help. SuchanaSaathi uses these parts to make things easier for people. Figure 2 shows how these parts work together to make the system work. SuchanaSaathi is about making government help available, to everyone and it does this by using these five main parts.

3.2 System Architecture

SuchnaSaathi has a simple yet effective architecture. It is designed to handle various types of input like for example image, audio, text etc. For all these input types there is a module dedicated. For image input we have the OCR module wherein the user uploads image and text can be extracted from it. Similarly, for the audio input we have the Smart Scheme Search module where users can query the system through voice input and in their regional language. An eligibility detection module is also included where users can enter their basic details like age, state, occupation, education, type of ration card, gender etc and get relevant schemes recommended for them.

All these modules can work in less connectivity areas as well enabling people from rural areas to also use this system anytime and anywhere.

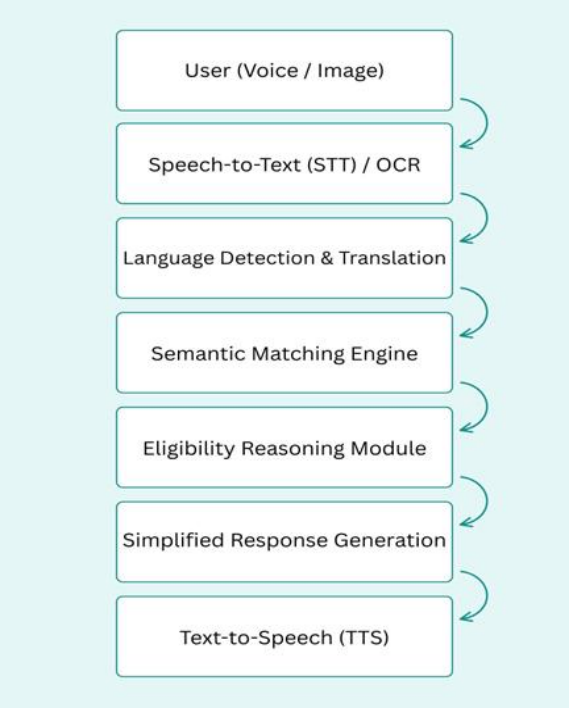

Fig 1: System Architecture

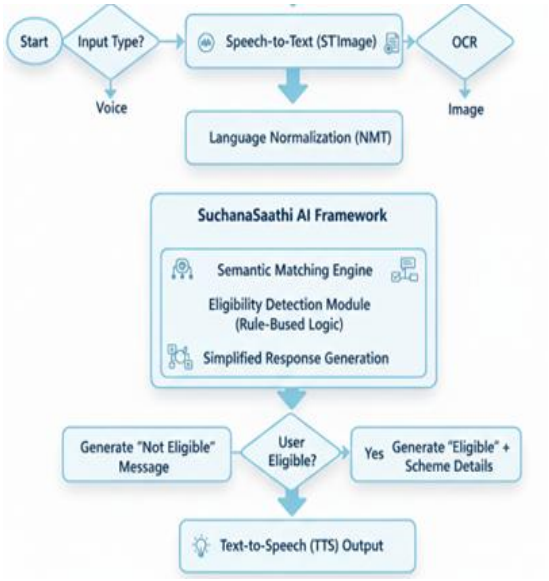

Fig 2 : Flow Chart of the System

3.3 Optical Character Recognition (OCR) Module

People have these papers. They cannot get the information they need from them by themselves. It is not easy to type all the information from these papers into a computer. The OCR module helps with this problem. When someone takes a picture or scans a paper the module looks at the picture. Get the words from it. The system uses a tool that is good at reading many Indian languages like Devanagari, Tamil, Telugu, Bengali and others that are used on government papers. The people who made this system made sure it can handle pictures that're not very clear and have bad lighting, which is common when people take pictures with simple phones in rural areas. Once the words are found they are cleaned up. Made nice before they are sent to the next step. This is very important because the words that come from the OCR module are often not perfect and have mistaken and if we send these mistakes to the translation module it will make everything that comes after it not very good. The OCR module is a useful tool for government documents.

3.4 Neural Machine Translation (NMT) Module

India is home to hundreds of languages and dialects. If a welfare system only speaks Hindi or English it is not really a system for people. The thing that makes SuchanaSaathi truly helpful to people who speak languages is the NMT module. When the OCR module takes out the text or when a person types a question the NMT module figures out what language it is and translates it into the language the person likes to use. This system can work with 22 languages that are officially recognized in India. SuchanaSaathi and the NMT module make this possible for languages, including the 22 languages that are part of the Indian system.

3.5 Text-to-Speech (TTS) Module

Not all people necessarily have reading skills. It basically leaves out the people that are unable to read a particular language. The TTS module converts translated text into audio.

The team paid attention to how the audio sounds when spoken. This is because in Indian languages small changes in tone can completely change what a word means. The module has regional language audios. For content like a full description of a government program the audio is sent in small parts. This way people do not have to wait for the whole thing to be ready before it starts playing.

3.6 Speech-to-Text (STT) Module

For people who do not like to type. There are a lot of them. The Speech to Text module lets them use their voice to put in information. A person can just say what they want to ask into the system and the module turns

what they say into text that the rest of the system can work with. The Speech To Text model was taught and improved using speech from people who talk differently from parts of the country so it can handle different accents, how fast people talk, background noise and the way people say things in their area. This was a deal: if the model was only taught using clean and clear speech from city people it would not work well with audio from a noisy countryside area or if someone spoke in a local dialect. The people who made the model made sure it could handle all the kinds of sounds that happen in real life from the very start.
The module also works with simple questions that use just a few words. A person does not need to say a sentence.

3.7 Eligibility Detection Module
You know about a scheme that is just the beginning. The big question is, can you actually get it? The Eligibility Detection Module is here to help you figure that out in a way that's just for you the Eligibility Detection Module. The Eligibility Detection Module starts by getting some information about you. How old you are, if you are a man or a woman, how much money you make, where you live, what kind of job you have and some other things like that. You can enter this information yourself or the Eligibility Detection Module can get it from a document you scan or it can ask you some questions. The Eligibility Detection Module only asks for what it needs to know so you do not have to fill out a lot of forms, which can be really annoying. Once the Eligibility Detection Module has all your information it looks at a list of government schemes. Like the ones from the Ministry of Rural Development, the Ministry of Agriculture and the Ministry of Health as well as some schemes that are just for your state. Each scheme on the list says who can get it and the Eligibility Detection Module can read that and match it with your information. Then the Eligibility Detection Module gives you a list of schemes that you can probably get. It explains what each one is and how you can get it. The Eligibility Detection Module tells you this in a way that's easy to understand in your own language and it can even read it out loud to you if you want. The Eligibility Detection Module gets updated all the time so it always has the information, about schemes and it can tell you about new ones that just started and ones that are not available anymore. This way the Eligibility Detection Module can always give you advice on the Eligibility Detection Module.

3.8 Implementation Details
The system was made using Python as the language for the backend. The part that can read text from images is built on Tesseract with some scripts to make it work better with Indian languages. The models that can translate text are based on a kind of architecture called transformer and they were fine-tuned using a library from Hugging Face on texts that are specific to certain domains and have both languages. The parts that can convert text to speech and speech to text use models that're open to everyone and were fine-tuned on recordings of people speaking Indian languages. The logic that checks if someone is eligible is made using a set of rules. The data is stored in a simple format called JSON so it is easy to update without having to retrain the models. The front end of the app is made using a simple framework that works on Android phones, even old ones, like Android 8.0 so it can be used on many different kinds of phones that people have in rural India. All the models are made to be small and work well on phones that only have a CPU and not a lot of memory 2GB of RAM so the system can still work on the kinds of phones that the people it is trying to help actually use.

IV. EVALUATION

Suchnasaathi is evaluated as a whole rather than module-by-module. It is done so that we can check whether the entire pipeline is working well as a whole in the real time and that it's actually helping people find the schemes they are eligible for with ease.

4.1 System Performance
SuchanaSaathi works well as a whole system. The main goal of SuchanaSaathi is to find government schemes for a citizen. It can find the schemes that a user is eligible for. On testing it with real users it was found out that this system can retrieve best results within a very few seconds approximately around 6-7 secs. This wait time is normal for web applications in low connectivity areas.
In the eligibility detection module. Similarly, in the smart scheme search module where the user can use their voice as input, is also working just fine with the retrieval time as 5-6 secs. On the other hand, in the

OCR and fraud detection module , the system takes some time like, more than 7 secs to translate the extracted text in the regional language and another 7-8 secs for converting it to audio.

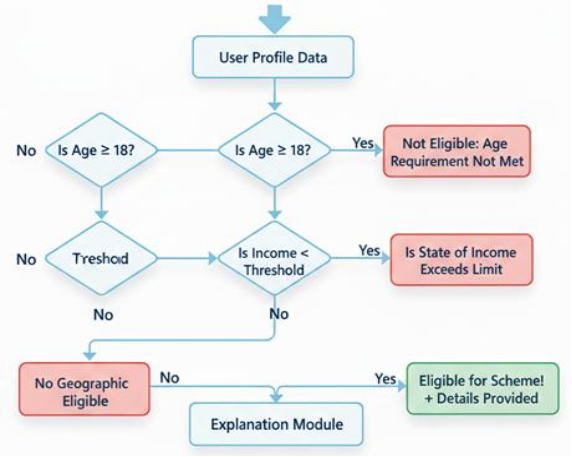

Fig3: Rule-based eligibility determination process

4.2 Usability Study

Technical scores, however good, do not confirm that a system is useful. To get that confirmation, a usability study was conducted with 60 participants across the three field sites rural residents with limited formal education, nearly half of whom were interacting with a smartphone application for the very first time. Participants were asked to complete three tasks: reading a physical document through the system, asking a question by voice, and finding out which schemes they were personally eligible for all without any assistance from the research team.

The results were up to the mark. Task completion rates ranged from 83% to 92% depending on the task, and the average time to generate a useful result was less than three minutes.

4.3 Comparative Context

Compared to an English-only chatbot and a general translation tool paired with a static database. Both used as references. SuchanaSaathi did better in every important area. These areas include accuracy of recommendations languages supported voice interaction capability and ability to work on devices with resources. The increase in F1 score over the reference from 0.74 to 0.89 shows the real advantage of creating a pipeline that is integrated and specific to a domain. This is than linking general tools that were not made to work together. SuchanaSaathi outperformed on every dimension. The improvement reflects the benefit of building an integrated pipeline. It is, then, connecting general-purpose tools that were not designed to work together.

V. RESULTS AND DISCUSSION

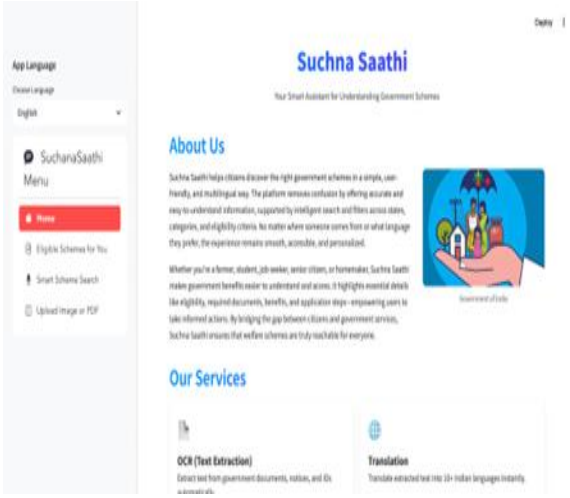

Fig 4.1

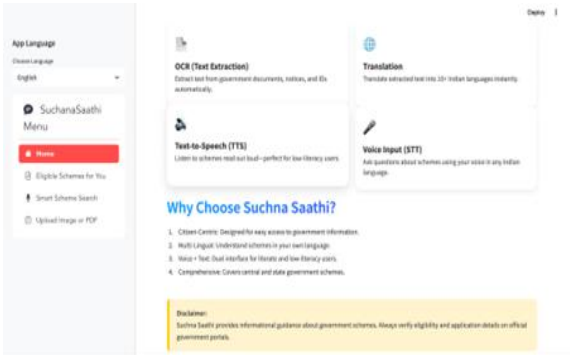

Fig 4.2

Fig 4.1,4.2 Home Page of the Application

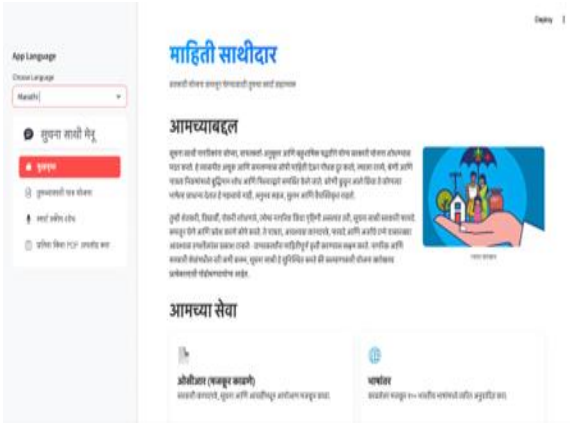

Fig 5 Home Page in a Different Language showcasing the Multilingual nature of the webapp.

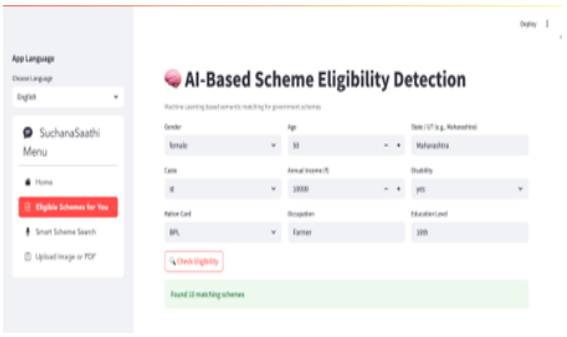

Fig 6.1

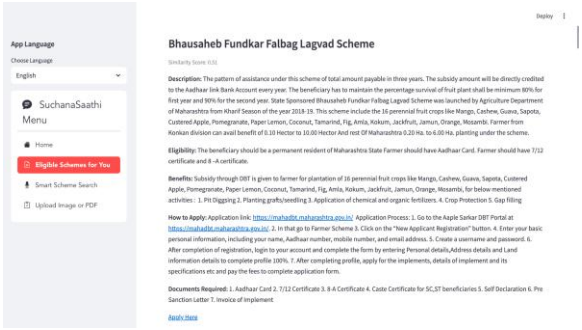

Fig 6.2

Fig 6.1,6.2- Eligibility Detection Module based on basic information of the user.

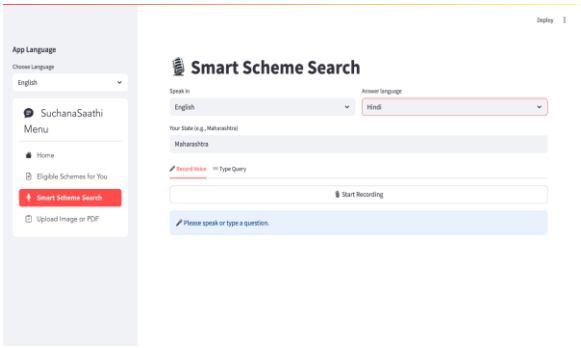

Fig 7.1

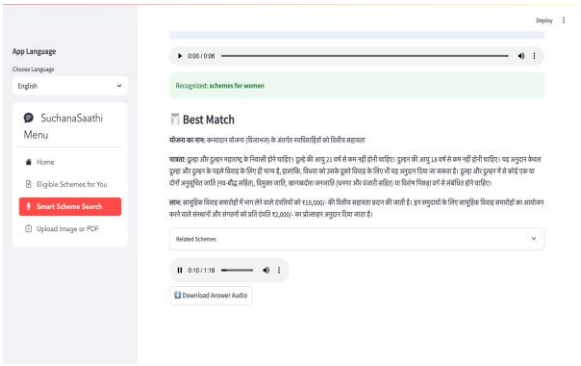

Fig 7.2

Fig 7.3

Figures 7.1,7.2,7.3 show the Voice based Scheme Search with inputs & results in the regional language selected by the user. The results are also converted into audio for users that are unable to read.

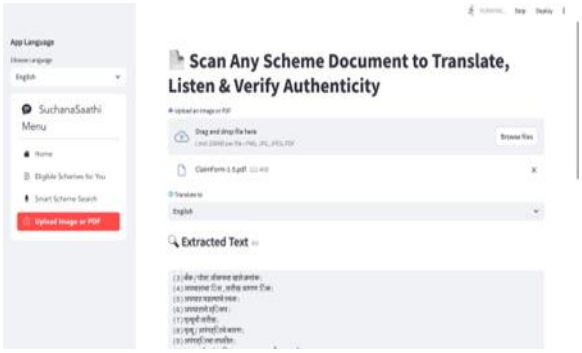

Fig 8.1

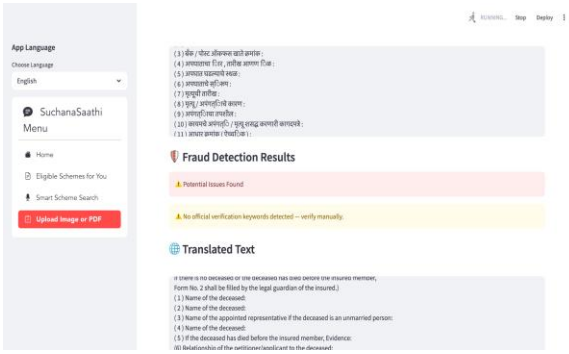

Fig 8.2

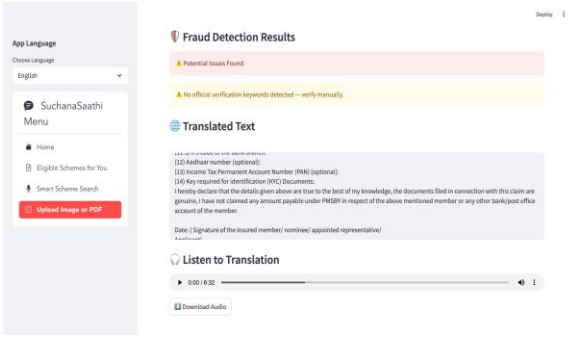

Fig 8.3

Figures 8.1,8.2,8.3 show the OCR based extraction with fraud detection module with inputs & results in the regional language selected by the user. The results are also converted into audio for users that are unable to read.

VI. CONCLUSION

A new way of helping people understand what welfare options are available is really working. This project combines talking to a computer, reading text, different languages, comparing things in a logical way and giving clear reasons why someone can get benefits. It does not use websites instead it works the way people normally ask questions. Things get better when the answers are given in a way through speech. It feels more natural to talk to a computer than to fill out forms and hope you are doing it right. Welfare options are easier to understand when you can just talk to a computer and get answers back, about your welfare options.

When you look at the results you can see that the proposed method really helps people get the information they need. The proposed method makes it faster for people to complete tasks because users can find the programs they need more easily. They also understand who is eligible for these programs better. The proposed method reduces the number of back-and-forth steps that people need to take. People who use the eligibility reasoning tool have confidence because it gives them a clearer understanding of things. This is true even when people ask questions or use everyday language to ask their questions. The proposed method is still able to find the scheme matches because it uses semantic search logic. The proposed method is really good, at understanding what people mean when they ask questions even if they do not ask them in a clear way. When we look at the results it becomes clear that it is really important to evaluate Artificial Intelligence systems that are built to be accessible, by looking at the tasks they can complete, not how accurate their predictions are. The Artificial Intelligence system is successful if we can see that users are able to use the Artificial Intelligence system without any problems. What is most important becomes apparent in how the interfaces of the Artificial Intelligence system work and how easily people can interact with the Artificial Intelligence system. Looking at it overall, this study contributes to growing work on AI designed for people and fairness in public services. Results indicate that combining multiple interaction methods helps reduce barriers for underserved communities in going online and accessing info.

REFERENCES

[1] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in Proc. 2019 Conf. Empirical Methods in Natural Language Processing (EMNLP), 2019. [Online]. Available: https://arxiv.org/abs/1908.10084

[2] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge, U.K.: Cambridge Univ. Press, 2008. [Online]. Available: https://nlp.stanford.edu/IR-book/html/htmledition/irbook.html

[3] M. Janssen, Y. Charalabidis, and A. Zuiderwijk, "Benefits, adoption barriers and myths of open data and open government," Information Systems Management, vol. 29, no. 4, pp. 258–268, 2012. https://scispace.com/pdf/benefits-adoption-barriers-and-myths-of-open-data-and-open-5ccio5woyu.pdf

[4] J.Van Dijk, The Digital Divide. Cambridge, U.K.: Polity Press, 2020. https://www.researchgate.net/publication/339751402_Jan_Dijk_2020_The_digital_divide_Cambridge_UK_Polity_208_pp_1799_paperback_ISBN_9781509534456

[5] I. Medhi, A. Sagar, and K. Toyama, "Text-free user interfaces for illiterate and semi-literate users," in Proc. Int. Conf. Information and Communication Technologies and Development (ICTD), 2007. https://www.researchgate.net/publication/339488825_Investigating_the_Use_of_Email_Application_in_Illiterate_and_Semi-Illiterate_Population

[6] D. Kakwani et al., "Indic NLP Suite: Monolingual corpora, evaluation benchmarks and pre-trained models for Indian languages," in Findings of EMNLP 2020, 2020. https://aclanthology.org/2020.findings-emnlp.445.pdf

[7] A. Banerjee and E. Duflo, Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty. New York, NY, USA: PublicAffairs, 2011. https://www.researchgate.net/publication/233756

407_Poor_Economics_A_Radical_Rethinking_of_The_Way_to_Fight_Global_Poverty

[8] S. Amershi et al., "Guidelines for human-AI interaction," in Proc. CHI Conf. Human Factors in Computing Systems, 2019. https://www.microsoft.com/en-us/research/wp-content/uploads/2019/01/Guidelines-for-Human-AI-Interaction-camera-ready.pdf

[9] B. W. Wirtz, J. C. Weyerer, and C. Geyer, Artificial Intelligence and the Public Sector Applications and Challenges," International Journal of Public Administration, vol. 42, no. 7, pp.596–615, Jul. 2018, doi: 10.1080/01900692.2018.1498103.

[10] I. Medhi, M. Cutrell, and K. Toyama, "It's not just illiteracy: Designing for low-literate users," in Proc. ICTD, 2010. https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/Medhi_IndiaHCI2010.pdf

[11] M. Plauché, U. Nallasamy, J. Pal, R. Wooters, and D. Ramachandran, "Speech interfaces for equitable access to information technology," Int. J. Human-Computer Studies, vol. 69, no. 9, pp. 633–646, 2011. https://itidjournal.org/index.php/itid/article/download/245/245-581-2-PB.pdf

[12] J. Sherwani, "Speech interfaces for information access by low literate users," Ph.D. dissertation, MIT, Cambridge, MA, USA, 2009. https://www.cs.cmu.edu/~jsherwan/JS-thesis.pdf

[13] F. Baldauf, "Towards conversational e-government: Requirements and opportunities of voice-based citizen services," in Proc. Int. Conf. e-Government, e-Business and e-Health, 2018. https://matthiasbaldauf.com/publications/Baldauf20b.pdf

[14] C. Savelli, M. Janssen, and A. Zuiderwijk, "From e-government to AI-enabled government: A systematic literature review," Government Information Quarterly, vol. 39, no. 2, 2022. https://doi.org/10.3390/informatics12030098

[15] J. Segura-Tinoco, F. L. G. Nunes and M. L. S. de Souza, "A conversational agent for argument-driven e-participation," in Proc. ICEGOV, 2019. http://arantxa.ii.uam.es/~cantador/doc/2022/dgo22.pdf

[16] A. P. Gurnani, S. R. Yadav, and R. Manmatha, "TRINS: Towards multimodal language models that can read," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2022. https://openaccess.thecvf.com/content/CVPR2024/papers/Zhang_TRINS_Towards_Multimodal_Language_Models_that_Can_Read_CVPR_2024_paper.pdf

[17] Y. Wang, Y. Chen, and H. Li, "OCR improves machine translation for low-resource languages," arXiv:2109.06453, 2021. https://aclanthology.org/2022.findings-acl.92.pdf

[18] D. Kakwani et al., "Indic NLP: Challenges in processing Indian languages," in Proc. ACL, 2020. https://arxiv.org/html/2501.13912v1

[19] A. B. Araya, D. T. H. Nguyen, and M. J. O'Grady, "Visual conversational interfaces to empower low-literacy users," in Proc. IFIP INTERACT, 2019. https://www.researchgate.net/publication/299700719_Visual_Conversational_Interfaces_to_Empower_Low-Literacy_Users

[20] S. Kusal et al., "AI-based conversational agents: A scoping review from technologies to future directions," IEEE Access, 2022. https://www.researchgate.net/publication/362899303_AI-based_Conversational_Agents_A_Scoping_Review_from_Technologies_to_Future_Directions/fulltext/63065eb01ddd447021071e98/AI-Based-Conversational-Agents-A-Scoping-Review-From-Technologies-to-Future-Directions.pdf