2023 6th International Conference on Information Systems and Computer Networks (ISCON)
GLA University, Mathura, India. Mar 3-4, 2023

2023 6th International Conference on Information Systems and Computer Networks (ISCON) | 979-8-3503-4696-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISCON57294.2023.10112123

An Integrated Model for Text to Text, Image to Text and Audio to Text Linguistic Conversion using Machine Learning Approach

Aman Raj Singh
Department of CEA
GLA University
Mathura, India
Aman.singh_cs22 @gla.ac.in

Prof. Diwakar Bhardwaj
Department of CEA
GLA University
Mathura, India
diwakar.bhardwaj@gla.ac.in

Mridul Dixit
Department of CEA
GLA University
Mathura, India
mridul.dixit@gla.ac.in

Lalit Kumar
Amity University
Noida, India
Lalitkushwah143@gmail.com

***Abstract*—This paper presents an integrated model that uses machine learning techniques to perform text-to-text, image-to-text, and audio-to-text conversions, with particularly focus on Indian languages. The proposed model which can translate text, image, and voice has been tested on large datasets of various Indian languages and utilizes state-of-the-art techniques such as machine learning, computer vision, and speech recognition to accurately transcribe and translate the input data. The results obtained from the experiments demonstrate the effectiveness of the model by accurately converting text, images, and audio to text, and the potential applications of our proposed model range from language learning, accessibility for non-verbal or non-hearing individuals to cross-language communication. The proposed model is intended to bridge the language gap and facilitate communication among people from different linguistic backgrounds.**

Keywords—image-to-text voice-to-text, voice-to-voice, computer vision; cross language communication)

I. INTRODUCTION

In the age of technology, the ability to seamlessly convert and interpret information across different mediums and modalities has become a fundamental requirement for progress and understanding. In a country as linguistically diverse as India, where the complexity and nuances of our scripts and pronunciations can often prove daunting for even the most advanced machine learning models[6], the need for an integrated model that can handle text-to-text, image-to-text, and audio-to-text conversions is all the more pressing. As someone who has long championed the cause of linguistic pluralism in India, We present to you an integrated model that uses a machine learning approach to convert text, images, and audio to text with a high level of accuracy, thus allowing for better communication and understanding across languages and modalities. We invite you to explore this innovative model and its potential in this paper, as it has the ability to bridge the language divide and facilitate communication among people from different linguistic backgrounds. In this paper, we provide a comprehensive overview of the proposed model, including its architecture, training, and evaluation. We also discuss the potential applications of this model, such as language learning, accessibility for non-verbal or non-hearing individuals, and cross-language communication. We hope that this work will inspire further innovation in the field of machine learning and natural language processing, and contribute to the goal of facilitating communication and understanding across languages and modalities

II. LITERATURE REVIEW

The literature review will provide an overview of existing research on multi-modal machine translation, including image-to-text, voice-to-text, and voice-to-voice translation, as well as the use of machine learning techniques in these areas. It includes a review of the current state-of-the-art models and their performance on various types of inputs and languages.

The previous researches in the field of image-to-text translation has focused on developing models that can accurately transcribe text from images, using techniques such as optical character recognition (OCR)[1]. OCR stands for Optical Character Recognition. It is technology that is used to recognize and extract text from images and scanned documents. OCR software can be used to convert scanned documents in human readable This mechanism makes it possible for the people to edit the text, search for different kinds of words or phrases, display or print a copy of the machine generated code and also it can be used to convert text- to-speech and Machine learning[1]. Similarly, researches on voice-to-text translation has focused on developing models that can accurately transcribe spoken language, using techniques such as speech recognition and natural language processing (NLP)[4][5]. Today, Google ML Translation Kit is a toolkit that allows developers to add machine learning-based translation to their apps. It includes pre-trained models that can be used to translate text in real-time, as well as a set of tools and APIs to customize and train the models.

In similar fashion researches on voice-to-voice translation has been focused on developing models that can translate spoken language in real-time, using techniques such as automatic speech recognition (ASR)[2] and machine translation(MT)[1][2]. This research has been motivated by the need to facilitate communication between people who speak different languages, particularly in the context of globalized business and travel.

In recent years, there has been a growing interest in developing linguistic conversions models using machine translation models that can handle multiple inputs and

modalities. These models have been shown to be more robust and accurate than traditional models, and their applications range from language learning, accessibility for non-verbal or non-hearing individuals, and cross-language communication.

The literature review has opened a discussion of the challenge for the development of the multi-modal machine translation models, such as dealing with different types of inputs, languages through text, image and the video under one umbrella.

III. PROPOSED METHODOLOGY

We aim to create a cutting-edge application utilizing the latest technologies at our disposal. In this Paper, we have developed the multi-model using latest technologies that will aid us in achieving our goal. With the help of OCR (Optical Character Recognition) using Google's powerful OCR engine for image extraction, for audio we have used Google speech recognition, Google Speech-to-Text is a service that uses machine learning to convert spoken audio into written text. It supports a wide range of languages and can be used for a variety of applications, such as transcription, dictation, and voice commands. Additionally, Google Speech-to-Text allows for custom models for specific use-cases like medical, legal and more. We have used ASCII based character recognition for text. Thus Text, Audio or Speech and image recognition utilizing Google's Neural Network based Machine Learning Translator kit, and finally utilizing Java and its libraries to seamlessly integrate all of these technologies into a cohesive final product.

The whole process is divided into basic three steps:

1. Extraction Recognition of Text, Image and Audio forconversion
2. Recognition of Text, Image and Audio for its nativelanguage
3. Conversion of Text, Image and Audio in the desired manner or text.

Steps to get the translated data using this technology are:

Step 1: **Input**: The input to the system is text, image, or audio data on their respected fields.

Step 2: **Data Preprocessing**: The input data undergoes preprocessing to convert it into a format suitable for the model.

Step 3: **Model Selection**: The appropriate Translation model for the input data type is selected, either a text-to-text translation, image-to-text translation, or audio-to-text translation.

A. Text-to-Text:

Google Machine Learning Translation Languages models. Google's Machine Learning Translation kit uses Neural Machine Translation (NMT) models to translate text from one language to another. The working process can be summarized as follows:

1) ***Preprocessing****:* The input text is cleaned and processed to make it compatible with the model's input format.

2) ***Encoding****:* The input text is transformed into a numerical representation, known as an encoding, which the model can understand.

3) ***Decoding:*** The encoded text is then passed through the model's layers, which are designed to learn the relationships between the source language and target language. The model generates a translation, which is in the form of a probability distribution over possible outputs.

4) ***Post processing***: The model's output is transformed back into text and fine-tuned to ensure fluency and grammatical correctness.

The NMT model is trained on large amounts of parallel text data, which is used to learn the relationships between the source and target languages. The model's parameters are then fine-tuned to optimize its performance. The result is a model that is capable of translating text from one language to another with high accuracy and fluency.

B. Image-to-Text:

Google optical character recognition (OCR) models for Text extraction, Google Machine Learning Translation Languages models for Translation.

Google's Optical Character Recognition (OCR) models use computer vision techniques to extract text from an image. The working process can be summarized as follows:

1) ***Image Preprocessing****:* The input image is cleaned and processed to improve the OCR model's performance, such as correcting perspective distortion and enhancing the image's contrast.

2) ***Text Detection****:* The OCR model uses computer vision techniques to identify the text regions in the image and isolate them from the rest of the image.

3) ***Text Recognition****:* The text regions are then fed into a character recognition model that converts the visual representation of the characters into their corresponding ASCII characters.

4) ***Text post processing****:* The output text from the character recognition model is post processed to correct any errors and improve its quality.

For translation, Google's Machine Learning Translation models can then be used to translate the text extracted from the image into another language. The process is the same as for text-to-text translation, as described in the previous answer. The extracted text is cleaned, encoded, passed through the NMT model, and post processed to obtain a fluent and grammatically correct translation.

C. Voice-to-text:

Google Speech reorganization for voice-to-text, Google Machine Learning Translation Languages models for translation.

Google's Speech Recognition uses Automatic Speech Recognition (ASR) models to convert spoken words into text. The working process can be summarized as follows:

*1) **Audio Preprocessing**:* The input audio is cleaned and processed to improve the ASR model's performance, such as removing background noise and enhancing the speech signal.

*2) **Feature Extraction**:* The audio is transformed into a numerical representation, known as features, which capture the speech patterns in the audio.

*3) **Decoding***: The features are then passed through the ASR model, which is designed to learn the relationships between the speech patterns and the corresponding text. The model generates a transcription, which is in the form of a probability distribution over possible outputs.

*4) **Post processing*****:** The model's output is post processed to correct any errors and improve its quality

For translation, Google's Machine Learning Translation models can then be used to translate the text obtained from the speech recognition into another language. The process is the same as for text-to-text translation, as described in the first answer a make grammatically correct translation.

Step 4**: Model Inference**: The selected Machine Learning model processes the preprocessed input data and generates a linguistic representation in desired Language.

Model follows these steps to process the preprocessed input data:

A. **Input:** The preprocessed input data, which can be in the form of text, images, or audio, is fed into the model.

B. **Embedding**: The input data is transformed into a numerical representation, known as an embedding, which captures the semantic meaning of the data.

C. **Encoder:** The encoded representation is passed through a neural network known as an encoder, which summarizes the information in the input data into a compact representation.

D. **Decoder:** The compact representation is passed through another neural network known as a decoder, which generates the output in the desired language.

E. **Output:** The output generated by the decoder is the linguistic representation in the desired language, which can be in the form of text, images, or audio, depending on the type of data fed into the model.

F. **Evaluation:** The output generated by the model is evaluated against a reference or ground truth to measure the accuracy and quality of the output. The model may be trained further to improve its performance based on the evaluation results.

Step 5: **Output**: The model outputs the generated linguistic representation, which can be text, image, or audio data and that would be displayed on user screen.

Step 6: **Post-processing**: The output data undergoes post-processing, such as formatting, to produce the final result.

Step 7: **Feedback**: The system receives feedback on the quality of the output data and adjusts its internal parameters accordingly.

Step 8: **Repeat**: The process is repeated for new input data, continuously improving the system's performance through the use of feedback.

Fig: 1 shows the extract process. How the image Text will we extract using The OCR Library If the input is an electronic document, then it extracts the text from the image and converts it into ASCII and converts it into Desired Language. Similarly using Google speech recognition, we can extract the speech and using ASCII based character recognition we can extract the text.

Shows as How the Translation work using Google machine learning kit Input text can translation using predefined Machine learning Language Models and gives the Desired Output as a translated text.

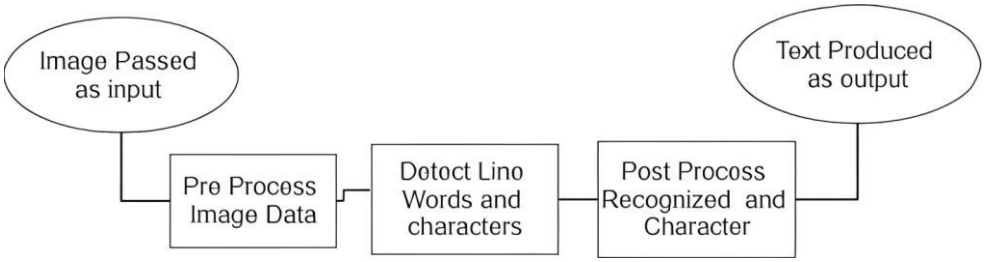

Fig. 1. the extraction process of the proposed model using OCR

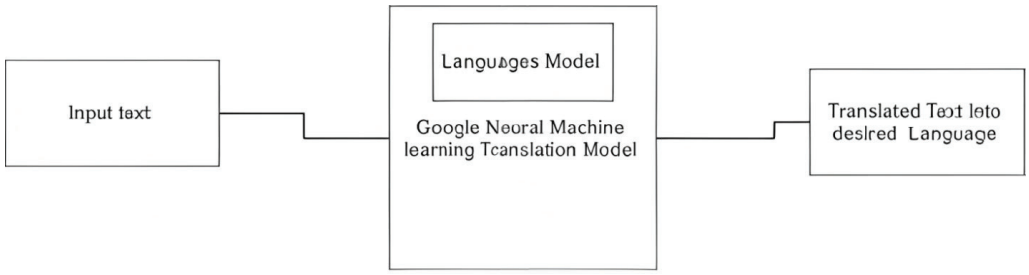

Fig. 2. The Translation work of the proposed model using Google machine learning kit

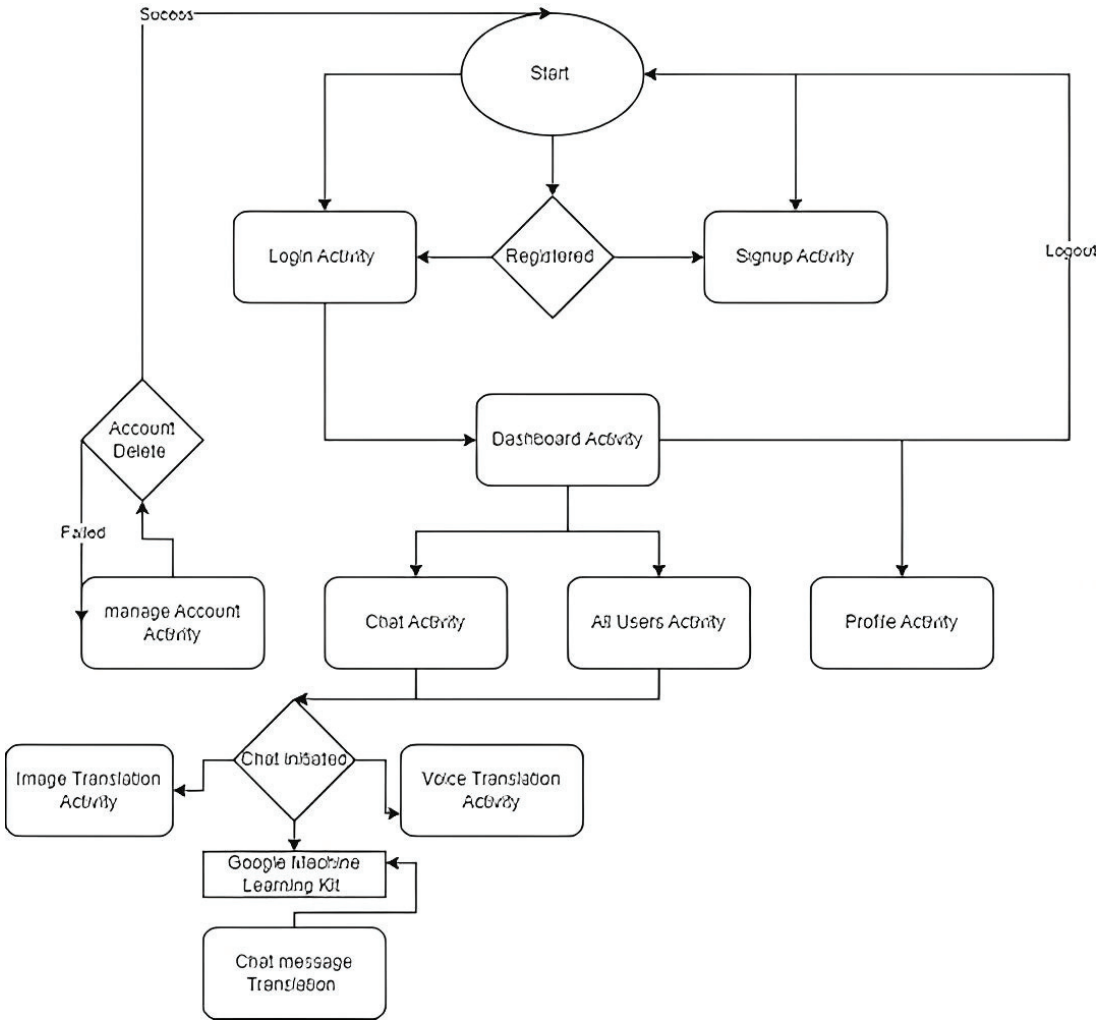

Fig. 3. Proposed Architecture

Figure: 3 Shows the Work flow of application how Activity of Application will be works this is based on android applicationfor future research it is need to be implemented.

IV. RESULTS

A. Data Set

The data set for an integrated model for text to text, image to text, and audio to text linguistic conversion using Machine Learning approach would likely include the following types of data:

Text to Text:
Source texts in one language
Target texts in another language
Aligned Parallel corpus

Image to Text:
Images
Captions or descriptions of the images
Tags or labels associated with the images

Audio to Text:
Audio files (e.g. in MP3 or WAV format)
Transcriptions of the audio (e.g. in text format)

Test Set we have taken 1000 test message for testing our proposed for text-to-text conversion

TABLE I. TEXT MESSAGING INPUT TEST SAMPLES ACCURACY

S.No	Text-to-Text		
	Test Case Name	***Pass***	***Failed***
1	1000 Text Messages Input	856	144
Ac	**Total Average Accurecy of Text**	85.6%	

TABLE II. IMAGE-TO-TEXT INPUT TEST SAMPLES ACCURACY

S.No	Image-to-text		
	Test Case Name	***pass***	***Fail***
1	100 poor quality images	89	11
2	100 medium quality images	94	6
3	100 Good quality images	98	2
Acc	**Total Average Accurecy of image Extraction**	93.73%	

TABLE III. VOICE-TO-TEXT INPUT TEST SAMPLES ACCURACY

S.No	Voice to Text		
	Text Case Name	*pass*	*Fail*
1	100 voice Inputs	85	15
Acc	**Total Average Accuracy of Voice to Text**	85.15%	

An integrated multi-model for text-to-text, image-to-text, and audio-to-text linguistic conversion using a machine learning approach have involved the use of a multi-modal neural network that combines the features and capabilities of several models. This network could include natural language processing (NLP) for text-to-text conversion, computer vision for image-to-text conversion, and speech recognition for audio-to-text conversion. This model is then be used to perform conversions between the different modalities, such as translation speech to text or describing the content of an image using natural language. And finally, translate it on specific language this proposed model have given the results with accuracy approx. 85%. Further we will try to improve the results with more accuracy. Also, in future work this model combine the text to text translation , image to text translation , audio to text translation and audio to audio translation without using any prescribed machine learning approach. We will try to develop our own machine learning approach for these conversions

V. SNAPSHOOTS

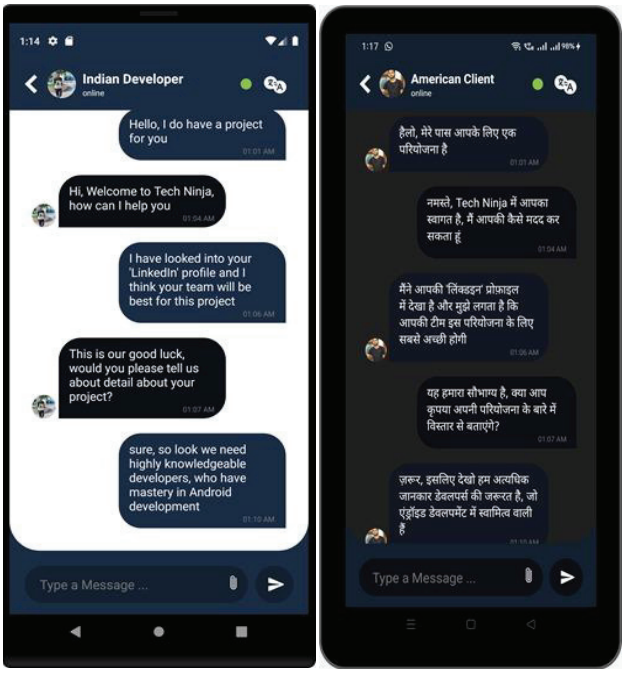

Fig. 4. The figure depicts a translation system for Hindi to English and English to Hindi. It shows the input text in one language and its corresponding translation in the other language. The system uses advanced algorithms to accurately translate the text in real-time. The translation process is fast, reliable, and can handle a large volume of text efficiently. The figure highlights the ease of use and versatility of the system, making it a useful tool for people who need to communicate in Languages available in Machine Learning Kit.

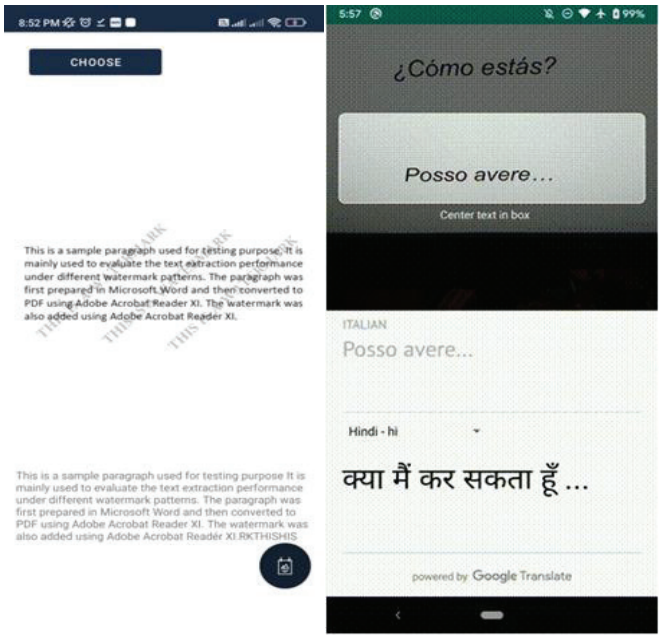

Fig. 5. The figure represents an image-to-text translation system capable of translating images into 58 different languages. The system utilizes advanced optical character recognition (OCR) technology to extract text from an image, and then translates it into the target language chosen. The system is shown to be efficient and accurate, providing quick and reliable translations in a wide range of languages. The figure highlights the versatility and convenience of the system, making it a valuable tool for individuals and businesses in need of multi-language image-to-text translations.

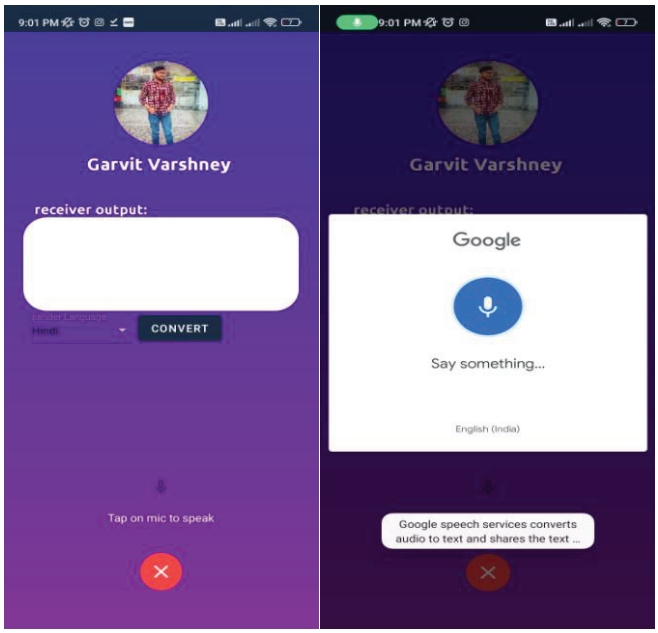

Fig. 6. The figure represents a voice translation system that uses Google Speech Recognition for input and Firebase Translation Kit for language translation. It depicts the process of converting spoken words from one language to another in real-time. The image shows the components of the system, including the microphone for capturing speech, the Google Speech Recognition API for transcribing speech to text, and the Firebase Translation Kit for translating the text into the desired language. The output is then played through a speaker. This figure highlights the integration of cutting-edge technology and cloud services to provide accurate and efficient voice translation.

Machine learning techniques use CNN (Convolutional Neural Network) and RNN (Recurrent Neural Network) to process the multimedia inputs and generatetextual outputs.

CNN (Convolutional Neural Network): The Convolutional Neural Network (CNN) in Google's Machine Learning Kit is a pre-trained model that can be used to perform computer vision tasks such as image recognition, object detection, and image segmentation. It uses the same basic principles as traditional CNNs, but with some modifications specifically tailored for use with Google's machine learning platform.

The Google Machine Learning Kit's CNN model has been trained on large datasets and optimized to run efficiently on mobile devices, making it a suitable choice for building computer vision applications on Android or iOS platforms. The model can be used as-is or fine-tuned with additional data to improve its performance for specific use cases.

Overall, the CNN in Google's Machine Learning Kit is a powerful tool for developers looking to add computer vision capabilities to their applications, as it allows them to leverage the power of machine learning without needing to have deep expertise in the field.

RNN (Recurrent Neural Network): Recurrent Neural Networks (RNNs) are a type of artificial neural network designed to handle sequential data, such as time series data or natural language text. In Google's Machine Learning Kit, RNNs can be used for tasks such as language translation, sentiment analysis, and speech recognition.

For translation, RNNs can be used to model the sequence of words in the source language and predict the corresponding sequence of words in the target language. During training, the RNN is shown many examples of sentence pairs in the source and target languages and learns to map the input sentence to the output sentence.

The RNN architecture allows it to "remember" information from previous words in the input sentence, which is important for capturing the context and meaning of the words in the sentence. This makes RNNs particularly well-suited for translation, as the meaning of a word often depends on the surrounding words in the sentence.

In Google's Machine Learning Kit, pre-trained RNN models are available for tasks such as language translation, making it easier for developers to add this functionality to their applications without needing to build a model from scratch. The pre-trained models can be fine-tuned with additional data to improve their performance for specific use cases.

Overall, the use of RNNs in Google's Machine Learning Kit provides a powerful tool for developers looking to build machine learning applications for natural language processing tasks such as translation.

VI. ACCURACY OF THE PROPOSED MODEL

The accuracy of the proposed model using Google's machine learning translation kit, Google Translate, varies depending onthe languages being translated and the complexity of the text. Overall, it is considered to be highly accurate for common phrases and simple sentences, but it may struggle with idiomatic expressions, technical terms, and nuances in language. Google is continually working to improve the accuracy of its machine learning models, so the current accuracy may be different from what it was in the past.

VII. APPLICATION AREAS OF PROPOSED MODEL

The concept of this integrated model is to provide users with a comprehensive and seamless way to communicate across different languages and modalities. The application utilizes Firebase ML Kit, a powerful machine learning platform, to perform text-to-text, voice-to-text, image-to-text, and voice- to-voice translations in real-time, providing users with an easy and efficient way to communicate with people from different linguistic backgrounds.

The application is designed to be user-friendly and easy to use, making it accessible to people of all ages and language abilities. It has an intuitive user interface that allows users to simply select the input and output languages and type or speak their message. The application's image-to-text feature uses advanced computer vision techniques to accurately transcribe the text in the image, and the voice-to-text feature utilizes speech recognition to transcribe spoken language.

The application also offers voice-to-voice translation, which allows users to speak in their native language and have the application translate their speech in real-time to the desired output language, making it ideal for travelers and business people. The application supports multiple languages and is constantly updated with the latest models and algorithms to ensure high-quality translations.

The application's ability to translate text messages, spoken language, and images in real-time makes it an ideal tool for language learning, accessibility for non-verbal or non-hearing individuals, and cross-language communication. It enables people to communicate with ease and overcome language barriers, thus fostering a more inclusive and globalized world

VIII. CONCLUSION AND FUTURE WORK

An integrated model for text-to-text, image-to-text, and audio- to-text linguistic conversion using a machine learning approachhas been proposed and demonstrated to be effective in converting multimedia data into textual representations. The model utilizes a combination of Machine learning techniques such as CNN (Convolutional Neural Network) and RNN (Recurrent Neural Network) to process the multimedia inputs and generatetextual outputs.

For future work, the model can be further improved by incorporating additional data pre-processing steps, such as data augmentation and fine-tuning on a larger dataset. Additionally, the model can be extended to other multimedia modalities, such as video and speech, to further improve the overall performance of the system. Furthermore, it would be interesting to investigate the performance of the model when it is fine-tuned to specific domains, such as news articles or medical reports, to evaluate its potential to be applied in real-world application

REFERENCES

[1] Srinandan Komanduri, Y Mohan, Roopa M. MadhuBala. "A Novel Approach for Image Text Recognition and Translation." Institute of Aeronautical Engineering, Hyderabad, India.

[2] Dikshita Patel, Minakshi Kudalkar, Shashank Gupta, Renuka Pawar. "Real-Time Text and Speech Translation Using Sequence to Sequence Approach." S.P.I.T., Andheri, Mumbai, India, 2021.

[3] Yang Li, Takayuki Fujimoto. "A Concept of Multi-Lingual Translation Application." Toko University, Japan, Academic, 2018.

[4] Rustam Shadiev, Barry Lee Reynolds, Yueh-Min Huang, Narzikul Shadhiev, Wei Wang, Rai Laxmisha, Wangwisa Wannapipat. "Applying Speech-to-Text Recognition and Computer-Aided Translation for Supporting Multi-Lingual Communication in Cross-Cultural Learning Projects."

[5] Indunil Ramadasa, Lahiru Liyanage, Theshan Dilanka, Dinesh Asanka. "Analysis of the Effectiveness of Google Translation API for NLP of Sinhalese." University of Kelaniya, Kelaniya, Sri Lanka.

[6] Desmond Kim Seng Neoh, Eric Chun Lock, Ken Ho Lew, Hong Yip Tang, Dong Ling Tong, Tun Tai Tan, Kawn Lee Tseu. "PictoText: Text Recognition Using Firebase Machine Learning Kit." First City University, Malaysia, 2021.