Early Results from Automating Voice-based Question-Answering Services Among Low-income Populations in India

Aman Khullar
aman.khullar@oniondev.com
Gram Vaani
India

M Santosh
m.santosh.mt118@maths.iitd.ac.in
IIT Delhi
India

Praveen Kumar
praveen.kumar@oniondev.com
Gram Vaani
India

Shoaib Rahman
shoaib.rahman@oniondev.com
Gram Vaani
India

Rajeshwari Tripathi
rajeshwari.tripathi@oniondev.com
Gram Vaani
India

Deepak Kumar
deepak.kumar@oniondev.com
Gram Vaani
India

Sangeeta Saini
sangeeta.saini@gramvaani.org
Gram Vaani
India

Rachit Pandey
rachit.pandey@oniondev.com
Gram Vaani
India

Aaditeshwar Seth
aseth@cse.iitd.ac.in
Gram Vaani, IIT Delhi
India

ABSTRACT

Question-answering systems where users can ask questions based on emergent needs which are then answered by experts or peers, have emerged as an important information seeking modality on digital platforms. Automating this process has been an active area of research since many years, to identify relevant answers from pre-existing question-answer databases. We report on the feasibility of running automated question-answering systems in the context of rural and less-literate users in India, accessed through IVR (Interactive Voice Response) systems. We use commercial speech recognition APIs to convert audio questions asked by users into their equivalent transcripts in real time, in Hindi, and use deep-learning based architectures to retrieve corresponding candidate answers which are instantly played to the users. We report several insights from an earlier phase of running question-answering programmes through a manual operation, to how it was transitioned to an automated setup, and document the user experiences during this journey.

CCS CONCEPTS

• **Information systems → Speech / audio search**; **Question answering**; • **Human-centered computing → Interactive systems and tools**; • **Computing methodologies → Machine learning**; • **Applied computing → Health care information systems**.

KEYWORDS

Interactive Voice Response systems, question-answering, FAQ retrieval, speech recognition, natural language processing

ACM Reference Format:
Aman Khullar, M Santosh, Praveen Kumar, Shoaib Rahman, Rajeshwari Tripathi, Deepak Kumar, Sangeeta Saini, Rachit Pandey, and Aaditeshwar Seth. 2021. Early Results from Automating Voice-based Question-Answering Services Among Low-income Populations in India. In *ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS) (COMPASS '21), June 28-July 2, 2021, Virtual Event, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3460112.3471946

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
COMPASS '21, June 28-July 2, 2021, Virtual Event, Australia
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8453-7/21/06…$15.00
https://doi.org/10.1145/3460112.3471946

1 INTRODUCTION

Question-answering is an effective technique to help people fulfill information needs on-demand, as and when they arise. Search engines, chatbots, and question-answering forums like Quora, Reddit, and Stack Overflow are widely used tools for query resolution. These tools however require internet access, smartphones, laptops or computers, and high digital literacy by the participants. Voice-based systems running on IVR (Interactive Voice Response) [15, 16, 19, 20, 24] can help bridge this divide and allow people to listen to audio recordings, ask questions by recording audio messages, and even provide answers to these messages for peer-to-peer knowledge sharing. This can create community based learning environments even among less-literate people without requiring the internet, and accessible through simple feature phones [23]. In this paper, we present our learning from running two question-answering programmes on voice-based discussion forums, which were transitioned from a manual question-answering process to an automated method of FAQ (Frequently Asked Questions) retrieval using machine learning techniques.

Gram Vaani, a social enterprise running several voice-based forums for low-income populations in India, operates two question-answering programmes: "*Mera Sawaal Hai (MSH)*" (My Question is...) and "*Poocho Aur Jaano (PAJ)*" (Ask and Learn), on the JEEViKA and Saajha Manch Mobile Vaani IVR platforms respectively. Both the programmes are in Hindi, the common local language in most

of North India. The JEEViKA platform runs in several blocks of the Nalanda district in the state of Bihar and is used by women SHG (Self Help Group) members to access and share information on health and nutrition practices for pregnant mothers and small children, livelihood, and agricultural best practices [7, 23]. The Saajha Manch platform, on the other hand, runs in the National Capital Region of Delhi, and is used mostly by male organized sector workers employed in automotive and garment factories to discuss issues on labour rights and working conditions [21, 23]. The MSH and PAJ programmes were originally run in a manual fashion. Users could call and record questions. A team of content moderators would listen to these questions, and route the valid ones to domain experts from among partner civil society organizations knowledgeable about the respective topics. On a weekly basis, answers would be collected from the experts, recorded, and pushed back over a call to the person who asked the question. The question and answer pairs would also be played on the JEEViKA and Saajha Manch platforms for other users to listen and benefit from them. During a deployment span of 12 for MSH and 41 months for PAJ, almost 400 and 1800 questions respectively were asked by the users. Not all the questions were unique, many questions could be answered by recordings already prepared when a similar question was asked earlier, and the experts kept this in mind when providing the answers.

We used this dataset of questions and answer pairs to prepare an automated version of the service. Users could record a question as before, for which we obtained a transcript in real-time through commercial Google Speech APIs for ASR (Automated Speech Recognition) [11]. This transcript of the question was used to obtain matching answers by formulating a machine learning task to query the question-answer database that had been accumulated so far. Three top-ranked answers provided by the machine learning models were then played back to the users. Our expectation was that being able to obtain instantaneous answers rather than wait for up to a week, would be useful for the callers. We also provided a feature for them to give feedback on whether or not their question had been satisfactorily answered. If not, then the questions were routed through the earlier manual process to obtain an answer through an expert, which also served to augment the database with new questions and answers.

In this paper, we present some early results about the relevance of question-answering systems, and the process of moving towards automated systems by utilizing recent advances in speech recognition and natural language processing in low-resource languages. We first describe user feedback on the manual question-answering system. We then present technical details of preparing machine learning models for automated question-answering, which gave us an accuracy of the order of 70% and 77% to find a matching answer among the top-3 results for a query. Finally we present early usage results and user feedback from deploying these automated models. Our findings will be useful for other researchers and practitioners, especially those working with voice-based forums, on the technological readiness and user capacity to deploy and scale such systems.

2 RELATED WORK

In recent years, IVR platforms have emerged as popular information sharing tools in several social development interventions, including for local news and announcements [15], behavior change communication [7], social accountability and grievance redressal [6, 14, 16], and citizen feedback on social welfare schemes [8]. Question-answering is an important format for information seeking and sharing, and Avaaj Otalo [19] was among one of the first applications to demonstrate this in the agricultural context. This has now been applied to many domains including question-answering about sexual and reproductive health issues, labour rights, health and nutrition, and eligibility for government schemes, among others [23].

Most such voice-based platforms have so far been operated manually, but with advances in speech recognition and natural language processing for low resource languages [4, 9, 13, 22], initial attempts have been made to apply automation on audio data for information retrieval tasks. Topic modeling on transcribed Hindi audio was used to develop a prototype chatbot on smartphones to retrieve learning videos [17, 27]. Another conversational agent, called FarmChat [12], used a large corpus of recorded phone conversations of an agricultural consultancy service to build a language model for retrieving factoid based information nuggets from a pre-existing knowledge base. Our work is related but rather than a chatbot for factoid queries or retrieving learning videos, we build and deploy an automated question-answering service accessible over IVR, for users to seek advice and information on specific topics by asking unstructured questions through natural speech. This we believe is relevant to serve the information needs of people without internet or smartphones, and limited capability to navigate user interfaces. We expand on the lessons learnt from our own prior work in automating IVR-based question-answering systems [5], to improving the retrieval techniques and deploying the system. In particular, we present the process we followed to build question-answering datasets, build-up on new question-answering models and data augmentation techniques to improve the system accuracy, and then present the lessons learnt from our initial experiences of the system's field deployment and user interviews.

3 MANUAL QUESTION-ANSWERING SYSTEM

We begin with describing the manual question-answering system which was used to build the dataset for subsequent automation. We also present user feedback collected through semi-structured phone interviews, which provides important insights on the advantages and limitations of question-answering systems.

3.1 Process Flow

Figure 1 illustrates the process flow for the MSH and PAJ manual question-answering programmes. The process involves a coordinated effort between different teams: the field team, content moderation team, and domain-experts. The field team runs training and demonstration workshops among the users, to inform them about the IVR systems, how they work, and how to ask questions or record messages to share updates. In the JEEViKA context, the field team trains community mobilizers responsible for coordination and book-keeping of regular income and saving activities of

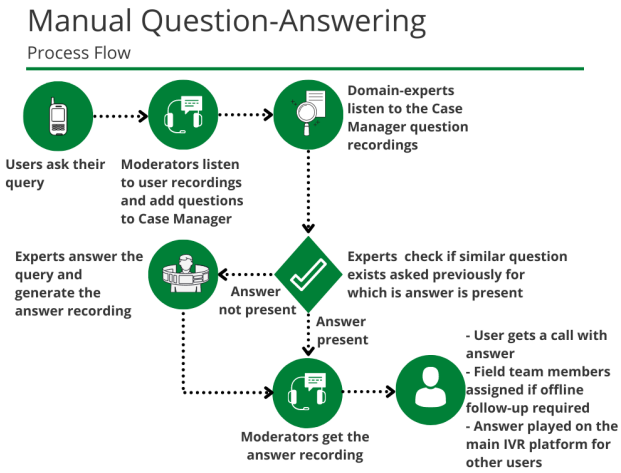

Figure 1: Process flow for the manual question-answering system

the SHGs, who in turn inform the women SHG members about JEEViKA Mobile Vaani and the associated MSH question-answering programme. In the Saajha Manch context, the field team identifies and trains volunteers from among workers employed in factories in urban areas, who in turn spread the word about the platform among other industrial sector workers and encourage them to share factory news, reports on violations of humane working conditions and other regulations, and ask questions to seek advice on possible steps they can undertake. Most such field activity occurred prior to the COVID-19 outbreak; it had to be suspended for most of 2020, during which time the platforms continued to be actively used mostly by those who were already familiar with them. Any questions and updates recorded on the IVR platforms are heard by a trained team of content moderators who understand the local language and are able to determine the next steps to undertake with the voice recordings. News or opinions shared on different topics are published on the IVR platforms if they pass certain editorial guidelines, while any recorded questions are passed on to a team of domain-experts. These experts, some of whom are internal to the Gram Vaani team and some are affiliated with partner organizations, are responsible to provide answers to these questions and create audio recordings.

The experts first need to decide whether they have encountered a similar question earlier and already provided a suitable answer for it, or whether this is a new question that has not been asked before. The first case simply requires the experts to provide the earlier answer available in the question-answer dataset composed by them so far. The new question is however added to the dataset, thereby building up a set of many questions all of which have the same answer. As we will show, this is important to build machine learning models that can recognize different ways in which the same question could have been asked. In the second case when a new question is encountered that has not been answered earlier, the experts provide a new answer and the Gram Vaani team creates an audio recording for it. The moderation team then uploads this recording to automatically push a call to the user with the answer, attached to which also is a short IVR feedback survey that the user can answer through keypresses to indicate whether they are satisfied with the answer or not. All question-answer pairs collected through the duration of a week are also played on the main JEEViKA and Saajha Manch Mobile Vaani platforms so that all users can hear and benefit from the questions and answers.

3.2 User Feedback

The keypress-based satisfaction survey revealed a 77% user-satisfaction rate, and we went deeper to gain an understanding through qualitative means about the merits and demerits of the question-answering programme. A user feedback exercise was carried out through semi-structured phone interviews with 64 users who had put up questions on PAJ and 19 users from MSH. Due to the COVID-19 outbreak, these interviews had to be conducted remotely.

Most responses indicated that a significant benefit seen by people was the ability to ask questions at any time of day. Some of the JEEViKA SHG members responded that the IVR information source was more convenient to use than resolving queries through community health workers who were not available at all times. Some people also said that they ask questions to cross-check the information they receive from some other sources, and pointed out that the information provided through the IVR platform came from experienced doctors and was more authoritative. They often shared this information with others in their family too:

"*I asked about what I should eat during pregnancy and I received the answer in very interactive manner which I also shared with my mother-in-law. I now eat according to the provided recommendation as much as possible*" — Female user, Nalanda, Bihar.

Some users also pointed out drawbacks of such one-shot question-answer systems. The JEEViKA users gave several examples where they did receive generic information corresponding to their question, but their motive behind the question had not been understood clearly and therefore the information was inadequate. One such example was a question about the complexities when a normal child delivery cannot happen. The answer provided detailed information on what sometimes hinders a normal delivery and what to do under such circumstances, but the user was looking for information on how to ensure a normal delivery:

"*All this information is good but I wanted to know what could be done during pregnancy, to prevent an operation*" — Female user, Nalanda, Bihar.

Another example cited was a question about "*how long does the Corona virus last on a surface*". The answer provided useful information about how the virus operates and why it can linger on surfaces, but the user was looking for specific information on how to handle metal surfaces, vegetables, packages, etc. in daily life.

This clearly shows that a single request-response question-answering system may not be adequate as compared to having an actual conversation with an expert, but asynchronous information exchange by consulting an archive does bring non-zero benefits.

We also learned that although women may receive sufficient actionable information, they may not be able to action the steps due to other challenges related to financial constraints or social-norms. Some users who asked questions pertaining to food and diet intake during pregnancy said that the answers provided useful information

but they did not have the resources to change their dietary practices. The job and income loss caused due to the pandemic created further stress among many households:

"*Pregnant women know what they should drink and eat but since there are no jobs because of Coronavirus, how will we purchase meat or fruits?*" — Female user, Nalanda Bihar.

This lack of ability to act was even more severe on Saajha Manch, where many workers were rendered out of jobs and required social protection measures but were unable to access them. The users indicated that they got useful information to questions about social entitlements, such as whom to approach to submit an application or file a grievance, but often these actions did not yield any benefit because of uncooperative administrative staff and previous employers who were not helpful. One user whose wages had not been cleared by his former contractor, submitted a written complaint in several offices but no action was taken by the authorities:

"*I have been working as a security guard in the State Bank of India since the last 7 years. My contractor fired me during the lockdown and did not even give my salary. My father is also unwell in my village. I have lodged complaints in every department but have not received any help. Can Saajha Manch provide any help?*" — Male user, Delhi NCR.

Another account was of a user who wanted to rectify his matriculation certificate as proof of his date of birth, required for subsequent paperwork. He shared that he underwent a lot of harassment at the hands of the state examination board. The officials first called him in-person to their head-office in Varanasi, but when he arrived he was informed that all his documents were supposed to arrive by post. Even after submitting by post, he did not have any information about the current status of his request.

Such cases additionally highlight the need for relief or offline grievance redressal assistance required by the users, beyond the information itself. The Gram Vaani team of local volunteers indeed channeled many such cases for action during the COVID-19 lockdown, to facilitate the delivery of food packages for vulnerable families, grievance redressal and guidance for access to social entitlements, cash transfers for workers who were stranded in cities without food or cash, demand registration for transportation of stranded migrant workers from cities to villages, among others [1, 2]. However, despite these caveats, question-answering systems were found to be useful by most users. We next describe how the dataset built through the manual question-answering process was used to build an automated system.

4 AUTOMATED QUESTION-ANSWERING

To automate the question-answering process, the key machine learning problem is that given a query, and a database of questions and answers, we need to obtain a ranked list of answers from the database that are relevant to the query. As framed in prior work [5], the database originally composed of audio recordings of questions and answers is manually transcribed word for word, incoming audio queries are transcribed through ASR APIs, and text-based FAQ retrieval models are then used to identify relevant answers. A variety of models can be evaluated, based on question-question similarity in which incoming queries are matched against questions in the dataset, question-answer similarity where incoming queries are matched against the answers, and combinations of these two approaches. The similarity matching itself can be done through keyword or word-vector or other approaches. In this section, we describe the evaluation of different methods on the MSH and PAJ datasets. The code for our experiments is available online[1].

4.1 Question-Answer Dataset

The datasets for MSH and PAJ created through the manual question-answering system are in the form of question-answer pairs. Questions that have the same answer are grouped together, to represent different ways in which the question was asked. The dataset is also divided into broad themes, so that if the theme information is available then the search for an answer can be conducted only within that theme. An improved performance was noticed in prior work when theme information was also provided to restrict the search space [5]. The MSH dataset spans six themes in health and nutrition: Maternal Nutrition, Child Nutrition, Menstruation, Coronavirus, Encephalitis, and Others. The PAJ dataset spans seven themes related to labour rights and livelihood: Provident Fund enrollment, Provident Fund withdrawal, Employee State Insurance, Pension schemes, Public Distribution System for subsidized food, Rural Employment Guarantee Act, and working conditions and wage related queries. Table 1 shows the number of themes, queries, and answers for both the datasets.

Since ASR APIs are not very accurate, prior work showed that models trained on manually transcribed questions and answers performed better than if the ASR output for question and answer recordings was used [5]. We therefore recruited a vendor to manually transcribe the entire dataset of questions and answers in MSH and PAJ. We also retained the ASR output for the questions for testing purposes, since in the real setting we would obtain an ASR output for user queries that would be matched against the question-answer datasets.

Prior work further showed that users may sometimes record very wordy questions with superfluous information, and that training models only on relevant portions of the questions would perform better than if the entire question text was used [5]. Based on advice from the domain experts, the Gram Vaani moderators further annotated the questions into relevant, possibly relevant, and irrelevant segments.

The final dataset has the following structure: (T, Q, RQ, A, RA, STT). Here, T denotes the theme of the questions, Q denotes the manual transcription of the questions, and RQ denotes the relevant portion of the question texts. Similarly, A and RA are the manual transcriptions of the answers and the relevant portion of the answers respectively. STT denotes the Speech-to-Text ASR transcripts of the audio recordings of the questions, obtained through

[1] https://github.com/ICTD-IITD/Voice_App_Automated_QnA

Table 1: Dataset Statistics

Dataset	Themes	No. of Queries	No. of Answers
PAJ	7	250	47
MSH	6	394	69

the Google ASR APIs. We use q to denote an incoming user query, obtained through the ASR API. In our experiments, we train the models using (T, Q, RQ, A, RA, STT) and test them on a query q not belonging to Q or STT.

4.2 Models

We implemented six models using the following techniques: Jaccard, Weighted Jaccard, Multilingual BERT (M-BERT), IndicBERT, Flair, and iNLTK. For each model, we evaluated three variants, for question-question similarity (q, Q), question-relevant portion similarity (q, RQ), and question-answer similarity (q, RA/A).

4.2.1 Jaccard. This uses the traditional Jaccard similarity method [25] which computes a similarity score between the words in the user query q against each of the questions Q in the dataset, and identifies the questions with the highest similarity.

4.2.2 Weighted Jaccard. This is a variant on the above method where keywords are weighted based on their TFIDF (Term Frequency Inverse Document Frequency) scores. The TFIDF scores are computed by treating each theme as a separate document composed through a concatenation of the questions (or answers) in the theme.

4.2.3 Multilingual BERT (M-BERT). Transformer based architectures such as BERT are known to improve many NLP tasks in information retrieval and entity extraction [10]. We use the pre-trained M-BERT model to define a classification task that takes a pair of questions (q1, q2) and returns true if the questions are similar, or false otherwise. The positive samples in this dataset are comprised of pairs of questions belonging to the same group, i.e. having the same answer. Negative samples are constructed through random sampling by forming pairs of questions that belong to different groups. We used the HuggingFace library [26] and PyTorch [18] for implementing the model.

4.2.4 IndicBERT. IndicBERT [13] is another transformer based architecture that is known to produce good results on NLP tasks in Indian languages. The input and output format is same to M-BERT.

4.2.5 Flair. Flair is word-vector based publicly available library with ready made functions for NLP tasks in several languages [3]. We compute the question-question or question-answer similarity as the dot product of sentence embeddings derived through the Flair architecture.

4.2.6 iNLTK. Similar to Flair, iNLTK helps obtain sentence embeddings based on a pre-trained language model for Indian languages [4]. A cosine similarity on the embeddings is used to obtain a score for question-question or question-answer similarity.

While evaluating the performance of these models, as suggested in our prior work [5], we assume that the theme is provided as an input by the user, to restrict the search space for each query.

4.3 Data Pre-processing

We experimented with several techniques for data augmentation to increase the size of our dataset. In the first setting which we refer to as *No Augmentation*, we use only the base dataset for MSH and PAJ for their respective models. In the second setting referred to as *Domain Adaptation*, we concatenate the datasets of MSH and PAJ together. We also include another dataset we used in prior work of questions and answers related to sexual and reproductive rights and health, called the KAB ("*Kahi Ankahi Baatein*" – said and unsaid stories) programme [5]. The third setting referred to as *Data Augmentation* consists of a list of synonyms for theme-specific words provided by the domain experts. Multiple copies of a query are made by replacing words with their synonyms. We also used the iNLTK library which provides a feature to generate sentences similar to a given sentence, and created two additional sentences for each question in our dataset.

For each of these experiments, the training and testing data for evaluation is constructed following the same methodology as used in prior work [5]. For each group of similar questions having the same answers, we do a 50/50 split if that group has more than 10 questions, else a 70/30 split, for training and testing respectively.

4.4 Post-processing

A post-processing step is included to ensure that the returned matches do not belong to the same group. This improves the results by preventing repetition in the retrieved answers.

4.5 Results

We evaluate the models using 4 metrics: Success rate in the top one (SR@1), three (SR@3), or five (SR@5) ranks, and the Mean Reciprocal Rank (MRR). Success rate @K is defined as the fraction of tests in which a correct answer was returned in the top-K results. The MRR metric further takes into account the position at which the first correct answer was returned in the tests. The test data uses the ASR transcripts of the user queries, while the variants of the models are trained on (Q, Q), (Q, RQ) and (Q, RA/A) on clean manually transcribed data. We choose the best variant of a model trained on either (Q, Q) or (Q, RQ), or by taking a model ensemble with the variant trained on (Q, RA/A). Since we return the top-3 answers to an input query, we use the SR@3 metric to select the best models.

Table 2: Results in No augmentation setting. The best and second best results are bold and underlined respectively.

Model	Dataset	SR@1	SR@3	SR@5	MRR
Jaccard	MSH	0.470	<u>0.664</u>	0.724	0.567
	PAJ	0.354	0.658	0.785	0.536
Weighted Jaccard	MSH	0.373	0.552	0.649	0.494
	PAJ	0.468	<u>0.722</u>	0.759	0.603
M-BERT	MSH	0.463	**0.687**	0.746	0.588
	PAJ	0.443	**0.734**	0.810	0.609
IndicBERT	MSH	0.291	0.560	0.642	0.445
	PAJ	0.392	0.709	0.848	0.570
Flair	MSH	0.313	0.530	0.604	0.449
	PAJ	0.215	0.557	0.658	0.421
iNLTK	MSH	0.351	0.507	0.590	0.458
	PAJ	0.278	0.595	0.772	0.477

Table 3: Results in Domain Adaptation and Data Augmentation on MSH Test Data. The best and second best results are bold and underlined respectively. M, P, K denotes MSH, PAJ and KAB respectively and Aug M, P, K denotes Augmented MSH, PAJ and KAB datasets.

Model	Train Dataset	SR@1	SR@3	SR@5	MRR
M-BERT	M	0.463	0.687	0.746	0.588
	M, K	0.493	0.679	0.784	0.610
	M, P	0.530	0.679	0.746	0.624
	M, P, K	0.433	**0.701**	0.769	0.579
	Aug M, P, K	0.321	0.604	0.716	0.480
IndicBERT	M	0.291	0.560	0.642	0.445
	M, K	0.396	0.590	0.679	0.524
	M, P	0.246	0.530	0.642	0.412
	M, P, K	0.381	0.612	0.709	0.516
	Aug M, P, K	0.194	0.485	0.657	0.377

4.5.1 No augmentation. In this setting, we observe that the best results are obtained on (Q, Q) or (Q, RQ) similarity. M-BERT performs better than other models, and the next best performance is by using the Jaccard model for MSH and Weighted Jaccard for PAJ. Table 2 shows the results for this setting.

4.5.2 Domain Adaptation. Tables 3 and 4 show the results of *Domain Adaptation* on MSH and PAJ respectively. M-BERT gives the best results on both MSH and PAJ when trained with datasets from all three programmes: MSH, PAJ, and KAB. IndicBERT also performs better on PAJ when trained on data from all 3 programmes. We can thus infer that including different domains of data from the same language generally has a positive impact on model performance.

4.5.3 Data Augmentation. Building upon the accuracy achieved through domain adaptation, we observed a further improvement in M-BERT's SR@1 and MRR accuracy on PAJ while the accuracy on MSH did not have a significant improvement and rather experienced an accuracy dip (as shown in tables 4 and 3 respectively) which shows that data augmentation does not always lead to an accuracy improvement and can also introduce noise when the dataset sizes are small.

Table 4: Results in Domain Adaptation and Data Augmentation on PAJ Test Data. The best and second best results are bold and underlined respectively. M, P, K denotes MSH, PAJ and KAB respectively and Aug M, P, K denotes Augmented MSH, PAJ and KAB datasets.

Model	Train Dataset	SR@1	SR@3	SR@5	MRR
M-BERT	P	0.443	0.734	0.810	0.609
	P, K	0.456	0.747	0.848	0.628
	P, M	0.456	0.734	0.848	0.622
	P, M, K	0.443	0.772	0.861	0.620
	Aug M, P, K	0.557	**0.772**	0.861	0.678
IndicBERT	P	0.392	0.709	0.848	0.570
	P, K	0.418	0.722	0.861	0.592
	P, M	0.392	0.734	0.823	0.580
	P, M, K	0.430	0.722	0.810	0.605
	Aug M, P, K	0.291	0.671	0.823	0.502

4.5.4 Automatic Theme Identification. In our previous work [5] we reported how theme identification leads to better system accuracy by restricting the model search space. In this experiment, we tested if we can automate this step of theme identification from the user query itself and do away with any manual theme selection by the user.

We trained six theme classifiers on the MSH and the PAJ datasets and tested the performance of the classifiers using the query STT. We observed that even the best performing classifier obtained an F1 score of 43% on the MSH dataset. We obtained an SR@3 accuracy of 55% using the M-BERT after automating this step of theme selection as compared to the 69% SR@3 accuracy in the No augmentation setting. We therefore did not attempt to field-test an automatic theme identification question-answering model and instead evaluated if keypress-based manual theme selection by users is feasible or not, which has been described in more detail in section 5.

4.5.5 Final models. The best performance for the two programmes was obtained as follows:

- **MSH**: M-BERT with (Q, RQ) similarity trained through *Domain Adaptation* on all the three programmes gave the best results, with an SR@3 of 0.701. Jaccard based similarity also performs well with an SR@3 of 0.664.
- **PAJ**: M-BERT with (Q, Q) similarity trained through *Data Augmentation* and *Domain Adaptation* on all the three programmes gave the best results, with an SR@3 of 0.772. The overall performance of Jaccard based similarity is also reasonable with an SR@3 of 0.658.

5 FIELD DEPLOYMENT

We used the models described in the previous section to deploy an automated version of the question-answering system.

5.1 System Design

Figure 2 describes the end-to-end flow of the automated question-answering system. Upon entering the IVR flow for MSH or PAJ, the user is first asked to select the theme on which they want to ask a question. This is put up as a multiple-choice question to be answered through a keypress on the phone. Next the user is asked to speak their question. The recording is instantly sent to the Google ASR API, and the question is looped in case the speech transcription returned by the API is empty. Once a non-empty transcription is obtained, it is used to query the question-answering model as explained in the previous section. Since the Jaccard model runs on CPUs and provides similar accuracy as the deep-learning models that require more expensive GPU infrastructure, we only used the Jaccard model for now. The model returns the top-three answers to the query. Rather than playing these answers directly one after the other, we also prepared a "sanitized question" recording for each answer and play the question-answer combination as follows: "*Did you mean [sanitized question], if not then press 1 to skip, else continue*

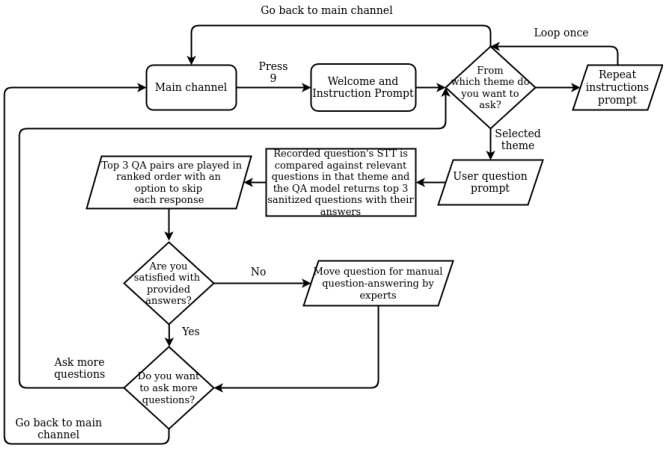

Figure 2: Automated question-answering system design

to listen to the answer", following which the answer plays, then the question-answer pair at the second rank, and finally the third rank. After having played the candidate questions and answers, the users are asked a keypress-based satisfaction question with two options, of whether they are satisfied with the answers or not. In case the user is not satisfied, the question is flagged for the moderators to take it through the regular manual question answering process. Finally, an option is given to the users of whether they want to ask more questions, or to return to the main IVR forum.

Along with this standard flow, in case the user did not select a theme or ask a question within three tries, we assembled a default set of FAQs from among the same set of questions and answers. These are played to the user with an intention of familiarizing them with the concept of questions and answers. The default selection is periodically changed based on temporal relevance, such as questions related to COVID-19 prevention practices on MSH and on social entitlements on PAJ were selected during the first half of the year 2020 when such information was important to reach the people.

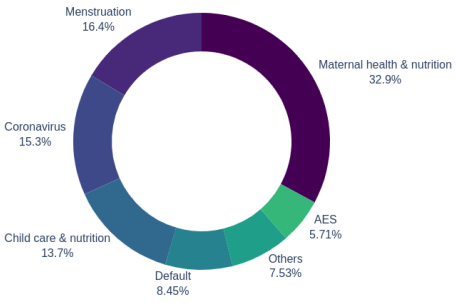

Figure 3: Themes selected by users to ask their questions

5.2 Usage Analysis

Within a span of 14 weeks of deployment of the automated setup, we received 156 relevant questions on MSH from among 438 calls by 241 unique callers.

The automated setup for PAJ was deployed very recently, and therefore the rest of the analysis is presented only for MSH.

Upon analyzing the usage logs, we found that in 92% of the calls a theme was selected from the given set of options, while in 8% cases a theme was not selected and users proceeded to the default theme. Only 36% of the calls led to users asking a question relevant to the health and nutrition focus of MSH, 1% asked a question outside the scope of the programme, and 63% users did not ask any questions and went on to listen the default FAQs from their selected (or default) theme. Among the cases where calls led to question recordings, these were asked across all the provided themes, as shown in Figure 3.

Given the large number of calls where users did not ask questions, we distinguished the accesses between *information seeking* calls and *information browsing* calls. *Information seeking* calls are those where users asked a relevant question, while *information browsing* calls are those where users did not ask a relevant question, or a question at all, but went on to listen to the default FAQs. As shown in table 5, in 83% of the *information seeking* calls the users responded positively to the survey question, similar to the 77% satisfaction rate noticed in the manual question-answering setup. Moreover, we observed that 77% of the *information browsing* calls also led to a satisfied experience, indicating that a substantial fraction of users found the default information to be useful as well.

On the last question of whether users wanted to "*ask more questions or go back to the main IVR forum*", as shown in table 5, we found that 77% of the "information seeking" calls and 75% of the *information browsing* calls went on to ask another question. 81% of these *information seeking* calls were those where the user was satisfied with the answers provided to the previous question, and went to ask more questions, indicating that these users were likely trying out and familiarizing themselves with the new system. Among the *information browsing* calls, of those who went on to ask more questions, 56% had not been satisfied with the default answers that were played to them earlier, indicating that they likely had understood the purpose of the system better by now and chose to give it another try.

We also carried out a production-level evaluation of the system by creating a ground-truth based on the questions asked by the users. We did this by passing on to experts even those questions which had been automatically answered to the satisfaction of the users, and compared the response of the experts against the response generated by the automated models. We also marked whether the recorded questions belonged to the theme selected by the user or not.

We found that among the 92% of the calls where a theme was selected, only 30% led to a question pertaining to that theme, 6% had an incorrect theme (i.e. asked a question relevant to some other theme), and 56% were calls where a question was not asked. The cases of incorrect theme selection suggest that users may find it hard to categorise their query upfront into a specific theme, and it

Table 5: Usage statistics based on the user responses to third and fourth question

User Type	User Experience	Level
Information Seekers	Satisfied	82.79%
	Re-asked	77.19%
Information Browsers	Satisfied	76.56%
	Re-asked	75.40%
Satisfied information Seekers	Re-asked	80.65%
Dissatisfied information seekers	Re-asked	55.56%

may still be worth considering to remove the question to trade-off some accuracy in exchange for easier usability.

Among the questions that were asked, 62% could have been answered from the database available with us, and among these the SR@3 accuracy was 70%. This is similar to the test accuracy reported in Section 4.5, indicating that the production environment is similar to the test environment based on which the models were trained. However, 38% of the questions asked were outside the database of questions and answers available with us. On the one hand, this indicates that similar to the manual question-answering process, an ongoing database expansion is occurring even with the automated setup. On the other hand, this also indicates the need to develop a method to detect questions that are outside our knowledge base so far, so that the users are not given incorrect answers but the question is straightaway parked for manual resolution.

5.3 User Interviews

Our field team additionally facilitated an experimental run with 14 users identified by them. The users were asked by the field team members to access the MSH platform and ask a question, and were also informed that an interviewer will call them to take feedback of their experience. Although a small-scale study, most of the users reported that they found the platform to be useful in terms of instantaneously providing an answer:

"*I asked a question and received the correct response right away*" — Male user, Nalanda, Bihar.

However, some users also expressed dissatisfaction with the accuracy of the answers that were provided. One user asked a question related to blindness, and another about chickenpox (both of which were not present in our set of topics), but the answers provided by the model were about something entirely different. 40% of the users also reported that they found the theme selection to be difficult. Additionally, some users preferred that only one answer should be played instead of three answers. This feedback, and the usage analysis, leads us to conclude that in the next iteration of the automated question-answering service, we may want to remove the theme selection, and incorporate a check to spot out of topic questions and pass them for manual processing. By and large however, despite the limitations of a single request-response question answering setup, and the limited accuracy of automated answer selection, such a design does seem to be useful and promising.

6 CONCLUSIONS

We presented early results from transitioning manually operated question-answering services on IVR systems in rural India, to an automated version. Users had reported a high satisfaction with manual question-answering, although the information they gained may not always have been easily actionable for various reasons, and they continued to report a similar satisfaction even with the automated version. In particular, being able to ask questions and listen to answers in an audio format instead of written text, and receiving an immediate answer instead of having to wait for a few days for domain-experts to respond, was appreciated by the users. Although the automated version is not able to currently achieve very high accuracy, but with some improvements planned in the design, and ongoing improvements with speech recognition and natural language processing techniques, we feel that such systems will get better over time and open up a rich field of conversational question-answering through a voice-based interface for less-literate and low-income users. Some insights we have presented on the design and implementation of such systems is likely to be useful for other researchers and practitioners working in this space.

ACKNOWLEDGMENTS

We would like to express our immense gratitude to the entire Gram Vaani team, without whom none of this work would have been possible. In particular we would like to thank the moderation team who helped in the initial development of the question-answering dataset, the field team, who helped in training users on how to ask their queries on our platforms and our mentor Sayonee Chatterjee who provided us with important design considerations for our on-field deployment. We would also like thank the Bill & Melinda Gates Foundation and Humanity United for funding this project and providing us with valuable feedback on this project. Finally, we also like to thank IIT Delhi for providing us access to its High Performance Computing (HPC) instances, which helped us train our machine learning models.

REFERENCES

[1] Mira Johri Aaditeshwar Seth, Aarushi Gupta. 2021. *Delivery of Social Protection Entitlements in India.* https://drive.google.com/file/d/1hrxF2UP3qZ8Ouo2IdfrTCZwGHLhcFjGX/view

[2] Orlanda Ruthven Aaditeshwar Seth, Sultan Ahmed. 2020. *#NotStatusQuo A campaign to fix the broken social protection systems in India.* https://drive.google.com/file/d/1q2TtBZanO_PhZuLfve9qVQRHav4Wuj5V/view

[3] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art

NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. 54–59.

[4] Gaurav Arora. 2020. iNLTK: Natural Language Toolkit for Indic Languages. arXiv:2009.12534 [cs.CL]

[5] Pranav Bhagat, Sachin Kumar Prajapati, and Aaditeshwar Seth. 2020. Initial Lessons from Building an IVR-based Automated Question-Answering System. In *Proceedings of the 2020 International Conference on Information and Communication Technologies and Development*. 1–5.

[6] Dipanjan Chakraborty, Mohd Sultan Ahmad, and Aaditeshwar Seth. 2017. Findings from a civil society mediated and technology assisted grievance redressal model in rural India. In *Proceedings of the Ninth International Conference on Information and Communication Technologies and Development*. 1–12.

[7] Dipanjan Chakraborty, Akshay Gupta, and Aaditeshwar Seth. 2019. Experiences from a mobile-based behaviour change campaign on maternal and child nutrition in rural India. In *Proceedings of the Tenth International Conference on Information and Communication Technologies and Development*. 1–11.

[8] Dipanjan Chakraborty and Aaditeshwar Seth. 2015. Building citizen engagement into the implementation of welfare schemes in rural India. In *Proceedings of the Seventh International Conference on Information and Communication Technologies and Development*. 1–10.

[9] Jeanne E Daniel, Willie Brink, Ryan Eloff, and Charles Copley. 2019. Towards automating healthcare question answering in a noisy multilingual low-resource setting. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 948–953.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[11] Google. 2021. *Speech To Text*. https://cloud.google.com/speech-to-text

[12] Mohit Jain, Pratyush Kumar, Ishita Bhansali, Q Vera Liao, Khai Truong, and Shwetak Patel. 2018. FarmChat: a conversational agent to answer farmer queries. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 1–22.

[13] Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*.

[14] Meghana Marathe, Jacki O'Neill, Paromita Pain, and William Thies. 2015. Revisiting CGNet Swara and its impact in rural India. In *Proceedings of the Seventh International Conference on Information and Communication Technologies and Development*. 1–10.

[15] Aparna Moitra, Vishnupriya Das, Gram Vaani, Archna Kumar, and Aaditeshwar Seth. 2016. Design lessons from creating a mobile-based community media platform in Rural India. In *Proceedings of the Eighth International Conference on Information and Communication Technologies and Development*. 1–11.

[16] Preeti Mudliar, Jonathan Donner, and William Thies. 2012. Emergent practices around CGNet Swara, voice forum for citizen journalism in rural India. In *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development*. 159–168.

[17] Ankur Pandey, Inshita Mutreja, Saru Brar, and Pushpendra Singh. 2020. Exploring Automated Q&A Support System for Maternal and Child Health in Rural India. In *Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies*. 349–350.

[18] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703* (2019).

[19] Neil Patel, Deepti Chittamuru, Anupam Jain, Paresh Dave, and Tapan S Parikh. 2010. Avaaj otalo: a field study of an interactive voice forum for small farmers in rural india. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 733–742.

[20] Agha Ali Raza, Mansoor Pervaiz, Christina Milo, Samia Razaq, Guy Alster, Jahanzeb Sherwani, Umar Saif, and Roni Rosenfeld. 2012. Viral entertainment as a vehicle for disseminating speech-based services to low-literate users. In *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development*. 350–359.

[21] Orlanda Ruthven. 2018. Labour Reform is Fine But Who Holds Employers to Account When Government Fails? *The Wire* (2018). https://thewire.in/labour/rights-at-work-who-holds-employers-to-account-when-the-government-fails

[22] Huda Sarfraz, Sarmad Hussain, Riffat Bokhari, Agha Ali Raza, Inam Ullah, Zahid Sarfraz, Sophia Pervez, Asad Mustafa, Iqra Javed, and Rahila Parveen. 2010. Large vocabulary continuous speech recognition for Urdu. In *Proceedings of the 8th International Conference on Frontiers of Information Technology*. 1–5.

[23] A Seth, A Gupta, A Moitra, D Kumar, D Chakraborty, L Enoch, O Ruthven, P Panjal, RA Siddiqi, R Singh, et al. 2020. Reflections from Practical Experiences of Managing Participatory Media Platforms for Development. In *Proceedings of the 2020 International Conference on Information and Communication Technologies and Development*. 1–15.

[24] Aditya Vashistha and William Thies. 2012. {IVR} Junction: Building Scalable and Distributed Voice Forums in the Developing World. In *6th USENIX/ACM Workshop on Networked Systems for Developing Regions ({NSDR} 12)*.

[25] Wikipedia contributors. 2021. Jaccard index — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Jaccard_index&oldid=1009813550 [Online; accessed 6-April-2021].

[26] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6

[27] Deepika Yadav, Mayank Gupta, Malolan Chetlur, and Pushpendra Singh. 2018. Automatic annotation of voice forum content for rural users and evaluation of relevance. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*. 1–11.