Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606

ICASET
2024

Research Article
Collection: Intelligent Computing of Applied Sciences and Emerging Trends

Audio-to-Text Urdu Chatbot using Deep Learning Algorithms RNN and wav2vec2

Areeba Khalid[1], Malik Daler Ali Awan[1], Nadeem Iqbal Kajla[2*], Amnah Firdous[3], Hafiz Muhammad Sanaullah Badar[2], and Malik Muhammad Saad Missen[1]

[1]The Islamia University of Bahawalpur, Bahawalpur, 63100, Pakistan.
[2]MNS University of Agriculture, Multan, 60000, Pakistan.
[3]The Government Sadiq College Women University Bahawalpur, 63100, Pakistan.
[*]Corresponding Author: Nadeem Iqbal Kajla. Email: Nadeem.iqbal@mnsuam.edu.pk

Academic Editor: Salman Qadri **Published:** April 01, 2024

Abstract: Advancement in technology limited the distances via communication. People globally exchange thoughts in different languages using many ways like text, audio, pictures, and videos to express their ideas. Among many languages Urdu language has more than 100 million people around the world. It is necessities the development of smart applications to facilitate Urdu language users that can communicate via audio instead of only text. A conversation bot system enables individuals and computers to communicate using natural language. Numerous Chabot's have been developed in English, German, Korean, Spanish, and Chinese languages. Because of the significant language barrier, those who do not speak English, German, Korea, or Spanish well cannot use these chatbots. In this research work we developed a smart chatbot system that can take voice as input for Urdu language using RNN a deep learning model. The proposed system is developed using two datasets UQuaD and custom dataset. A pretrained model is used to convert Urdu audio to text named as "wav2vec2-large-xls-r-300m-Urdu". The proposed system on UQuaD and custom achieved an accuracy of 68.30% on the UQuaD dataset and 89.6% on the custom dataset.

Keywords: Natural Language process (nlp); Urdu question answer dataset (Uquad); Wav2vec; Automated Speech Recognition(asr).

1. Introduction

The modernity and revolution in technology reshape the landscapes. The priorities of the world are evolving to align with trends, innovation and digitization in everyday life. In the past, countries were connected primarily for economic and commercial purposes. However today, these connections play crucial role in technology revaluation to facilitate the human being all around the world. The inexorable forces of digital globalization is steadily propelling as a global wave. The ultimate purpose is the advancement in education, research, technology and betterment of human life standard by collaboration of innovative ideas, research findings and hooking varieties of scientific knowledge.

Totally autonomous, intelligent artificial intelligence systems that can converse with us in a way that is really "human-like" in terms of technological development may yet be several decades away. But, in many ways, this future scenario is approaching gradually and unexpectedly swiftly because to the continued development of what is known as automated speech recognition technology. It also appears to promise some genuinely helpful improvements in user experience for a variety of applications, at least so far.

There are numerous chat platforms accessible worldwide that allow for natural language conversation. Human-human dialog systems and human-computer dialog systems are the two basic categories into which these chat systems can be roughly divided. The Human-Human Interaction System acts as a go-between for two people. and does not use machine learning. The Human-Human Dialog System does not rely on artificial intelligence in robots. Skype and Whats App are the most widely used Human-Human Dialog chat systems worldwide. However, the Human-Computer Dialog System, another conversation option, is known as a chat-bot. [1].

A chat bot is essentially a computer program that pretends to have a conversation with humans using natural language. This technology can communicate with people on any platform, including mobile, the web, and desktop applications [12]. When conversing with a human, a chatbot pretends to be a person.. The chat-bot converses and interacts with thousands or even millions of people at once while a human only converses and interacts with one person at a time.

Modern era problems require modern solutions, We all know that Language is a basic thing to start anything when we are unknown of any language it is a big barrier to cross through, My research motivation is to make a model or translator that will ease and finish the language barrier for our Pakistani community, Many kinds of research have been done on different language translators, and no doubt they have done brilliant work for their languages like English to Turkish, English to English Translators, I'll work on the translator for Urdu to Urdu so that it eases our society because the majority of us speak Urdu as our national and common language, I hope my work will be helpful to many of us. We decide to process Urdu language audio into Urdu language text to design a conversation bot for Urdu language.

The study's primary goal is to create a system that can process audio in the Urdu language to Urdu language text.

1.1 Contributions

- Developed Urdu audio to the text processing system.
- We have developed customized dataset for the conversation bot.
- Comparison of different datasets.
- Comparison of base model dataset to custom dataset.

2. Literature Review

2.1 Urdu Speech Recognition

There is a description of continuous Urdu speech recognition across a vast vocabulary in [5]. The development of acoustic and linguistic models for Urdu, English, and other language pairings using the CMU Sphinx Tool Set. Pre-ordering sentences is how improved machine translation is presented[6]. A multilingual collection of documents is suggest to handle using n-gram-based language modeling.

2.2 Review of different papers

In this article [7] The authors suggested a chatbot called dinus intelligent assistance for use in university admissions. because discussions at their university are conducted exclusively in bahasa indonesia, DINA chose Bahasa Indonesia as her primary language of understanding. They concentrate on the university visitor book, which features questions and responses about the UDINUS admissions process. By entering questions, this system is tested. The author evaluated it with ten randomly selected sample questions from 166 intentions. It answered eight out of ten test questions accurately. The researcher's proposed The results of DINA Chatbot suggest that the new approach will help users find the data they need.This significant fault in present chat-bots, Using the suggested technique and the paper's constraints, the problem, which is that they are unable to understand the connection between entities and characteristics, can be resolved.

This article [8] examined the potential of an English- and Urdu-supporting chat bot system for the mehran university of engineering technology jamshoro examination department. English voice-based chatbots, Urdu typing-based chatbots, and English typing-based chatbots are the three chat-bot interfaces that have been developed. According to a poll of engineering graduates, the ideal layout is one that requires English writing since it is the most user-friendly and has a general grasp of queries.

In this article [9] the authors recommend a deep neural network-based Urdu language translator. Three pipeline module automatics speech recognition , text-to-text translation, and text-to-speech conversion make up the suggested model of the urdu-to-english speech translator. Deep neural networks make up the ASR module in the proposed pipeline, which is simple than standard ASR, which calls for labor-intensive manual work such as feature extraction and the use of resources like phoneme dictionaries. The suggested system can deliver correct real time outcomes while demonstrating resilience to loud environments and voice tone fluctuations after being trained on more than 8 GPUs using 12–13 hours of audio data. Online apis for Text-to-Text and Text-to-Speech conversion have been successfully combined to deliver accurate results promptly. The proposed model put out the idea that by boosting computation power and linguistic corpus size, effective and precise results might be obtained in real-time.

The writers of this paper [10] developed an Urdu Text-to-Speech System Letter-to-Sound Conversion. The architecture of the Natural Language Processing component of an Urdu Text-to-Speech system is briefly cover in this work. It provides information on the consonantal and vocalic systems of Urdu as well as Urdu letters. Because Urdu behaves predictably, the phonemic forms can be deduced from the textual input. The letter-to-sound rules specify this mapping and are hence necessary for building Urdu TTS.

The authors of this article [11] suggested creating a system for automatic audio-to-text conversion. The purpose of the paper is to identify the gender of the user to increase the human-likeness of the chatbot's response. The automated conversion of auditory impulses into text is the subject of the study.The created software item is a chatbot that transforms the audio message received into text and sends it back as a sms, is the paper's intended outcome. This project has produced a system dubbed Harry Bot for automatic audio-to-text conversion with enhanced performance for analog systems. Using self-learning and continuous improvement techniques, new artificial intelligence and machine learning technologies that are simple to use for the transformation's outcomes RNN and other kinds of neural networks are emerging so quickly that it is getting harder to keep up with the latest, most intriguing, and most advanced models for tackling the most challenging and tough jobs. These sequential techniques for training neural networks are not just applicable to machine translation. Basic illustrations include models that can describe a visual verbally, identify voices, and carry on a dialogue. They believe that the advancement of RNNs will result in the introduction of intelligent personal assistants that can recognize the owner's speech and accurately understand the work. Machine translation now uses RNNs most commonly, and experts predict that this technology will soon be improved. The experiment's findings indicate that the Keras library-based model the more effective regarding the present training data set.

3. Materials and Methods

3.1 Dataset

In the world of technology, data is power, and reliable data empowers the system. Data is a basic building block of every machine learning and deep learning-based intelligent system. In our research task, we explored various sources of information, like Kaggle, GitHub, google, websites, and many more. In this research work, to achieve our objective, We choose to employ a pre-trained model that was develop using a well-used voice dataset. The audio dataset known as Common Voice is made up of a single MP3 and matching text file. The dataset contains 9,283 hours that have been collected. Age, sex, and accent-related demographic metadata are also included in the collection. 7,335 validated hours in 60 language families make up the dataset.

The second dataset that is publicly available [2] named "UQuAD-Urdu-Question-Answer-Dataset" is available on GitHub for researchers to develop more accurate systems. The details of a dataset contain there are 27 different Urdu sentences that have been collected from a variety of sources, including news articles, YouTube videos, and Urdu Wikipedia. Every paragraph that was chosen has between three and seven questions with a range of one to three responses for each. Most of the terms in the data are from Urdu, although there are also some words from English. I recorded also all these questions text into audio from a mobile phone recording in mp3 format.

We prepared another dataset for testing and training purposes. that dataset contains questions and answers. that dataset consists of eight numbers of classes with different questions. The majority of the phrases in the data are from Urdu; however there are a few English words as well. We also recorded all these questions text into audio from a mobile phone recording in mp3 format.

3.2 Statistics of Dataset

Table 1. Dataset Info

Statistical Information	**Numbers**
Amount of questions overall	139
Number of paragraphs throughout	27
Word count for the entire paragraph	555
Word count for all sections combined	1631
Whole number of words in all questions	1237
Number of individual words in the entire paragraph	395

3.3. Proposed methodology

To achieve our research objectives, literature provided us vision to design our methodology. In our best knowledge yet there exists not enough referential work for Urdu audio to Urdu text. In the figure 3.1 that is given , we presented a complete overview of our proposed system.

This model is an improved version of facebook/wav2vec2-xls-r-300m for the common voice dataset.Because it is read speech, CommonVoice is a simpler task than BABEL, but because we employ a smaller labelled data arrangement, a new problem is presented. We specifically adopt Riviere et al. (2020)'s train/dev/test divides, which match to a few-shot scenario where only one hour of training data is provided per language. A novel cross-lingual speech representation model called XLS-R scales the number of languages, the volume of training data, and the size of the model. The training dataset comprises 128 languages in 436K hours of recorded speech audio, which is an order of magnitude more than previous efforts.[3].

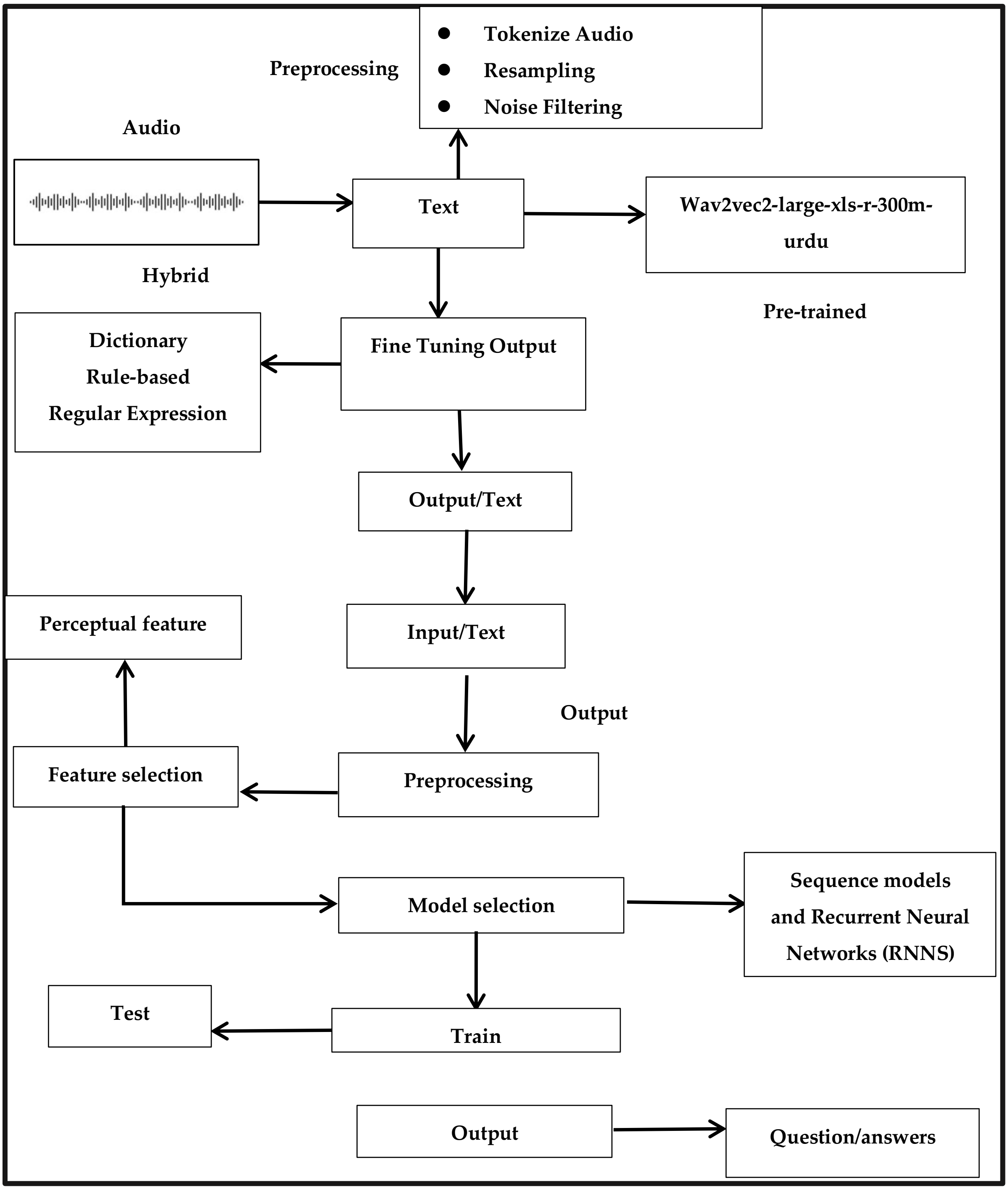

Figure 1. A Comprehensive Overview of Designed System

3.4 Models

3.4.1.wav2vec2-large-xls-r-300m-Urdu

This model is an improved version of facebook/wav2vec2-xls-r-300m for the common voice dataset.Because it is read speech, CommonVoice is a simpler task than BABEL, but because we employ a smaller labelled data arrangement, a new problem is presented. We specifically adopt Riviere et al. (2020)'s train/dev/test divides, which match to a few-shot scenario where only one hour of training data is provided per language. A novel cross-lingual speech representation model called XLS-R scales the number of languages, the volume of training data, and the size of the model. The training dataset comprises 128 languages in 436K hours of recorded speech audio, which is an order of magnitude more than previous efforts.[3].

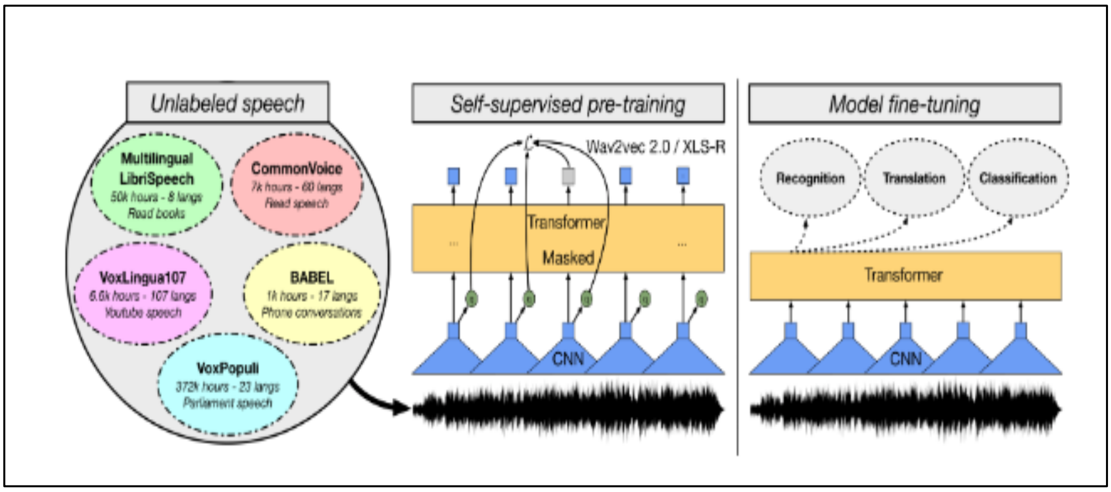

Figure 2. Self-supervised cross-lingual representation learning [13]

3.4.2. Recurrent neural networks (RNNS)

Recurrent neural network is used to address the sequence problem (RNN). Recurring meaning occurring over a predetermined period of time. In this instance, the model will translate each new word into Urdu as it is received from the phrase. This process will be repeated for all remaining words, making the translation effort ongoing. Also, it will preserve the last translated word in memory so that we have the context for translating the next word.

3.5 Training Dataset

The training data refers to a portion of the dataset that is utilized to train the Uquad dataset common voice dataset. In our research work, we decided to use 70 % of the dataset for training purposes. In research tasks like audio classification and text classification, the training dataset is the collection of labeled/annotated instances. The simulation is prepared on a textual dataset that consists of many questions and answers related to general knowledge.

3.6 Testing Dataset

The unseen and split part of the data set examined the developed (trained) model is called the testing dataset. We used the 30% portion of the entire corpus for testing purposes. In the testing dataset, many questions are unseen for trained classifiers and are used to test the performance of the trained system. The model is tested on a textual dataset , that consists of many questions and answers related to general knowledge.

4. Experiments Results

To accomplish our research task, we decided to use python language and its various libraries because of its versatility. Python language highly preferable for scientific tasks. GoogleColab is an online platform that provides the cloud services for storage and computation.

4.1 Standard Performance Evaluating Parameters

To evaluate and report the performance of developed system, we rely on the standard parameters that we found in the literature. It can be observed that precision equation (1) , recall equation (2), equation (3) f1-

measure and equation (4) accuracy are general parameters to report the overall performance of the system [4].

$$\Pr ecions = \frac{TP}{(TP+FP)} \quad (1)$$

$$\operatorname{Re} call = \frac{TP}{(TP+FN)} \quad (2)$$

$$F1 - Measure = 2 * \frac{\Pr ecision * \operatorname{Re} call}{\Pr ecision + \operatorname{Re} call} \quad (3)$$

$$Accuracy = \frac{TN+TP}{TN+FP+TP+FN} \quad (4)$$

4.2. Result Discussion

This section covers the chat bot's outcomes. the outcome of the training and testing data sets and the dataset's correctness.

4.2.1 Training Accuracy

We achieved 99.28% accuracy in the training dataset of UQuAD---Urdu-Question-Answer-Dataset.

Table 2. Training Accuracy of UQuAD

Parameters	Values
accuracy	0.993
Loss	0.057

Table 3. Training Accuracy of custom Dataset

Parameter	Values
Accuracy	1.000
Loss	0.001

4.2.2 Testing Accuracy

We achieve 68.35% accuracy in testing dataset of UQuAD.

Table 4. Testing Accuracy of UQuAD

Parameter	Values
Loss:	1.626
Accuracy:	0.683

We achieve **89.6%** accuracy in testing dataset.

Table 5. Testing Accuracy of custom Dataset

Parameter	Values
Accuracy	0.896
Loss	0.454

4.2.3Testing Result Graph of custom data

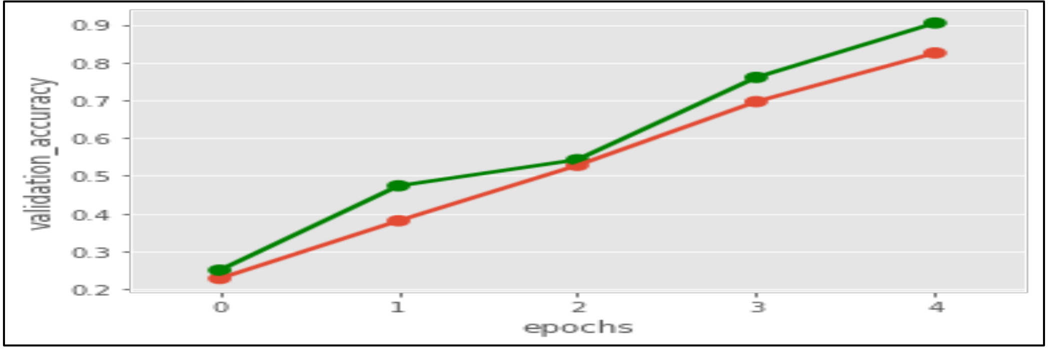

Figure 3. Validation Accuracy of Conversation Chatbot System

5. Discussion

Social media, meanwhile, expanded the scope of communication. A significant portion of people has switched to using online social networks for communication, including sharing, questioning, praising, and expressing opinions. Technology opened huge space for many kinds of data. such as text, photos, videos, and audio.In today's world, audio is use for communication. In this proposed study, we decide to process Urdu language audio into Urdu language text to design a conversation bot for Urdu language. This model is only specific to Urdu language. Based on greeting, shopping, payment, and delivering general knowledge, for a small dataset. The proposed system on UQuaD and custom dataset showed 68.30% and 89.6% accuracy respectivel**y.**

The designed system demonstrated up to 89.6% accuracy; going forward, we will use a large number of datasets for testing in order to enhance the performance of the testing system. Further , we have plan to design our model to convert audio to text and vice versa. A well-structured and huge dataset set can be developed.Text to audio model can also be developed .

6. Limitations

- Unavailability of data
- Lack of processing resource
- Based on greeting, shopping, payment, and delivering general knowledge, for a small dataset.
- This model is only specific to urdu language.
- This model is based on only audio-to-text.

7. Conclusions

Language translation is one of the interesting and challenging tasks of Natural Language Processing. Especially in case of lacking resources, translating audio data to text is exhaustive tasks. It demands highly skilled and language experts to develop such system. In our research journey we explored the Audio and text data of the Urdu language. We faced lot of challenges while converting Urdu language audio to text. As, the purpose of our research work is to provide a base and take initiative to develop an audio-to-audio chatbot for Urdu language. It is pertinent to mention that lack of dataset is yet major challenge to develop such proposed system. While the testing on same dataset i.e., communicating with developed chatbot showed 100% accurate results.

8. Future Work

- The designed system demonstrated up to 89.6% accuracy going forward, we will use a large number of datasets for testing in order to enhance the performance of the testing system.
- Further we, have plan to design our model to convert audio to text and vice versa.
- A well structured and huge dataset set can be developed.
- Text to audio model can also be developed.

References

1. Hettige, B., & Karunananda, A. S. (2006). First Sinhala chatbot in action. Proceedings of the 3rd Annual Sessions of Sri Lanka Association for Artificial Intelligence (SLAAI), University of Moratuwa, 13.
2. https://github.com/ahsanfarooqui/UQuAD---Urdu-Question-Answer-Dataset
3. Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., ... & Auli, M. (2021). XLS-R: Self-supervised cross-lingual speech representation learning at scale. arXiv preprint arXiv:2111.09296.
4. Malik. D.A.A, Malik S.S, and Hussnain M. "Multiclass Event Classification from Text", Scientific Programming, 2021.
5. Ali, S. A., Khan, S., Perveen, H., Muzzamil, R., Malik, M., & Khalid, F. (2017). Urdu language translator using deep neural network. Indian J. Sci. Technol, 10(40), 1-7.
6. Visweswariah, K., Rajkumar, R., Gandhe, A., Ramanathan, A., & Navrátil, J. (2011, July). A word reordering model for improved machine translation. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (pp. 486-496).
7. Santoso, H. A., Winarsih, N. A. S., Mulyanto, E., Sukmana, S. E., Rustad, S., Rohman, M. S., ... & Firdausillah, F. (2018, September). Dinus Intelligent Assistance (DINA) chatbot for university admission services. In 2018 International Seminar on Application for Technology of Information and Communication (pp. 417-423). IEEE.
8. Arain, M., Memon, M. A., Bhatti, S., & Arain, M. (2019). Feasibility of chatbot for mehran UET examination department. Review of Information Engineering and Applications, 6(2), 17-28.
9. Ali, S. A., Khan, S., Perveen, H., Muzzamil, R., Malik, M., & Khalid, F. (2017). Urdu language translator using deep neural network. Indian J. Sci. Technol, 10(40), 1-7.
10. Sarmad, H. (2004). Letter-to-sound conversion for Urdu text-to-speech system. In Workshop on Computational Approaches to Arabic Script (pp. 74-79).
11. Basystiuk, O., Shakhovska, N., Bilynska, V., Syvokon, O., Shamuratov, O., & Kuchkovskiy, V. (2021). The Developing of the System for Automatic Audio to Text Conversion. In IT&AS (pp. 1-8).
12. Orin, T. D. (2017). Implementation of a Bangla chatbot (Doctoral dissertation, BRAC University)
13. Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., ... & Auli, M. (2021). XLS-R: Self-supervised cross-lingual speech representation learning at scale. arXiv preprint arXiv:2111.09296.
14. Kajla, N. I., Missen, M. M. S., Luqman, M. M., & Coustaty, M. (2021). Graph neural networks using local descriptions in attributed graphs: an application to symbol recognition and hand written character recognition. IEEE Access, 9, 99103-99111.
15. Kajla, N. I., Missen, M. M. S., Luqman, M. M., Coustaty, M., Mehmood, A., & Choi, G. S. (2020). Additive angular margin loss in deep graph neural network classifier for learning graph edit distance. IEEE Access, 8, 201752-2