Neural Networks based Dialogue System for Customer Support

By

Amna Noor

000000276430

Supervisor

Dr. Rabia Irfan

Department of Computing

School of Electrical Engineering and Computer Sciences (SEECS)
National University of Sciences and Technology (NUST)
Islamabad, Pakistan

(July 2022)

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis entitled "Neural Networks based Dialogue System for Customer Support" written by AMNA NOOR, (Registration No 00000276430), of SEECS has been vetted by the undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial fulfillment for award of MS/M Phil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature: ______________________

Name of Advisor: Dr. Rabia Irfan

Date: 26-Jul-2022

HoD/Associate Dean:_______________________

Date: ___________________________________

Signature (Dean/Principal): __________________

Date: ___________________________________

Approval

It is certified that the contents and form of the thesis entitled "Neural Networks based Dialogue System for Customer Support" submitted by AMNA NOOR have been found satisfactory for the requirement of the degree

Advisor : Dr. Rabia Irfan

Signature: Irfan

Date: 26-Jul-2022

Committee Member 1:Dr. Muhammad Ali Tahir

Signature: Ali

Date: 27-Jul-2022

Committee Member 2:Dr. Muhammad Khuram Shahzad

Signature: MKhurram.

Date: 27-Jul-2022

Committee Member 3:Dr. Muhammad Imran Malik

Signature: ImranMalik

Date: 28-Jul-2022

Dedication

This thesis is dedicated to my beloved parents who supported me and believed in me.

Certificate of Originality

I hereby declare that this submission titled "Neural Networks based Dialogue System for Customer Support" is my own work. To the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST SEECS or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEECS or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project’s design and conception or in style, presentation and linguistics, which has been acknowledged. I also verified the originality of contents through plagiarism software.

Student Name: AMNA NOOR

Student Signature: Amna Noor ______________

Acknowlegement

I would like to express my gratitude to my supervisor Dr.Rabia Ifran for allowing me to conduct research and for her invaluable advice during this process. It was a great privilege to complete research under her guidance and to my committee members (Dr. Ali Tahir, Dr. Muhammad Khuram Shahzad and Dr. Imran Malik) for their constant support and guidance.

I want to thank my parents and siblings from the bottom of my heart for supporting me and providing me with the resources I need to achieve my goals.

I would like to say my special thanks to my friends Ammarah Irum, Ezza Shaukat, Almas Shabbir and Rabia Ghafoor who supported and provided me with a wealth of happy memories.

Contents

List of Abbreviations

CA	Conversational Agent
ML	Machine Learning
DL	Deep Learning
NLP	Natural Language Processing
RNN	Recurrent Neural Network
LSTM	Long Short Term Memory

List of Tables

List of Figures

Abstract

Customer service is one of the most important components of online services. However, as natural language processing methods are on the rise, the market is looking at automated conversational models based solutions to deliver high-quality services to a user base that is always expanding. The Deep Learning based conversational agent (CA) is a challenging Natural Language Processing (NLP) task in a language with poor resources like Urdu, as well as the scalability and generalisation capacities of the neural conversational models that were lacking in previously employed manually annotated and rule-based systems. Although conversational agents have been developed for other languages, recent state-of-the-art neural network-based techniques have not yet been investigated for conversational agents in Urdu. We have compiled a dataset of about 12000 question-answer pairs and implemented two basic deep learning architectures: Long Short Term Memory (LSTM) with and without Attention mechanism. These have been used in our work to examine the powerful deep learning techniques for an Urdu conversational agent in the customer support area. In this study, we developed an Urdu conversational agent model based on Transformer that entirely follows the attention mechanism. The suggested and baseline methodologies were implemented on Urdu and English customer care datasets from Amazon, where the suggested model outperformed all other deep learning techniques when the results of these techniques were examined. The Transformer attained a BLEU score of 38.13, 40.2, and 31.7 on the small, large, and English data sets, respectively, outperforming the basic deep learning models.

Keywords: *Urdu Conversational Agents, Deep Learning, NLP, Attention.*

Chapter 1

Introduction

In this new era of artificial intelligence (AI), data-driven solutions are becoming indispensable for our daily lives. As there is a significant rise in the accessibility of textual data from various sources, the scientific community has shown a great deal of interest in a variety of new and more challenging topics in the field of Natural Language Processing and Computer Vision. This chapter gives an overview of the study that has been conducted on Conversational Agents (CA).

The motivation for conducting this research is described in section 1.1. Afterwards, The applications of conversational agents and a brief classification of chatbots is then stated in section 1.2 and 1.3, respectively. The problem statement will be detailed in section 1.4. The research objectives, novelty and contributions will be explained in sections 1.5 and 1.6, respectively. We shall conclude with the thesis outline in the section 1.7.

1.1 Motivation

One of the most important part of the consumer experience for digital services is customer assistance. With the advancements in Natural Language Processing (NLP) and Artificial Intelligence (AI), the companies are turning to autonomous conversational services. The conversational agents, or chatbots, are one such current NLP research subject. Conversational agent, often known as a chatbot, is an AI based language generation system that can conduct a dialogue with a user through the use of a question-and-answer mechanism. Natural Language Processing (NLP) demands the computational representation of a language's complex semantic and syntactic links.

The languages like Urdu, Arabic, Turkish, Hindi, and Persian, that are resource-scarce but linguistically rich, have been used the least for NLP applications in contrast to the resource-rich languages, since it is difficult to adapt to the word-level complicated linguistic structure of certain languages [1]. Since the inception of NLP, the English language has gotten a lot of attention from researchers all over the world. Because it is computationally affordable to perform testing in a resource-rich language. Most large-scale research on each component of NLP has been conducted in English for the past two decades. Urdu's unique features, its complicated morphology, and the fact that there aren't many linguistic resources have all made it hard to do research on Urdu chatbots [2].

The majority of Urdu chatbot research is done using either handwritten rules or information retrieval (IR) approaches. The major issue with IR or rule-based algorithms is that they lack human-like responses. Machine learning, on the other hand, is based on extracting patterns from data. Although machine learning algorithms can detect the direct relationship between features, they are not scalable to big datasets and can only learn a few selected features.

In today's world, practically every company, no matter how big or small, has some kind of customer service department. The customer has come to anticipate that they will be able to obtain information and a resolution to their issue at any time of the day or night, seven days a week. A variety of customer service strategies are utilised by businesses in order to develop close relationships with their clients and enterprises. Due to the fact that data is continuously expanding, we also require solutions that are easily scalable.

Deep Learning models have shown over the course of the previous few years that they are capable of showing the highest outcomes, particularly in the field of conversational agents, as customer service is the most pressing issue in the modern world. One of the widely used models for sequential problems is called a recurrent neural network (RNN), and the reason for this widespread use is that the RNN excels at managing sequential data such as text, audio, and forecast. In this study, we also employed LSTM and attention-based LSTM models. Long short-term memory networks (LSTMs) are utilised here rather than a standard recurrent neural network or one of its variants since they are able to solve the issue of vanishing gradients and provide superior mapping even for lengthier questions and responses. The encoder and decoder now include attention mechanisms to improve the performance of the seq2seq model. In order to address the sequence-related conundrum, the transformer model has since emerged as a high-performance model with numerous attention mechanisms. In addition to achieving state-of-the-art performance for sequence transduction, our model shortens training time as compared to RNN-based models.

1.2 Applications of Conversational Agents

The Internet is now the incubator from which a limitless quantity of information emerges, which is to the benefit of people all over the world. Users are encouraged to discuss any subject, product, movie, dish, advertisement, or service that they have used, in order to provide others with the opportunity to learn about products or book an appointment. The majority of the applications for conversational agents, also known as chatbots, are tied to various industries and corporations monitoring user experience through their priceless input [3]. The following are a few CA applications:

1.2.1 Customer Support

Multinational companies and organisations place a high premium on improving their services in order to satisfy clients. Customer input obtained from websites, contact centres, surveys, chat rooms, and emails is used to improve customer service in general. Customer service management presents a variety of difficulties that make it difficult even for top-tier businesses to resolve.

1.2.2 Healthcare

The use of chatbots in the healthcare industry has been shown to be quite beneficial, particularly in recent years when hospitals and other types of institutions have been obliged to reduce the number of lines and crowds that form outside their facilities. Here are some use cases. Patients can use chatbots to schedule their appointments without leaving the comfort of their own homes. In addition, patients have the option of rescheduling their visits with the assistance of a chatbot, which helps to cut down on the number of in-person interactions that are necessary. A significant volume of tickets can be avoided thanks to chatbots' ability to analyse the inquiries from your patients and offer prompt, precise answers.

1.2.3 Education

Universities and educational institutions may find it difficult to support more than a thousand pupils. Nevertheless, these educational institutions can certainly handle the high frequency of enquiries by implementing AI-enabled chatbots. Here are a few education-related chatbot applications. With regards to the library, facilities, fees, technology, and device management, chatbots can help students with a variety of problems. To make sure that all students are aware of any updates or policies that are being implemented, chatbots can be utilised to broadcast updates.

1.2.4 Entertainment

Chatbot applications are now essential in the entertainment and media sectors, helping with everything from providing quick summaries of the current developments to helping clients manage their subscriptions. News outlets and publishers have started using chatbots to improve the reader experience. The chatbot can be used by customers to find content that intrigues them.

1.3 Classification of Chatbots

Task-oriented chatbots and non-task-oriented chatbots are the two basic groups that make up the chatbot classification system [4]. The purpose of a task-oriented chatbot is to facilitate the completion of a certain task. These chatbots are designed to hold brief discussions, typically occurring within a restricted sphere. The primary objective of task-oriented chatbots is to provide the user with assistance in completing a certain activity [5]. They are intended for dealing with particular circumstances, such as booking a hotel or flight; reserving lodgings; making an order for goods; organising a schedule for an event; or assisting users in gaining access to certain information, among other things. Voice-based task-oriented chatbots/conversational agents that try to provide a response to the task they are given include personal assistants like Cortana, Alexa, and Siri. Chatbots focused on certain tasks perform particularly well in constrained environments. There is no way to test these chatbots' trivial skills because they are not knowledgeable about general topics. Instead, they are goal-oriented chatbots that are centred on assisting you in accomplishing a particular objective that you have set for yourself.

Systems that are constructed for lengthy conversations are known as non-task-oriented chatbots. These systems are programmed to imitate the informal conversational or chat that is characteristic of human-to-human interaction rather than concentrate on a particular task such as booking plane tickets. These systems frequently provide some form of entertainment. Non-task-oriented chatbots, as opposed to task-oriented chatbots, have the ability to imitate a dialogue and provide the appearance of engaging in idle small talk for the goal of enjoyment in open domains. The creation of a task-oriented or non-task-oriented chatbot can be accomplished through the application of a variety of distinct methodologies. There is potential for crossover between these two primary classifications of chatbots and these methodologies. These methods can essentially be broken down into three distinct types of approaches: rule-based, retrieval-based, and generative-based.

1.4 Problem Statement

In today's world, practically every company, no matter how big or small, has some kind of customer service department. The customer has come to anticipate that they will be able to obtain information and a resolution to their issue at any time of the day or night, seven days a week. A variety of customer service strategies are utilised by businesses in order to develop close relationships with their clients and enterprises. Due to the fact that data is continuously expanding, we also require solutions that are easily scalable.

It is also difficult for researchers to find substantial, publicly open, documented data in native language like Urdu for conversational agents in the form of question-answers since the Urdu language has a sophisticated lexical composition. Deep learning-based models in the realm of conversational agents or chatbots are currently underdeveloped due to a lack of annotated Urdu data. People nowadays prefer to communicate in their native tongue. As a result, it is critical to deploy chatbots in native languages. **A scalable and automated customer service based conversational agent for Urdu with sophisticated lexical composition of annotated Urdu data is an interesting problem.**

1.5 Research Objectives

In addition to the intricate morphological structure and one-of-a-kind characteristics of the Urdu language, the availability of an annotated Urdu corpus is one of the primary hurdles that researchers face when attempting to develop Urdu conversational agents. It is difficult for researchers on this subject to broaden the scope of their studies because there are not many large standard corpora that are freely available to the public and do not cost anything. Due to the fact that people generally prefer to communicate in their native languages, the application development of an Urdu chatbot has been hampered by this barrier. In addition, deep learning techniques have not yet been thoroughly investigated in the field of Urdu. The main objective of this research is to develop an end-to-end deep learning based conversational model for Urdu language. The basic architecture of the chatbot is described in Fig 1.1.

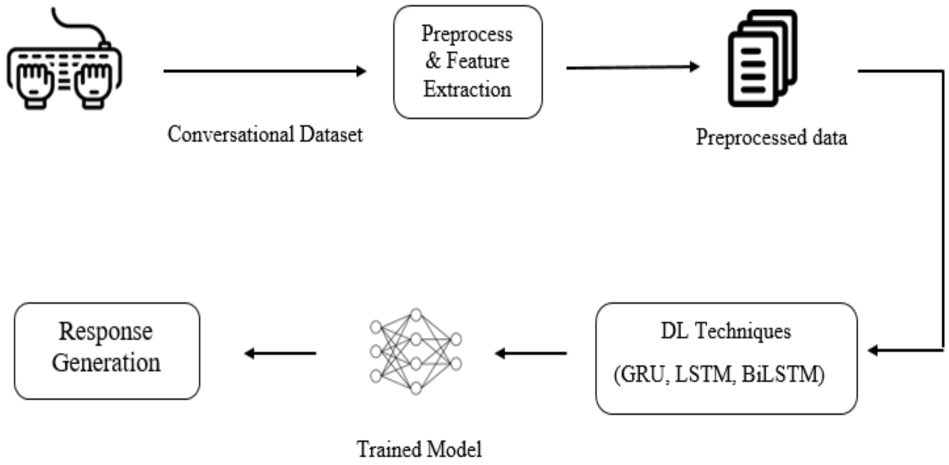


Figure 1.1: Chatbot Framework

1.6 Contributions

In a language with limited resources like Urdu, the Deep Learning-based conversational agent (CA) is a difficult Natural Language Processing (NLP) task. The scalability and generalisation capabilities of the neural conversational models were lacking in previously used manually annotated and rule-based systems. Despite the fact that conversational agents have been created for other languages, new cutting-edge neural network-based methodologies for conversational agents in Urdu have not yet been researched. We compiled a dataset of roughly 12000 question-answer pairs in the Urdu customer support domain. In this study, we developed a Transformer-based Urdu conversational agent model that entirely captures the attention process. The transformer model outperformed all other Deep Learning techniques when the outcomes of these techniques were analysed.

The key contributions of this research are filled in these gaps:

- A conversational dataset in Urdu language was compiled for customer support domain.
- For Urdu based conversational agents, a latest Transformer-based architecture is developed.
- The performance of proposed and baseline Deep Learning models is analyzed on variable sized data sets.
- On a variable sized data sets, the effectiveness of suggested and baseline Deep Learning models is examined.
- A comparison of the performance of the transformer with well-known Deep Learning models is performed.

1.7 Thesis Outline

The remaining chapters of the thesis are categorised as follows.

1.7.1 Literature Review

In this chapter, we will discuss the work that has been done in Urdu NLP up to this point. The categorization study carried out for Urdu conversational models utilising a variety of Machine Learning and Deep Learning approaches is extensively discussed here.

1.7.2 Background

A brief summary of popular Deep Learning approaches and techniques for conversational models are given in Chapter 3. Additionally, a few of the foundational ideas covered in the methodology chapter are described.

1.7.3 Research and Methodology

This chapter goes into thorough detail of the proposed model for Urdu chatbot. Additionally, it outlines the procedures for data collecting, data translation, and Deep Learning model building.

1.7.4 Results and Discussion

This chapter comprises of illustrating how our suggested model functions using various graphical representations. Exact results and their thorough discussion round out the chapter.

1.7.5 Conclusion and Future Work

A brief summary of this research project is provided in Chapter 6 along with potential tasks that could be completed in the future for additional research goals.

Chapter 2

Literature Review

This chapter provides in-depth details on conversational agents as well as related studies. It recalls the origin and evolution of conversational agents briefly. It also examines the research in terms of traditional and neural network methodologies. Lastly, we will conduct a critical examination of the work on conversational agents in the Urdu language.

2.1 Evolution of Chatbots

One of the key hurdles in Artificial Intelligence (AI), according to Turing, is giving machines the ability to converse with humans using normal language [6]. A German computer scientist, Joseph Weizenbaum, created a programme named ELIZA that is renowned as being one of the first conversational systems that was capable of partially passing the Turing Test [7]. The model of ELIZA interprets the data from the user and attempts to recognise key-terms with predefined responses and answer the question by rephrasing it. This created a sense that ELIZA comprehended the context of the question, but it lacked the structure for contextualising activities. It was designed to illustrate the shallowness of communication between humans and machines. Due to the fact that it is a rule-based framework, it is incapable of having intelligent dialogue with humans. Also, as described previously, it had standard replies to statements that did not conform to rules. Weizenbaum, was surprised by the people's response that believed in ELIZA's intelligence although he insisted it wasn't the case. Many chatbots, such as PARRY and A.L.I.C.E were inspired by this straightforward rule-based framework. Psychiatrist Kenneth Colby created PARRY in 1972 to design the dialogue framework of an individual that suffers with paranoid schizophrenia [8]. It was the first conversational model that passed the Turing test. Although PARRY is a rule-based conversational model, it is significantly more complex than ELIZA.

The Artificial Linguistic Internet Computer Entity, better known by its acronym A.L.I.C.E., is a well-known and widely used free chatbot that was created in 1995 by Dr. Richard Wallace. It was created in response to the novel ELIZA, that presently uses AIML, or Artificial Intelligence Mark-Up Language, to formulate responses to questions asked of [9]. It does this by using some template matching rules in order to provide responses to the user's inquiry. The design of a conversational agent consists of two clearly distinct pieces, which are referred to as the ”chatbot engine” and the ”language model,” respectively This chat robot has won multiple awards and is open source. It failed the Turing test as sometimes it was unable to respond to the normal queries of the user.

The British computer programmer Rollo Carpenter is responsible for creating the chatbot known as Jabberwacky [10]. Although it is considered to be the first computer programme that attempted the use of artificial intelligence, it has been indicated that the game does not use any fuzzy logic or neural networks, instead relying solely on heuristics-based methods. Its objective is to provide amusement in the form of a conversational bot that attempts to simulate normal human interaction. Rather than relying on rules, it works with the user's input by using contextual pattern recognition techniques. It has also learned to speak new languages.

With the surge in commercial chatbots, or personal assistants, as they are typically built into electronic devices such as mobile phones, smartwatches, etc. These assistants speak to the user by typing or by speaking to them. The advent of artificial intelligence has allowed for an expansion in the use of chatbots in various settings.

The creation of intelligent personal assistants, such as those offered by Amazon's Alexa, Apple's Siri, Google's Google Assistant, Microsoft's Cortana, and IBM's Watson, is one of the newest and most fascinating applications of artificial intelligence [11]. It is not possible to acquire information regarding the specifics of the deployment of any of these technologies because they are all commercial products. In these conversational agents you can complete a task, ask about the weather, or any general knowledge question, they also have improved capabilities such as image-based search., but natural language is shown to be rather difficult due to the numerous variances in human speech that occur at the regional, geographic, and localised levels. The detailed taxonomy of conversational agents is shown in Figure 2.1; a further literature review will be conducted in the following sections using this framework.

2.2 Conventional Techniques

Traditional methods are known as rule-based techniques because they generate replies based on predetermined rules. These rules have gotten increasingly intricate and complicated over time. When the conversation's domain is closed, i.e. when the discourse is focused on a single topic or activity, this method works well. However, rule-based techniques lose efficiency as the input grows more naturalistic or the topic becomes more open.

2.2.1 Semantics Based Techniques

Semantic Graphs are a logical hierarchy of real-world taxonomies. Most taxonomies are structured around concepts, which are also known as classes. The knowledge base consists of instances from various classes together with the taxonomy. It also includes a slot-based method where class properties are defined by slots. These categories can be linked together to form a hierarchical network. It has the advantage of being able to search across the nodes as well as suggest new responses using unique reasoning principles. It is also possible to establish

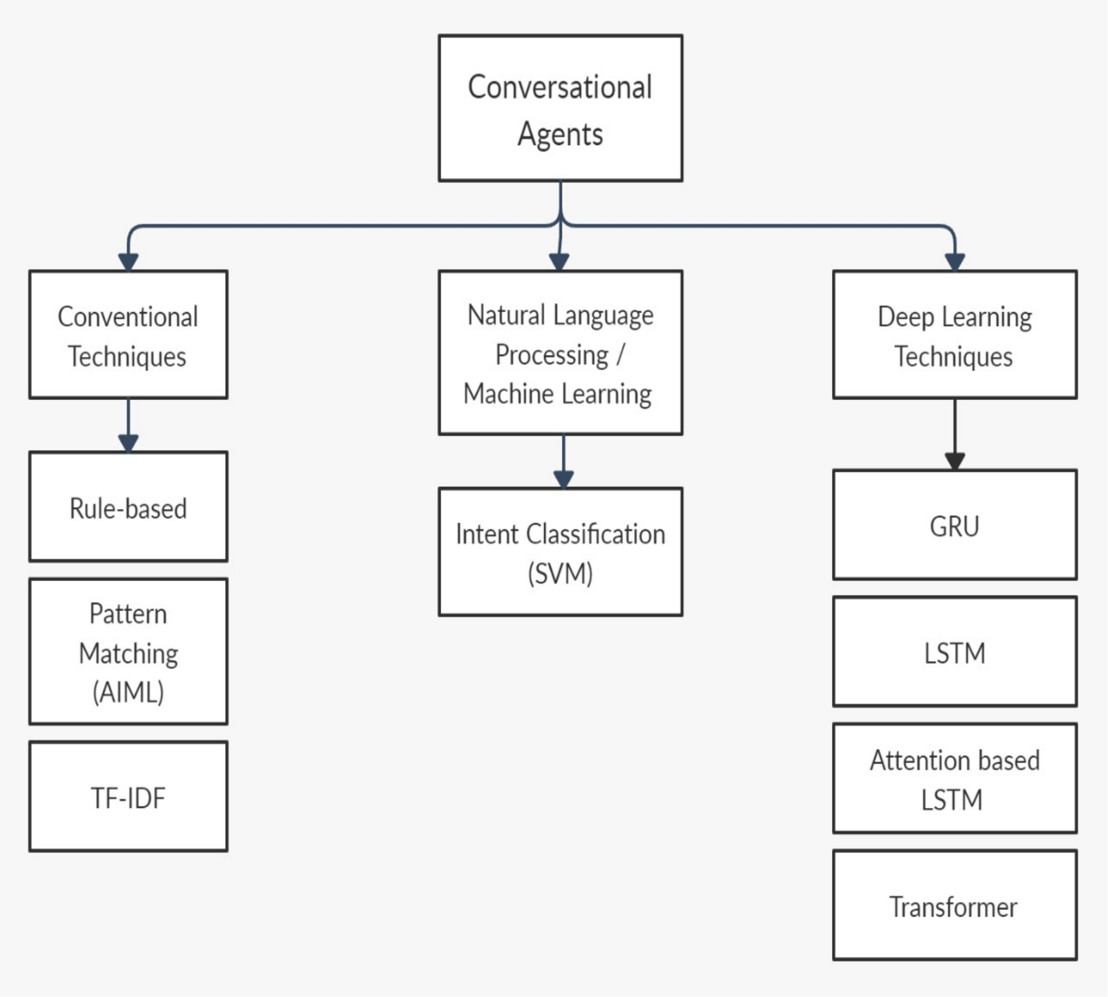


Figure 2.1: Taxonomy of Conversational Agents

different aspects of the slots. Chatbots have used the Wordnet ontologies [12]. Chatscript is a language for creating chatbots. The chatbot created by Wilcox, named Suzette, was based on this scripting language. It is a scripting language that is essentially a reworked version of AIML. Chatscript looks for a comparable context rather than looking through large numbers of categories for a match. A concept is a type of setting in which rules are defined [13].

2.2.2 Relational Database and SQL

In the construction of the chatbot, a relational database management system (RDB) can be used. Data is stored in a relational database to employ the TF-

IDF approach. N-gram TF-IDF can be used to extract keywords from the user query, where keywords are weighted to determine the correct response [14]. Each phrase has a weight associated to it and that weight is determined by using TF-IDF. Term Frequency is abbreviated as TF, and IDF stands for Inverse document frequency. TF calculates the occurrence of the word in a sentence while the weight of rare terms is calculated using IDF. Then tf and Idf scores are multiplied to get the weighted score. An N-gram TF-IDF can calculate a sentence of variable length.

2.3 Deep Neural Network

The time-consuming process of establishing rules is no longer necessary because of the chatbots that are powered by neural networks. The output of a neural network can be achieved in one of two ways: either by retrieving information from a huge dataset retrieval-based or by constructing something from scratch, known as generative output. There have also been some new combination strategies that combine the two of these approaches. Deep Neural Networks, often known as DNNs, are extremely effective models that have demonstrated outstanding performance on challenging learning tasks. Despite the fact that DNNs perform admirably whenever big and labelled training sets are accessible, they cannot be used to map one sequence to another.

2.3.1 Sequence to Sequence

Deep Neural Networks only work when we have fixed dimension vectors for both inputs and targets. So it creates a severe constraint when we are dealing with sequences of variable length, as in the case of machine translation, speech recognition, or question answering. This makes the limitation quite problematic. Therefore, it is clearly evident that a method that is not exclusive to a domain and that trains and maps sequences would be beneficial. Sutskever et al. [15] devised the solution to the sequence-to-sequence issue by demonstrating the very basic implementation of the Long Short Term Memory (LSTM) architecture. In this study, A multilayered LSTM maps the input to a vector of fixed dimensionality, and then the target sequence is decoded using a deep LSTM layer. In order to produce a huge vector representation of fixed-dimension, first an LSTM is used to read and encode the input sequence one timestep at a time. A second LSTM is then used to decode the output sequence from the vector. Due to the significant delay in time between the inputs and their associated outputs, the LSTM is a good option for this purpose because it can effectively learn the data with lengthy dependencies.

In [16], Vinyals et al. carried out research that demonstrates how to train a conversational engine using a straightforward language model based on the seq2seq framework, where the model converses by anticipating the subsequent sentence given the prior statement or sentences in a conversation. It can be trained as an end-to-end conversational model, necessitating far fewer manually created rules. Both a domain-specific dataset and a sizable, noisy, and broad collection of movie subtitles can be used to extract knowledge. This model can execute basic types of common sense reasoning on a dataset that is noisy and open-domain. It can also implement a model that solves technical issues and it is trained on a domain-specific dataset. In this study, the authors used single-layered LSTM for training of the IT helpdesk dataset and two layered LSTM for the Open Subtitles dataset. This model lacks consistency.

An abundant amount of research has been carried out on conversational agents for the English language. Research on local languages is challenging because they lack vocabulary that has been generated specifically for them. Boussakssou et al. [17] presented the research in the Arabic language by building a chatbot that is trained on a large open-domain using Gated Recurrent Units (GRU) and LSTM. This model employed a dataset of around 81,659 Arabic conversation pairings that were manually constructed without the use of any custom rules. It would be beneficial to employ additional deep learning models in order to evaluate the correctness of the model. Larger models call for large amounts of data and are more expensive to construct and keep up-to-date, where responses of high quality are typically not required for various tasks the majority of the time. In this research model, Mathur et al. [18] uses fewer resources and provides a technique to enhance conversation data without increasing the amount of vocabulary.

A comprehensive examination of the best configurations for the Seq2Seq model was carried out by Palasundram et al. in [19]. The Gated Recurrent Unit (GRU), an alternative form of RNN, was utilised in this model with the comparative performance of word vs. character embedding and the effect of the dropout layer on the GRU was measured. A small dataset of 100 QA pairs of questions and answers was curated by hand. Due to the limited size of the dataset, it may lead to overfitting. In [20], Dzikien˙e et al. presented the authors have built efficient ways to create generative chatbots based on the seq2seq algorithm with very little data. As a result of the fact that the trials were conducted in both English and the morphologically difficult Lithuanian language, they were able to compare the findings of languages that have extremely distinct qualities. Typically, companies own only a few limited datasets that are specific to their domain. This paper offers a solution to the problem of creating a generative chatbot with only a limited amount of data. They explored the LSTM, stacked LSTM, and BiLSTM techniques, all of which are RNN-based.

In [21], authors have implemented an LSTM based model on a manually curated small dataset for university education. Whereas, BiLSTM was implemented on large movie data to conduct the research on chatbots in [22].

2.3.2 Attention

By focusing on certain areas of the source text during translation, Bahdanau et al. [23][27] introduced the attention mechanism for the neural network that improved neural machine translation. The foundation for Luong et al.[24][27] presentation was the global attention and local attention mechanisms for machine translation.

Abdullahi et al. [25] suggested a chatbot system that simulated the use of seq2seq based encoder and decoder and the attention mechanism. To transfer the input sequence to a vector of a specified dimensions, the Seq2Seq architecture employed, used a multi-layered and modified recurrent neural network GRUs. The target sequence is decoded from the vector by yet another stacked deep GRU. For input and output sequences, the model employed two distinct GRUs for input and output sequences. When considering the long dependency terms, however, attention mechanisms are required to contextualise a portion of the information. It is used as a supplement to the gradient descent vanishing problem [26]. The goal is to create a context vector that tells us about all of the inputs at a global level and highlights the most crucial information

The Attention-based Transformer model was initially suggested by Vaswani et al. [26], and it has demonstrated very good performance for machine translation. For instance, it reached the highest results on the translation dataset. Encoder and decoder components are included in the Transformer, just as they are in the sequence-to-sequence paradigm. The encoder and decoder are composed of a stack of layers that are all the same, and they operate on the principle of multi-head self-attention. For the purpose of figuring out the best solution in the domain of seq2seq chatbot, Hardalov et al. [28] conducted a study on automating Twitter customer care using two models. The first was focused on information retrieval (TF-IDF), while the second method was focused on generative neural networks that were further trained on the sequence-to-sequence and Transformer models. However, results in this research demonstrated that generative neural models perform better than retrieval-based models, although generative models do suffer when the amount of data is too limited. Masum et al. [29] conducted the research on the Bengali general knowledge Question Answer (QA) dataset to train a chatbot using the transformer model. This achieved state of the art results on the Bengali dataset as compared to seq2seq models. The Transformer model scored an 85.0 BLEU score, whereas Seq2Seq achieved a maximum of 23.5 BLEU score.

2.4 Urdu Based Conversational Agents

Since Urdu is a language with limited resources, no attempt has been made yet to develop a conversational model for a Urdu using cutting-edge deep learning techniques. The majority of well-known ones are task-oriented and employ traditional techniques.

Kaleem et al. [30] presented the conversational agent named UMAIR, which is a goal-oriented, rule-based conversational agent that incorporates string similarity measurements. The architecture of UMAIR is made up of a number of parts that work together to address the special challenges of the Urdu language. To address Urdu's specific linguistic issues, UMAIR's architecture combines a scripting language and the WOW (Word Order Wizard) string similarity algorithm. The production of precise intents for the associated unstructured Roman Urdu data is presented by Shabbir et al. [31] in this research. It integrates the RASA Framework with a knowledge graph to retain the dialogue history for an intent-based natural language mechanism for chatbot communication. it acquired the accuracy of 82.1%. The work is done based on semantic technologies for the development of the responses.

In [32], attention-driven, deep encoder-decoder-based neural conversational agent for the Urdu language is built in this research. The model was trained on a small, manually curated dataset of 5K lines on Pakistan's general knowledge. It achieved a Bleu score of 56 with an LSTM-based model.

2.5 Critical Analysis

According to the comprehensive literature review of the relevant studies and based on comparative analysis in Table 2.1, we have drawn following conclusions.

- The majority of studies in the field of deep learning based conversational agents are carried out in the English language.

Study	Technique	Dataset	Data Size	Language	Domain
Vinyals et al.	LSTM	Opensubtitles/ IT Helpdesk	Large	English	Open/ Close
Boussakssou et al.	GRU, LSTM	Arabic Blogs	Large, 81K	Arabic	Open
Mathur et al.	GRU, LSTM	OpenSubtitles	Large	English	Open
Palasundram et al.	GRU	Manually Curated	Small, 100QA	English	Closed
Dzikieṅe et al	LSTM, BiLSTM, Stacked LSTM	Tilde Company	Small	English Lithuanian	Closed
Abdullahi et al.	GRU	Cornell Movie	Large	English	Open
Hardalov et al.	TF-IDF, LSTM	TwitterCustomer Support	Large	English	Open/ Close
Kaleem et al.	Rule-based	Manually Curated	Medium	Urdu	Closed
Shabbir et al.	Knowledege-Graph Intent based	Manually Curated	Large	Urdu	Closed
Alam et al.	BiLSTM	Maunally Curated	Small 5k	Urdu	Closed

Table 2.1: Comparative Analysis Literature Review

- We have also concluded that there are primarily two categories of conversational agents: those that are rule-based and those that are end-to-end chatbots that are based on deep learning techniques.
- It was observed that the domain-specific conversational models tend to be trained on smaller datasets.
- Lack of openly available Urdu Dataset

Deep learning is a field that has not yet been fully studied for the purpose of creating conversational agents in a native language such as Urdu, which is morphologically a complex language, requires further research because of its complexity. Also, the previously deployed manual annotation and rule-based approaches lacked the generalisation and scalability capabilities of the neural network-based conversational agents. The implementation of our suggested architecture and baseline deep learning models tends to overcome the existing gap that is indicated in the literature by adopting an end-to-end approach for conversational agents on a dataset that was compiled specifically in Urdu. We will also conduct results on different sizes of datasets to cater to the problem of scalability. Therefore, in this study, we will concentrate on an end-to-end neural network-based Urdu chatbot. Additionally, we will investigate a range of methods to develop a more accurate model.

The following chapter includes a thorough analysis of the literature on the development of conversational agents, conventional building techniques for chatbots, and the research that has been done on deep learning-based techniques. At the end of the chapter, it will also include a comparative table, a taxonomy of conversational agents, and a critical analysis.

Chapter 3

Background

As it is a low-resource language, Urdu has a restricted supply of data sets and lacks the fundamental natural language processing techniques. In addition, the vast majority of the text resources for Urdu text are not easily available. In addition to these facts, Urdu is gradually becoming more mature for NLP-based applications [33], as a significant amount of work is being done at the native level.Researchers are extensively implementing deep learning models for the purpose of conversational agents in resource-rich languages. Researchers' faith in artificial intelligence has been reinforced as a result of these models (AI).

In this chapter, the characteristics of the Urdu language and Deep Learning models, in particular LSTM, LSTM with Attention, and Transformer, are broken down and examined in further detail. The chapter will begin with a brief overview of the Urdu language.

3.1 Characteristics of Urdu Language

Urdu is a language native to South Asia and is recognised as one of the most extensively used languages on the subcontinent. It is a prominent language all over the world, spoken by around 300 million people in different parts of the world [34]. Urdu has not received nearly as attention or work as it deserves because there is a shortage of resources. It has a character set that consists of 38 different symbols, as seen in Figure 2.1. In addition to having a complicated morphological structure, it possesses the following distinctive properties, all of which combine to make it a difficult language to work with while doing computational tasks. The components of the pre-processing tasks, such as removing stop words and diacritics, normalising the text, and stemming, are broken down and displayed here.

- The vocabulary of Urdu is significantly influenced by Persian, Turkish, and Arabic, Urdu also draws heavily from Sanskrit, Portuguese, in addition to English.. It has significant influence from other languages. For example, Arabic(ذکر), Turkish(جواب),Persian(بہار) and English(ٹیلیفون).
- In the language of Urdu, the idea of capitalization does not exist. In the following statement, for example, (کیا یہ فون بالکل نیا ہے؟ , Is this a new phone?) there is not a single instance of capitalization. Therefore, as a result of this characteristic, the beginning of the phrase as well as proper nouns cannot be easily detected. [35].
- Urdu is written Right to left and has preferably Nastalique writing style as shown in Fig. 3.1. Nastalique writing style is complex in its nature.

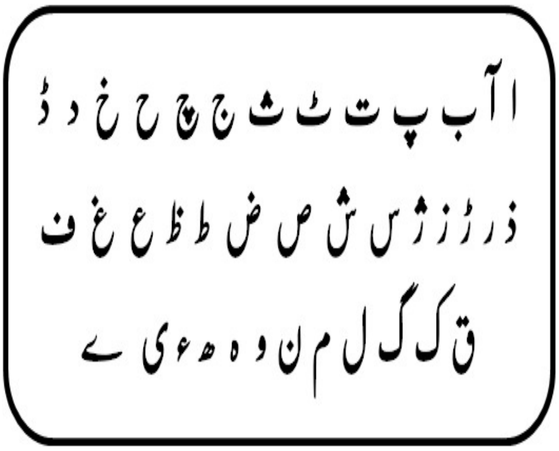


Figure 3.1: Alphabets of Urdu Language

- As there is more than one possible word order in Urdu for sentences with the same meaning, this language is sometimes referred to as a free word order

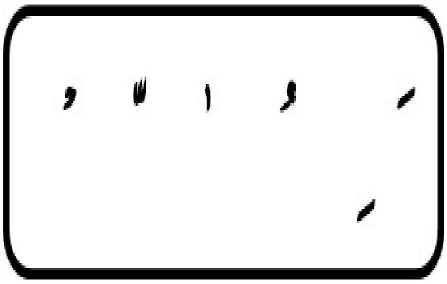


Figure 3.2: Diacritics of Urdu Language

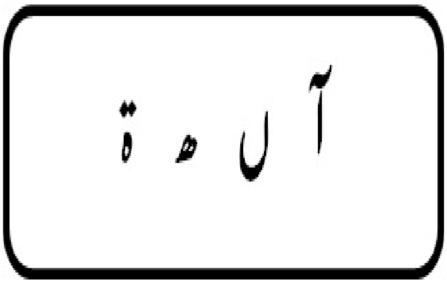


Figure 3.3: Secondary Urdu Alphabets

language. The sentence, 'Your car is good' can be written as, (تمہاری گاڑی اچھی ہے , اچھی ہے تمہاری گاڑی , and as گاڑی اچھی ہے تمہاری). All of these sentence are correct.

- The use of diacritics can alter the meaning of words, despite the fact that they are spelled and spoken the same as shown in Figure 3.2. e.g. 'اِس', pronounced as 'Iss' though it means 'This', 'اُس', pronounced as 'Uss' and it means 'That'.

- Urdu's greatest distinguishing quality is its sensitivity to context, which manifests itself in the letters' ability to alter their form depending on the letters that come before and after them. Due to the fact that spaces between whole words do not always signal word boundaries, context-sensitivity ends up causing difficulties that are connected to word segmentation. It's possible to get these words confused with others. Also there are secondary Urdu Alphabets as described in Figure 3.3 that are derived from the primary words.

3.2 Deep Learning Models

Deep Learning is an advanced subfield of Artificial Neural Networks (ANN). DL is now frequently used in well-liked NLP fields as a result of its enormous success in recent years in a number of practical applications. Artificial neural networks are used in deep learning to learn tasks using a multi-layered network. Artificial neural networks (ANNs), which are modelled after the human neural system, are employed to learn features of enormous amounts of observational data in order to make predictions about data that has not yet been seen. By controlling the weights among neurons, it mimics the way a biological brain learns to accomplish various tasks. The higher layers of the deep neural network often learn complicated features, whereas the lower layers, which are closer to the input data, learn basic features.

When deep learning is used for text generation tasks like machine translation or conversational agents, it is reasonable to anticipate that the algorithm will learn and make use of complex features such as word interactions and word patterns. This is in contrast to traditional machine learning algorithms, which only consider the score of word occurrences as a feature. For the purpose of attaining the highest possible level of model performance, these models go through an extensive amount of parameter tuning and optimization, as well as ongoing architectural fine-tuning.

Deep learning models are distinct from conventional machine learning models in the sense that, prior to applying a predictive model, they convert an initial form of prediction variables into an abstract set of features. In this way, deep learning models differ from conventional machine learning models. When generating a text with deep learning, the model is given both the outcome and a logical vector representation of the predictive variables. Before passing on the information to the model layer, the deep learning model will first learn the most important characteristics of the incoming data. During this stage of the process, deep learning models automatically develop complex feature representations by fitting a large number of parameters to their respective values. As it is a difficult task for humans to precisely specify predictive textual features in advance, this capability of deep learning models allows them to create the most effective predictive features from the initial text representation.

In the numerous sections that are to follow, the primary deep learning models and related approaches, the majority of which are used for text generation tasks, will be reviewed.

3.2.1 Recurrent Neural Network

RNNs, which stand for recurrent neural networks, are the most effective solution for neural network challenges because RNNs are capable of sequence modelling [36]. ANNs have inputs that are not dependent on one another in any way, but RNNs have inputs that are dependent on one another. Due to the fact that RNNs have their own memory, they are able to retain not only the most recent input but also the input that came before it, which makes sequence modelling tasks far simpler [37]. It is extremely capable of performing tasks such as language production, language translation, and sentiment analysis, as the result at any given time step depends not only on the most recent input but also on the output produced at previous time steps. The following equation describes the output y_{t} of the system at a specific time step 't', where current input is defined by X_{t}, prior output is defined by $\mathrm{y}_{\text{t-1}}$, the weight connected to $\mathrm{y}_{\text{t-1}}$ is described by W, and b is the bias term.

- The hidden layer receives X_1 as an input, and it outputs y_1.
- In the next step, y_1 along with X_1 is the input of next step.

$$y_{\mathrm{t}} = \sum_{t=0}^{T} X_{\mathrm{t}} w_{\mathrm{t}} + W y_{\text{t-1}} + b \tag{3.2.1}$$

RNNs provide unique challenges when it comes to training, one of which is the modelling of long-term sequences like vanishing gradients. Vanishing gradients occur if the algorithm assigns a greater value or too small a value to weights, which causes the model to stop learning. As a direct result of this, the length of the sequential data will be restricted. Figure 3.4 presents the RNN architecture for inspection.

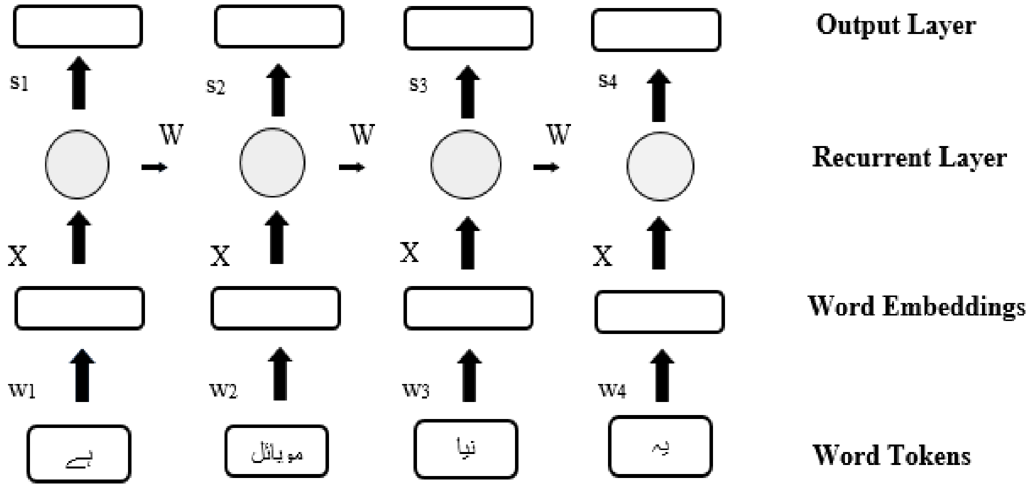


Figure 3.4: RNN Model

3.2.2 Long Short Term Memory

In order to solve the RNN issue of vanishing gradients, the authors in [38] proposed a Long-Short Term Memory Model. Using particular input, LSTM is an RNN-based architecture that is capable of learning long-term dependencies using forget, update, and output gates. LSTM introduces special hidden units known as memory blocks. The historical state of the neural network is preserved in memory cells within the memory block that have recurrent connections. The special multiplicative units in this block, known as gates, regulate the information flow via each unit. According to the weights, the model is learning, the input gate approves or rejects sequential data, and the forget gate activates or deactivates a neuron. The output gate chooses the units' output value for the LSTM.

Each memory block has an input and an output gate in the traditional LSTM architecture. The output gate regulates the flow of activations from the current memory cell to the remainder of the neural network, while the input gate regulates the flow of activations into the memory cell. The original LSTM architecture has the issue of making it difficult for LSTMs to analyze continuous input streams if the input flow is not sub-sequenced. Later, a forget gate was developed to enable adoptive forgetfulness as a solution to this problem. The LSTM cell's memory can be reset or forgotten with the help of the forget gate. Fig. 3.5 presents the LSTM's basic design.

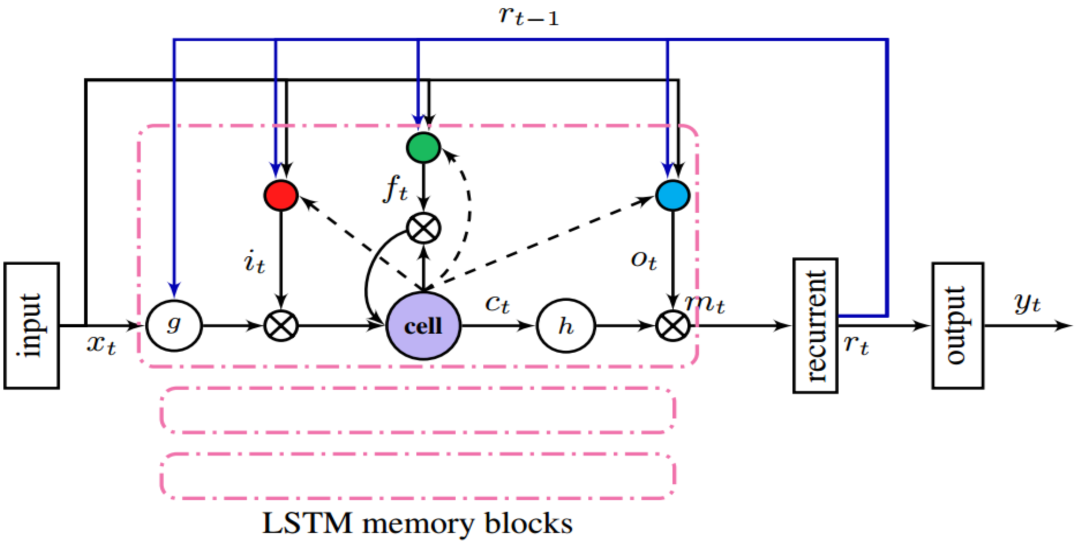


Figure 3.5: LSTM Cell [43]

3.2.3 Seq2Seq with Attention

A neural network includes an attention mechanism. It determines which source elements are more crucial at each decoder phase. In this configuration, the encoder does not need to reduce the entire sentence into a vector representation; rather, it provides representations for each source token, such as the complete set of RNN states rather than the most recent one. The fundamental concept is that a network may discover which input components are more crucial at each stage. A model based on attention can be trained from beginning to end because everything in this situation is differentiable. The model will learn to choose important information on its own; you don't need to explicitly train it to choose the words you desire. As Figure 3.6 shows that he attention layer is added separately.

Attention gets input at each decoder step h_t, including each encoder state (s_1, s_2,

s_3,...,s_t), and computes attention scores. Attention calculates the significance of each encoder state for this decoder state. Essentially, it performs an attention function that takes one encoder and one decoder state as inputs and outputs a scalar value. The most common methods for calculating attention scores are:

- The simplest way is dot-product.
- The approach suggested in the original research is a multi-layer perceptron, sometimes known as ”Bahdanau attention.” [23].
- Bilinear function, sometimes known as ”Luong attention” Attention output is weighted sum of the encoder states with attention weights [24].

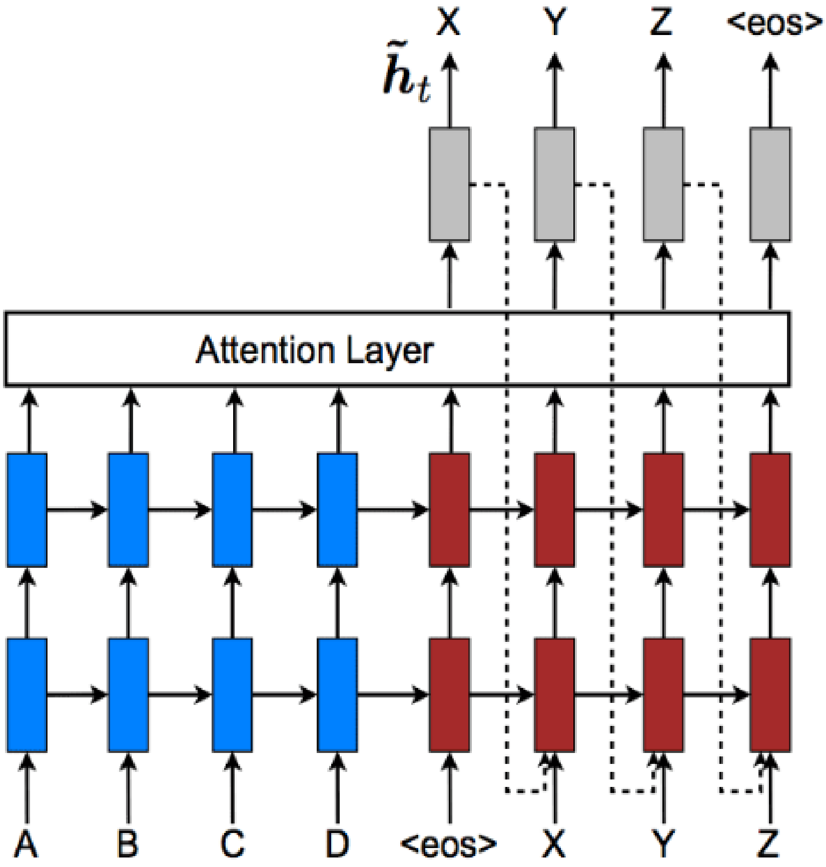


Figure 3.6: Seq2Seq with Attention [42]

Bahdanau Model

- In Bahdanau based attention, encoder is bidirectional.

- The encoder uses two RNNs that read data in opposing directions—forward and backward—to more effectively encode each source word.
- The two RNNs' states are combined for each token.
- Multi-layer perceptron is used to calculate the attention score

Luong Model

- In Luong based attention, encoder is unidirectional.
- Bilinear function is used to calculate the attention score
- For this step, there is a focus on the relationship between decoder RNN state and prediction.

3.2.4 Word Embedding

Word embeddings, neural network-based models that represent words as dense and dispersed vectors, have become more popular as a result of recent advancements in NLP. Performance improvements in numerous NLP applications have been made possible by these embeddings. By assisting Deep Learning algorithms in learning textual patterns more easily with better representations of words and obtaining better generalization of output from less information, In addition to enhancing the performance of cutting-edge algorithms in numerous NLP applications, the use of word embedding has sparked critically needed studies on resource-scarce languages. The outcomes of word embedding are frequently used as input characteristics in deep learning models for NLP [39]. Semantic characteristics of the wWords are encoded as vectors represented in word embeddings. Vocabulary words are converted into vectors of continuous real values using the language modelling and feature learning process known as word embedding. This method typically entails embeddings—where each word is treated as a dimension—from a high-dimensional dispersed vector space, such as one-hot encoding, to a low-dimensional compact vector space. These embedding vectors each have a dimension that represents a latent property of a word in text. The language's regularities and patterns are encoded inside these word vectors.

Word2Vec The two most common methods for learning word embeddings are the use of neural networks and matrix factorization [40]. A popular word embedding system is Word2Vec, a powerful model that successfully extracts word embeddings from text. There are two models in it: the Skip-gram model and the Continuous Bag of Words (CBoW) model. In contrast to the SG model, which predicts context words from the target, the CBoW model uses context words to predict the target word. The CBoW approach, which works well with tiny datasets, treats the full text context as a single observation. The SG model, on the other hand, treats every context and the target words in pairs as a fresh observation and performs best with big datasets. Pre-trained word embeddings for the Urdu language can be created using the Word2Vec model [41].

Next chapter provides an illustration of the methods that will be used in the proposed architecture. The chapter Proposed Model and Dataset will cover the specifics of the development of the proposed English to Urdu tanslated benchmark corpus for the conversational agents, as well as the annotation process, the steps involved in data preprocessing, and the proposed deep learning architecture for end-to-end chatbot.

Chapter 4

Research Methodology

This chapter discusses in detail the methods that will be used in the proposed architecture. The domain of deep learning has not yet been investigated for native Urdu based conversational agents. The implementation of our suggested architecture and baseline deep learning models tends to overcome the existing gap that is indicated in the literature by adopting a deep learning framework.

This chapter elaborates on the development of a proposed benchmark corpus for conversational agents based on the Urdu language, the procedure for translating and annotating the data, the steps involved in data pre-processing, as well as the proposed deep learning architecture and evaluation metric that were used to test the model.

4.1 Datasets

The conversational agent's training data must follow a natural conversational flow and should be domain-specific for customer support. It must be in the form of a phrase or a question that may be answered. We were able to locate data sources that met the requirements for this.

Two different data sets are utilised in the process of training the suggested model

Q	Is there a SIM card in Samsung Galaxy?
A	Yes. The Samsung Galaxy accommodates a micro SIM card.
Q	Why hasnt it upgraded to latest Android OS. Is it because it is unlocked?
A	My Samsung S was able to upgrade to Android last week, my service is with ATT and it shouldn't matter
Q	can in it be used abroad with a different carrier?
A	Yes, Iphone can be used with ATT

Table 4.1: Sample of Amazon QA Data Set in English

Q	کیا یہکیا سیمسنگ گیلیکسی میں کوئی سم کارڈ ہے؟
A	جی ہاں. سیمسنگ گیلیکسی ایک مائیکرو سم کارڈ کو ایڈجسٹ کرتا ہے۔
Q	اسے جدید ترین اینڈرائیڈ او ایس پر اپ گریڈ کیوں نہیں کیا گیا؟ کیا اس کی وجہ یہ ہے کہ یہ غیر مقفل ہے؟
A	میرا سیمسنگ ایس پچھلے ہفتے اینڈرائیڈ میں اپ گریڈ کرنے کے قابل تھا، میری سروس اے ٹی این ٹی کے ساتھ ہے اور اس سے کوئی فرق نہیں پڑتا
Q	کیا اس میں ایک مختلف کیریئر کے ساتھ بیرون ملک استعمال کیا جا سکتا ہے؟
A	ہاں، آئی فون کو اے ٹی این ٹی کے ساتھ استعمال کیا جا سکتا ہے۔

Table 4.2: Sample of Amazon QA Data Set in Urdu

for conversational agents. The data from Amazon's Question-and-Answer[1] service is the initial data collection that was utilized in this investigation. It was obtained from an online source and is based on the interactions that took place between customers and retailers regarding the product. Review websites such as Amazon offer QA systems that let people ask questions about specific products to other customers and retailers in order to assist users with their answers. The Amazon Question-Answer dataset was initially made available in JSON format; however, we were able to convert it into a text file, as shown in Table 4.1, with an odd number of lines representing questions and an even number of answers. It was carried out for scalability reasons. It covers approximately 100K lines of data from a single domain, namely cellphones. We had to resort to translating and annotating a subset of the English dataset because there is no publicly available question-and-answer based dataset in the Urdu language as demonstrated in Figure 4.2. The native Urdu dataset was translated using the Google Translator API. This dataset was initially annotated up to about 10k lines, but a smaller dataset of about 3k lines was developed for comparison with domain-specific conversational

[1] http://jmcauley.ucsd.edu/data/amazon/qa/

Properties	Amazon Urdu Large	Amazon Urdu Small	Amazon English
Size	Large	Small	Large
Questions	5000	1500	50000
Answers	5000	1500	50000
Sen. Max Length	40	40	65

Table 4.3: Comparative Analysis of Data Sets Properties

agents that are trained on smaller datasets, as seen in the literature review comparative study. The comparison of data set sizes also served as a test to see if deep learning models perform equally well with large or small data sets. Table 4.3 provides an explanation of the specifics of the dataset in both English and Urdu.

4.1.1 Data Translation and Annotation

The basic objective of data annotation is to obtain ground truth, which means achieving a state in which the annotated and translated data flawlessly satisfies the requirements. Automatic data annotation and manual data annotation are the two primary varieties of this form of annotation. When compared to manual annotation, automatic annotation is regarded as having a lower level of precision, yet it is able to annotate many more datasets in a more time-efficient manner than individuals. Manual annotation, on the other hand, is more accurate but requires diligent annotators to be attentive and precise.

For the purposes of research, we constructed an Urdu dataset repository using a freely available dataset of Amazon QA questions and answers in English from an online resource. Although it was translated using the Google API, certain words, such as Samsung Galaxy and iPhone, as well as the names of several companies, were not translated. These terms required a search and then needed to be replaced manually. Table 4.4 presents the initially translated version, whereas Table 4.2 displays the annotated version.

Q	کیا SamsungGalaxy میں کوئی سم کارڈ ہے؟
A	جی ہاں. Galaxy Samsung ایک مائیکرو سم کارڈ کو ایڈجسٹ کرتا ہے۔
Q	اسے جدید ترین AndroidOS پر اپ گریڈ کیوں نہیں کیا گیا؟ کیا اس کی وجہ یہ ہے کہ یہ غیر مقفل ہے؟
A	میرا Samsung ایس پچھلے ہفتے Android میں اپ گریڈ کرنے کے قابل تھا، میری سروس ATT کے ساتھ ہے اور اس سے کوئی فرق نہیں پڑتا
Q	کیا اس میں ایک مختلف carrier کے ساتھ بیرون ملک استعمال کیا جا سکتا ہے؟
A	ہاں، iPhone کو ATT کے ساتھ استعمال کیا جا سکتا ہے۔

Table 4.4: Sample of Translated Urdu Data Set

4.1.2 Pre-Processing

Before feeding the input into a neural network, a fundamental step in natural language processing known as preprocessing helps to organise the dataset by performing fundamental operations on it. Tokenization of translated sentences and the creation of a vocabulary for both the source language and the target language are the fundamental operations of the process. The words that are not in the vocabulary are represented by the special token <OOV>. It organises the unprocessed data and assigns it a value so that it can be used in subsequent tasks.

The text is preprocessed before any generation or classification tasks are performed on it. This ensures that only clean, normalised, and structured data is used, which ultimately leads to more accurate results. Data can be maintained in a form that is free of redundancy and noise with the assistance of preprocessing. Researchers have used the procedure to a large extent in order to obtain data that has been cleaned for improved interpretation of applied models. All of the data sets that were being used in the implementation of our suggested model have been subjected to text preprocessing in order to achieve standardisation.

To begin, we have separated the punctuation from the words because, when words are tokenized, each word is treated as its own separate token. For instance, in (کیا یہ فون بالکل نیا ہے؟ , Is this a new phone?), tokenization is a process in which each word of a document is treated as its own separate token. Consequently, we have separated the punctuation from the words. After that, additional preprocessing is done, which involves removing any alphanumeric characters, URLs, and converting English alphabets and words to Urdu. This ensures that the document only contains words from the language that is being processed. After the data has been preprocessed, it is input into the neural network model so that the performance of the model can be assessed.

4.2 Methodology

The main objective of this dissertation was to create a conversational agent based in Urdu that could assist with customer service. The Transformer concept was initially presented in a 2017 research article titled ”Attention Is All You Need” [26]. It is completely reliant on attention mechanisms; there is no recurrence or convolutions involved in its construction. In addition to producing better translations, the model can be trained significantly more quickly. Transformers and their variants are currently the default models for sequence-to-sequence tasks.

In comparison to earlier models in which processing across encoder and decoder was accomplished through the use of recurrence or convolutions as depicted in Table 4.5, the transformer functions only through the utilisation of attention. In a sentence, 'I need a new phone', RNNs must read the entire sentence in order to grasp what the word 'phone' means while encoding a statement, which can take some time for lengthy sequences. On the other hand, in the encoder of Transformer, tokens are encoded simultaneously.

On the basis of these facts, the model for an Urdu-based conversational agent that was discussed earlier in this paragraph is the one that our research suggests. We have implemented a Transformer-like architecture so that we may test the robustness of the system that has been designed. In this study, our objective is to determine how transformer-like design compares to RNN-based architecture in terms of the impact it has on Urdu conversational agents.

In addition to the model that we have proposed, both LSTM and LSTM with Attention have been applied to Amazon QA in both English and Urdu. The methodology that has been developed achieves better results on data sets of varying sizes.

Processing	Seq2Seq	Seq2Seq-Attention	Transformer
Encoder	RNN	RNN	Attention
Decoder	RNN	RNN	Attention
Encoder-Decoder	Fixed sized vector	Attention	Attention

Table 4.5: Processing with Seq2Seq Models

After obtaining new data, translating it, and annotating it, the next phase in the process is to preprocess the data that has been collected. Regarding the establishment of parameters, we have optimised parameters for both the baseline models that we have presented, specifically LSTM and LSTM with Attention. Self-attention, multihead attention units, and the whole architecture of the Transformer model have all been detailed here. Here is a list of the several layers that make up the hierarchy:

4.2.1 Self Attention

The encoder of the Transformer can be viewed naturally as a chain of logical layers. Tokens exchange information at each step and to better comprehend one another in the context of the entire sentence, as they share information at each step, and we need self-attention for that. One of the main elements of the model is self-attention. Token interaction occurs in the model's self-attention section. Using an attention mechanism, each token in the sentence examines the all the other tokens in the statement in order to collect context. The encoder of the Transformer can be viewed naturally as a chain of logical layers. Tokens exchange information at each step and to better comprehend one another in the context of the entire sentence, as they share information at each step, and we need self-attention for that. One of the main elements of the model is self-attention. Token interaction occurs in the model's self-attention section. Using an attention mechanism, each token in the sentence examines all the other tokens in the statement in order to collect context.

It then modifies its prior representation of itself. Essentially, a query-key-value approach is used to implement this understanding. There are three representations for each input item in self-attention, query, key, and value, each of which corresponds to a different role that it might perform:

- information is requested by the query.
- key responds to the query by computing attention.
- value is the information itself.

4.2.2 Masked Multi Head Attention (Decoder)

There is a self-attention mechanism in the decoder that is responsible for carrying out the function of looking at the preceding tokens. Self-attention operates somewhat differently in the decoder compared to how it does in the encoder. The decoder generates one token at a time; during generation, we are unsure of which tokens we will generate in the future. In contrast, the encoder obtains all the tokens in an input sentence simultaneously, and the tokens can view each other. The model employs disguised self-attention, masking out future tokens, to restrict the decoder from seeing forward.

It cannot during generation since we are unsure of what will occur next. However, we employ reference translations during training (which we know). In order to avoid the tokens seeing the future without masks, we feed the decoder the entire target text during training.

4.2.3 Multi Head Attention

Typically, in order to comprehend the function that a word serves within the context of a sentence, one must first comprehend the connections that exist between the word and the various components of the sentence. Not only is this vital while processing the source sentence, but it is also important when generating the target. The reasoning behind multi-headed attention is that we need to provide the model the freedom to concentrate on a variety of issues. Multi-head attention is a type of attention mechanism that, rather than having a single attention mechanism, has multiple heads or parts that operate independently.

When it comes to the implementation, all you have to do is divide the query, key, and the values that you calculate for a self-attention into a few different sections. The models with a single attention head or numerous of them will have the same dimension; having several attention heads does not result in an increase in the size of the model.

4.2.4 Transformer Architecture

Intuitively, the model achieves precisely what we explained earlier: that the tokens interact with one another and update their respective representations, in the encoder. While a target token first examines an already generated target token, then it examines the source token, and eventually updates its representation, in the decoder. This takes place on a number of different levels. Due to the fact that the Transformer does not possess a repetition or recurrence layer, all of the tokens are processed all at once, which results in an increase in the amount of efficiency achieved by the processing. One training step for recurrent models takes *O(l(source) + l(target))* steps, whereas for Transformer, the required number of steps is *O(1)*, often known as constant as it can be seen in Figure 4.2.

Feed-forward blocks: Each layer also comprises a feed-forward network component, which consists of two linear layers connected by a ReLU non-linear layer, in addition to attention. A model makes use of a feed forward network component in order to manage this additional information after first examining other tokens through the use of an attention mechanism. The weights for the feed-forward layer are trained during the training process, and the same identical matrix is used to apply them to each corresponding token position. It is a massively parallel component of the model due to the fact that it is applied independently of any communication with or inference from the other token positions.

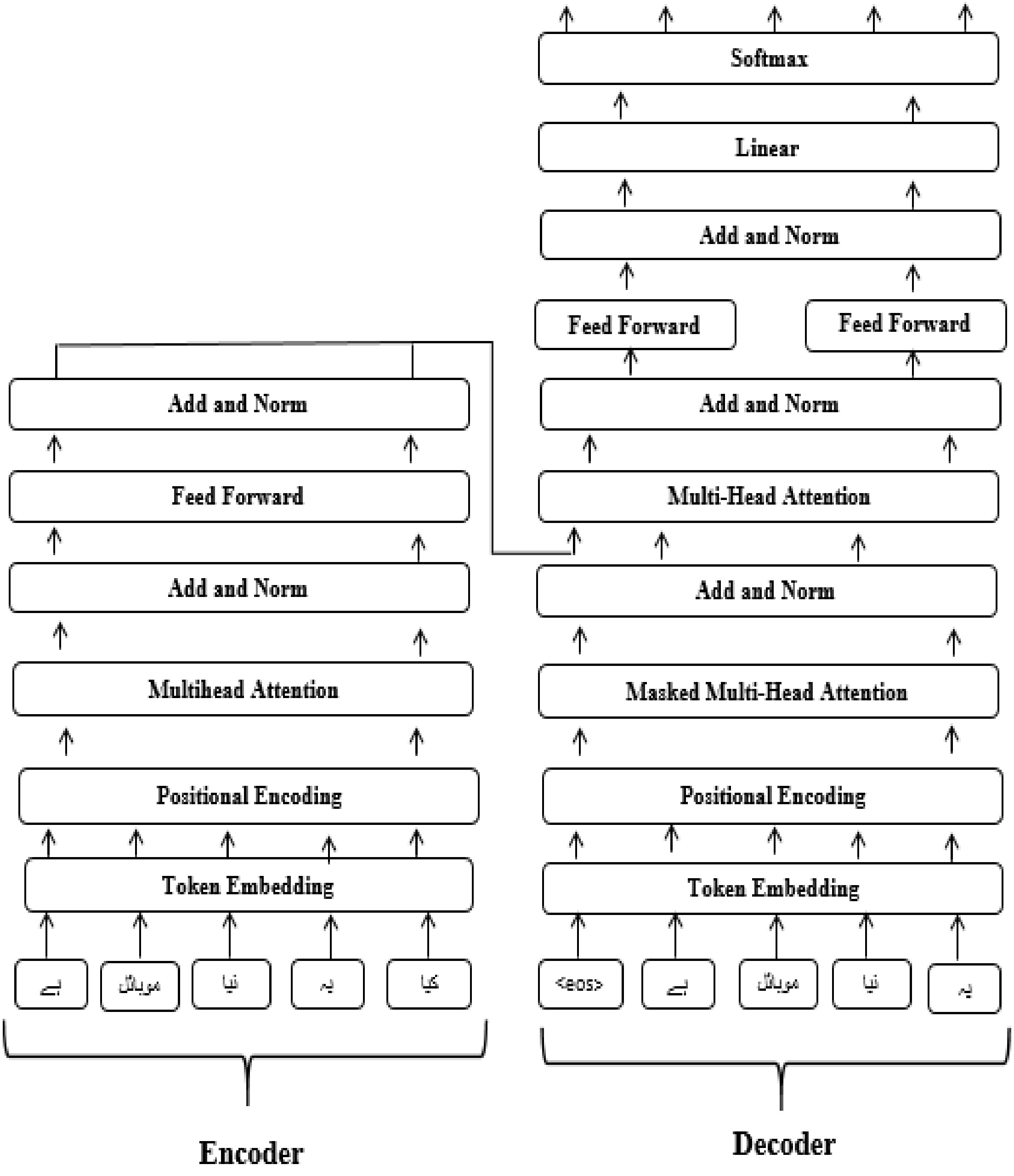


Figure 4.1: Transformer Architecture

Normalization: Layer Normalization is shown by the Norm component of the Add Norm layer. To regulate flow to the following layer, the vector representation of every token in the batch is separately normalized. Convergence stability and, occasionally, quality are improved by layer normalization. Each token's vector representation in the transformer must be normalised. Furthermore, the layer's

Parameters	Amzon Urdu Small	Amazon Urdu Large	Amazon English
Dropout	0.3	0.5	0.5
No. of Layers	2	2	2
Optimaization Function	Adam	Adam	Adam
Learning Rate	2e-05	0.0001	0.001

Table 4.6: Optimized Parameters for Transformer

trainable parameters, scale and bias, are utilised here to rescale the outputs of this layer or the inputs of the next layer. These layer properties, scale and bias, are equivalent.

Positional encoding: Keep in mind that the transformer does not know the sequence of the input tokens because it does not have recurrence or convolution. As a result, we must explicitly tell the model where the tokens are located. We have tokens, as we always do, and locations are the two sets of embeddings. Therefore, the representation of an input token is the product of its token and positional embeddings. It is possible to learn positional embeddings, but the authors discovered that using fixed values does not degrade the quality.

The optimized parameter list of the suggested Transformer model is described in Table 4.6.

The impact of data set size, language, and annotation on outcomes and evaluation will be covered in the following chapter. We will also conduct a detailed analysis of the attention-based and no-attention-based deep learning models' performances on Urdu and English data.

Chapter 5

Results and Discussion

The suggested model and various baseline models were used on two datasets in this research. The outcomes of these methods are described in detail in this chapter. For all of the data sets, the suggested model has outperformed baseline models. The primary goal of this research is to evaluate the efficiency of deep learning models on a set of Urdu data and also compare the results with English data as well. As far as we are aware, this is the first study in the field of Urdu conversational models that focuses on scalable Urdu chatbots using deep learning models. The results are thoroughly contrasted and discussed on the basis of the following crucial considerations.

In this chapter we will discuss impact of data set size, language and annotation on results and evaluation. In addition to this, we will conduct an in-depth analysis of the impact that Attention-based and No-Attention-based deep learning models have had on Urdu and English data with regard to both time and performance.

5.1 Experiments

The encoder-decoder based LSTM model was initially trained on all three datasets. The output of the simple LSTM model was then compared with that of the LSTM-attention model. The LSTM model was trained using three different datasets; we then fine-tuned it by using a batch size of 64 and setting the dropout probability to 0.3. For the purpose of optimization, an Adam optimizer was utilised, and the initial learning rate was set at 0.001 for the English dataset, 0.0001 for the Urdu Large dataset, and 2e-05 for the Urdu Small dataset. The training of the model trained for a total of 100 epochs. A neural network with 2 layers and 512 hidden units was trained to perform the task. Following that, we trained the same fine-tuned model, but this time we included an attention layer. As with the attention layer, there is a chance that the GPU will throw an out of memory error. In order to resolve this issue, we decided to employ the softmax sampling method, which involves selecting a particular vocabulary size. Due to unidirectional LSTM, 'loung' attention for the LSTM model that incorporates attention was utilized. The parameters for the LSTM model are similar to those of the LSTM model with attention to the addition of softmax sampling.

We have trained a two-layer transformer using positional encoding and multi-head attention. As optimizers, Adam and Adam beta2 with a value of 0.98 are utilised, and the initial learning rate and dropout ratio are both set to 0.5.

5.2 Model Comparison

The experimental research demonstrates that Transformer performs quite well even when applied to data sets of pretty small dimensions. The research community has always taken into consideration the importance of large data sets; nevertheless, it is not always possible to collect massive data sets in the native language of the researcher. In order to overcome this obstacle, we have carried out this experiment in the hopes that smaller data sets will one day be accorded the same level of significance as larger data sets in the realm of deep learning research.

Transformer achieves the best results when compared to LSTM and LSTM with attention model, especially since it handles long sequences better than the other two models, as shown in a comparative analysis of Question and Answers between ground truth and model predictions in Tables 5.1 and 5.2.

Question	یقین نہیں ہے کہ کون سا ماڈل ہے ۔ کیا یہ کے گلیکسی نوٹ میں فٹ ہو گا ؟
Actual Ans	ہاں، ہے ۔ یہ کیس ٹیبلیٹ سے بڑا ہے، میں نے اسے صرف گھر میں استعمال کیا کیونکہ مجھے یقین نہیں ہے کہ ٹیبلیٹ اس میں محفوظ ہے
LSTM	جی ہاں
LSTM Atten	کیس یا ہولسٹر میں ہے میں نے اسے اس کے بعد اس کی سفارش نہیں کر رہا تھا
Transformer	مجھے امید ہے کہ یہ آپ کی مدد کے لیے چارج کیا جائے گا ۔ کیس سے بڑا ہے ایسا نہیں ھے اور پاور بٹن ڈھکا ہوا ہے

Table 5.1: Deep Learning Model Comparison for Urdu Data Set

Question	Is this a hard case? Or a gel case?
Actual Ans	It's more like a hybrid of the two. It's made of rubber but feels hard like plastic
LSTM	It is a real otterbox case
LSTM Atten	It is a real case. it is a soft plastic and hard plastic.
Transformer	It's a hard plastic case, like a hard shell that can lightly be used to.

Table 5.2: Deep Learning Model Comparison for English Data Set

5.3 Evaluation Metrics

When conducting quality assessments of conversational agents models, researchers frequently make use of a wide variety of evaluation measures. The Bleu Score, and METEOR are the two performance metrics that are utilised in this particular piece of research. BLEU, which stands for "Bilingual Evaluation Understudy," is a score that is derived by contrasting a proposed text translation with the reference translations. Despite the fact that it was designed for translation, it may also be used to analyse the text that is generated for a variety of NLP. Scores in the blue zone range from 0 to 1 . Even when two humans work together to solve a problem, there is a good chance that they will come up with various possible solutions to the issue, and they will only seldom find a solution that is an exact fit. Because of this, a score that is closer to one is not realistic in practise and should serve as a warning sign indicating that the model is overfitting the data. Bleu score is calculated as follows

$$BLEU(N) = Brevity\,Penality * Average\,Precision\,Score(N) \tag{5.3.1}$$

Let's begin by gaining an understanding of N-grams, Brevity penality and Precision. Bleu Score can be calculated for a variety of different N values, where

BLEU-1 uses the unigram precision and BLEU-2 uses the unigram and bigram precision, where precision is calculated as.

$$Precision = \frac{Correctly\, Predicted\, Words}{Total\, Predicted\, Words} \tag{5.3.2}$$

The brevity penalty assigns a lower score to the translations that are generated shorter than the length of the reference that is the closest match, and this score decreases exponentially.

$$Brevity\, Penality = min(1, exp - (1 - \frac{reference\, length}{Candidate\, length})) \tag{5.3.3}$$

METEOR is an acronym that stands for ”Metric for Evaluation of Translation with Explicit ORdering.” It is a metric that is used to evaluate the output of machine translation. The metric relies on the harmonic mean between precision and recall of the unigram, with a greater emphasis placed on recall than precision. Along with the more common exact word matching, it also offers a number of capabilities that are not included in similar metrics, such as the ability to match synonym and performing stemming, among other things. METEOR M is calculated by using following formula

$$M = F_{\mathrm{mean}}(1 - p) \tag{5.3.4}$$

Precision and Recall are defined as

$$Precision = \frac{m}{n_{\mathrm{c}}} \tag{5.3.5}$$

The amount of unigram words in the candidate set are represented as $\mathrm{n_c}$, where m denotes the number of unigram words shared between the candidate translation and the reference.

$$Recall = \frac{m}{n_{\mathrm{r}}} \tag{5.3.6}$$

whereas unigram words in reference are denoted by $\mathrm{n_r}$. The harmonic mean F1 is used to combine precision and recall, with recall weighted nine times higher than precision:

$$F_{\mathrm{mean}} = \frac{10\, P\, R}{R + 9P} \tag{5.3.7}$$

Where penalty in METEOR is computed by

$$p = 0.5(\frac{c}{u_{\mathrm{m}}})^3 \tag{5.3.8}$$

5.4 Results

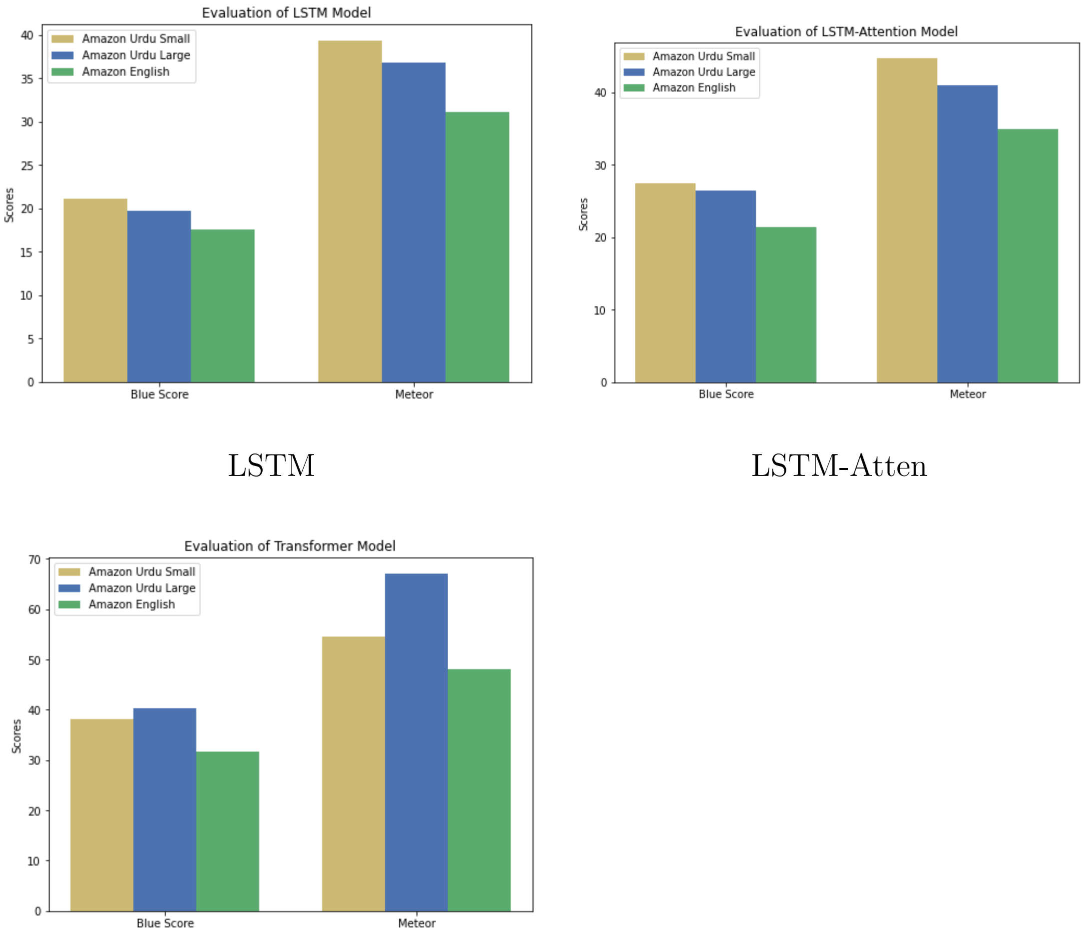
In spite of being nearly as extensive in size, the Amazon English data set has exhibited a significantly lower Bleu score than the Amazon Urdu data set. On the other hand, the performance of Transformer on all three data sets demonstrates that the Urdu small data set is superior to the other two data sets in terms of the Bleu and Meteor score. This would seem to show that the size of the data collection does not affect the overall outcome of the model. It is more dependent on the dataset's level of cleanliness and annotation, particularly for the closed-domain conversational model.

In order to investigate how the size of the data sets influences prediction scores, a variety of data sets of varying sizes have been utilised. As may be deduced from the obtained results, it is expected that Transformer demonstrates the best accuracy on the large-scale Urdu data set. Because a large data set has more features and varied sentences, it is possible for the model to learn new terms and predict them in a more accurate manner. This is due to the reason that large data sets include more features. According to what is displayed in Table 5.3, the Bleu score for Amazon Urdu Small is 21.1, 27.4, and 38.13, respectively, for the LSTM model, the LSTM model with attention, and the Transformer model. While the English dataset showed gains in blue score from 21.45 to 31.7, Amazon Urdu Large projected results of 19.7, 26.43, and 40.2 on all three datasets. It is evident from Table 5.3, that the transformer model has a significant advantage over the other two DL models and performs well on data sets of all sizes, whether they are small, medium, or huge.

In Figure 5.2 shows the training and validation accuracy of all three models. As the the curves were really noisy so we trained all 3 models 3 times each for 100 epochs. Average of every 3 curves is plotted, and Smoothing function is used to smooth the learning curves.

Dataset	Amazon-Urdu Small		Amazon-Urdu Large		Amazon-English	
Models	Bleu	Meteor	Bleu	Meteor	Bleu	Meteor
LSTM	21.1	39.3	19.7	36.8	17.6	31.1
LSTM-Atten	27.4	44.7	26.43	41	21.45	35
Transformer	**40.2**	**54.42**	**38.13**	**67**	**31.7**	**48**

Table 5.3: Transformer and other Deep Learning Results on Urdu and English Data sets



LSTM

LSTM-Atten

Transformer

Figure 5.1: Comparison Chart of Bleu and Meteor across All Data sets

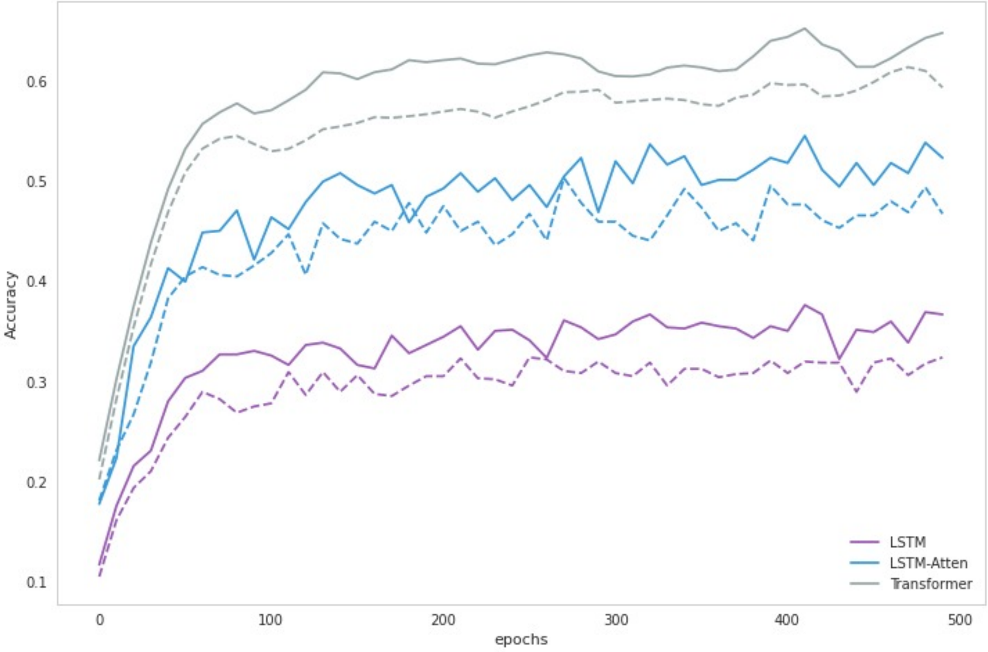


Figure 5.2: Training and Validation Accuracy Comparison

5.5 Discussion

While Amazon's English data set refers to customer retailer communication, it was preprocessed but no hand annotation was performed, thus it may not have obtained the desired results. In contrast, the Amazon Urdu data set is highly specific because it was translated, cleaned, and annotated properly. as preferred. As we shown in the Figure 5.1, smaller the dataset better the results are as the smaller dataset was annotated properly. For this reason, Transformer performs better on Amazon Urdu data sets than Amazon English sets. When different deep learning architectures are applied to the Amazon Urdu data set, it gives better results because it was thoroughly translated and annotated. In addition to Transformer, both LSTM and LSTM with Attention have demonstrated impressive performance on both sets of data. The capacity of LSTM, a type of deep learning model, to extract text features with the assistance of recurrence layers makes it an effective model. In addition, the Attention mechanism on LSTM typically demonstrates strong performance. This is because it assists in improving results for variable length sequences, which is a significant challenge that conversational models are currently encountering.

In conclusion, the model that was proposed Transformer has the ultimate advantage over other deep learning models due to the fact that the technique used in it is relatively simple and automatic in the sense that feature engineering is not involved in the whole process. This gives it a distinct competitive edge over other deep learning models. As a result, it reduces the cost and time spent on the execution of the model, and it also has a tendency to increase its overall performance. The research that was carried out serves as a benchmark study in the Urdu conversational model that makes use of deep learning approaches in addition to the contribution of a strategy that was proposed.

The following chapter concludes this study by providing a summary of the work carried out during this research. It also details the issues and challenges encountered specifically in compiling the Urdu dataset. At the end, we discussed the future work that needs to be done.

Chapter 6

Conclusion and Future Work

A summary of the work completed throughout this research is provided in the chapter. Additionally, it describes the problems and difficulties specifically in compiling the Urdu dataset. We will describe the future work at the end of this chapter.

6.1 Summary

In this study, by implementing our newly developed neural network–based deep learning model called Transformer to two data sets in Urdu, the purpose of this study was to make an effort to attract the attention of the academic community toward the language Urdu, which has a limited amount of available resources. A baseline study in the field of Urdu natural language processing has been carried out in the form of a Transformer-based Urdu Conversational agent. In order to evaluate how well our model performs in comparison to existing deep learning models, we have used two other models, namely LSTM and LSTM with Attention, on both datasets. We have given the experimental study that was chosen for the purpose of generating results on two different data sets, namely the Amazon Customer Support Urdu Q A data set and the Amazon English Q A data set. The results demonstrate that our proposed method performed better than previous baseline models when it was applied to both sets of data. No matter the size of the data in any of the three distinct size versions, our model received the highest score possible in both bleu and meteor. We have implemented an attention mechanism in the seq2seq model. However, despite the fact that it is more effective on lengthy sentences, using it requires a lot of time and effort. As a result, we have trained a Transformer, which is a linear attention-based model in which the encoder and decoder are stacked linearly; this model helps to improve performance while also reducing the amount of time and cost. As we have developed a scalable model, the method that has been suggested can also be used for a great number of different Urdu chatbots or conversational models, for example in the entertainment, health, and education industries. In the same way that Urdu is struggling with a lack of resources, there is a significant opportunity for additional research.

6.2 Issues and Challenges

The major challenge encountered while conducting this research was the lack of Urdu dataset. As the training data for the customer support based conversational agent should follow a normal conversational flow. It must be expressed as a sentence or a question that may be addressed. To address this issue we had to translate the data from the Amazon English QA customer review dataset because there was no publicly accessible Urdu data in the form of questions and answers. Despite using the Google API, some words, including iPhone and Samsung Galaxy, as well as the names of other firms, were not translated. These terms needed to be searched for, then manually changed. Another challenge was to convert back words like Galaxy to گیلیکسی whereas the translator translated it as کہکشاں. Due to lack of human resources it was difficult to annotate the large dataset as we can see from the results that the smaller the dataset was better the results as smaller dataset was annotated properly.

During training of LSTM with the attention model, out of vacablury errors frequently occured. To encounter that, we have used a softmax sampling of 5K words instead of almost all the vocabulary as in the case of LSTM.

6.3 Limitations

- Urdu dataset in QA pairs is insufficient.
- Translation and annotation of urdu data is a time consuming task.
- Translation of balanced dataset.
- There are no representative pre-trained word embeddings for Urdu data.

6.4 Future Work

As we experienced limitations in finding the representative pretrained word embeddings for the Urdu language, in the future we want to integrate a lexicon with pretrained neural word embeddings in Transformer in order to investigate the impact of utilising lexicons in conjunction with deep learning techniques. The strategies that are going to be explained in this research can also be compared to other data sets that are balanced. In addition, implementing BERT based on an additional dropout layer in order to further increase the accuracy of the model when applied to data sets of varying sizes is a research dimension that is also worth pursuing. We also want to translate and annotate some more Urdu dataset in the form of questions and answers to create a large Urdu repository freely available for research.

Bibliography

[1] Abdul-Mageed, M., Korayem, M., “Automatic identification of subjectivity in morphologically rich languages: the case of Arabic.” in 1st Workshop on Computational Approaches to Subjectivity and Sentiment analysis, WASSA, 2010.

[2] A. Daud, W. Khan, and D. Che, ”Urdu language processing: A survey”, Artif. Intell. Rev., vol. 47, no. 3, pp. 279-311, 2017

[3] B. Luo, R. Y. K. Lau, C. Li, and Y.-W. Si, “A critical review of state-of-the-art chatbot designs and applications,” Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 12, no. 1, 2022.

[4] A. S. Lokman and M. A. Ameedeen, “Modern Chatbot Systems: A Technical Review,” in Proceedings of the Future Technologies Conference (FTC) 2018, Cham: Springer International Publishing, pp. 1012–1023, 2019.

[5] E. Adamopoulou and L. Moussiades, “Chatbots: History, technology, and applications,” Machine Learning with Applications, vol. 2, no. 100006, p. 100006, 2020.

[6] A. M. Turing, “Computing Machinery and Intelligence,” in Parsing the Turing Test, Dordrecht: Springer Netherlands, pp. 23–65, 2009.

[7] J. Weizenbaum, “ELIZA - a computer program for the study of natural language communication between man and machine,” Communications of the ACM, vol. 26, no. 1, pp. 23–28, Jan. 1983.

[8] S. Reshmi and K. Balakrishnan, “Implementation of an inquisitive chatbot for database supported knowledge bases,” Sādhanā, vol. 41, no. 10, pp. 1173–1178, Oct. 2016.

[9] R. Wallace. ”The anatomy of ALICE.”, Parsing the turing test, Springer, Dordrecht, pp. 181-210, 2009.

[10] S. Singh and H. Beniwal, “A survey on near-human conversational agents,” Journal of King Saud University - Computer and Information Sciences, Nov. 2021.

[11] G. Caldarini, S. Jaf, and K. McGarry, “A Literature Survey of Recent Advances in Chatbots,” Information, vol. 13, no. 1, p. 41, Jan. 2022.

[12] H. Al-Zubaide and A. A. Issa, ”OntBot: Ontology based chatbot,” International Symposium on Innovations in Information and Communications Technology, pp. 7-12, 2011.

[13] R. Agarwal and M. Wadhwa, “Review of State-of-the-Art Design Techniques for Chatbots,” SN Computer Science, vol. 1, no. 5, Jul. 2020.

[14] H. Al-Zubaide and A. A. Issa, ”OntBot: Ontology based chatbot,” International Symposium on Innovations in Information and Communications Technology, pp. 7-12, 2011.

[15] I. Sutskever, O. Vinyals, and Q. V. Le, ”Sequence to sequence learning with neural networks,” Advances in neural information processing systems, 27, 2014.

[16] O. Vinyals and Q. Le, “A Neural Conversational Model,” Proceedings of the 31st International Conference on Machine Learning, Lille, France, volume 37, Jul. 2015.

[17] M. Boussakssou, H. Ezzikouri, and M. Erritali, “Chatbot in Arabic language using seq to seq model,” Multimedia Tools and Applications, Nov. 2021.

[18] S. Mathur and D. Lopez, ”A scaled-down neural conversational model for chatbots,” Concurrency and Computation: Practice and Experience, 31, 2019.

[19] K. Palasundram, N. Mohd Sharef, N. A. Nasharuddin, K. A. Kasmiran, and A. Azman, “Sequence to Sequence Model Performance for Education Chatbot,” International Journal of Emerging Technologies in Learning (iJET), vol. 14, no. 24, p. 56, Dec. 2019.

[20] J. Kapočiūtė-Dzikienė, “A Domain-Specific Generative Chatbot Trained from Little Data,” Applied Sciences, vol. 10, no. 7, p. 2221, Mar. 2020.

[21] N. N. Khin and K. M. Soe, ”Question Answering based University Chatbot using Sequence to Sequence Model,” 2020 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), pp. 55-59, 2020.

[22] Y. Zhang, T. Xu and Y. Dai, ”Research on Chatbots for Open Domain: Using BiLSTM and Sequence to Sequence,” Proceedings of the 2019 International Conference on Artificial Intelligence and Computer Science, pp. 145-149, July, 2019.

[23] D. Bahdanau, K. Cho, and Y. Bengio, ”Neural machine translation by jointly learning to align and translate,” Technical report, arXiv preprint arXiv:1409.0473, 2014.

[24] T. Luong, H. Pham, and C. D. Manning, “Effective approaches to attention-based neural machine translation,” in Proc. Conf. Empirical Methods Natural Lang. Process., pp. 1412–1421, 2015.

[25] S. S. Abdullahi, S. Yiming, A. Abdullahi, and U. Aliyu, “Open domain chatbot based on attentive end-to-end Seq2Seq mechanism,” ACM International Conference Proceeding Series, pp. 339–344, 2019.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, ”Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., pp. 5998–6008, 2017.

[27] Z. Niu, G. Zhong, and H. Yu, “A review on the attention mechanism of deep learning,” Neurocomputing, vol. 452, pp. 48–62, Sep. 2021.

[28] M. Hardalov, I. Koychev, and P. Nakov, ”Towards automated customer support,” in Proc. Int. Conf. Artif. Intell. Methodol. Syst. Appl., pp. 48–59, 2018.

[29] A. K. M. Masum, S. Abujar, S. Akter, N. J. Ria and S. A. Hossain, ”Transformer Based Bengali Chatbot Using General Knowledge Dataset,” 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 1235-1238, 2021.

[30] M. Kaleem, J. O’Shea, and K. Crockett, ”Development of UMAIR the Urdu Conversational Agent for Customer Service,” e, in Proceedings of the World Congress on Engineering, 2014.

[31] J. Shabbir, M. U. Arshad and W. Shahzad, “NUBOT: Embedded Knowledge Graph With RASA Framework for Generating Semantic Intents Responses in Roman Urdu,” February 2021. [Online]

[32] M. Alam, ”Neural Encoder-Decoder based Urdu Conversational Agent,” 2018 9th IEEE Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON), pp. 901-905, 2018.

[33] A. Nawaz, M. Bakhtyar, J. Baber, I. Ullah, W. Noor, and A. Basit, ”Extractive text summarization models for Urdu language,” Inf. Process. Manage., vol. 57, no. 6, Nov. 2020.

[34] K. Riaz, ”Comparison of Hindi and Urdu in computational context” in Int. J. Comput. Linguist. Nat. Lang. Process., vol. 1, no. 3, pp. 92-97. 2012.

[35] A. Daud, W. Khan, and D. Che, ”Urdu language processing: A survey,” Artif. Intell. Rev., vol. 47, no. 3, pp. 279–311, 2017.

[36] T. Iqbal and S. Qureshi, “The survey: Text generation models in deep learning,” Journal of King Saud University - Computer and Information Sciences, Apr. 2020

[37] A. Sherstinsky, “Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network,” Physica D: Nonlinear Phenomena, vol. 404, p. 132306, Mar. 2020.

[38] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[39] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. and Kuksa, P., ”Natural language processing” in Journal of Machine Learning Research, 2011.

[40] Mikolov, T., Chen, K., Corrado, G. and Dean, J., ”Efficient estimation of word representations in vector space” in Proceedings of International Conference on Learning Representations (ICLR), 2013.

[41] Sajadul Hassan Kumhar, Mudasir M. Kirmani, Jitendra Sheetlani, Mudasir Hassan, ”Word Embedding Generation for Urdu Language using Word2vec model” in Materials Today: Proceedings, 2021.

[42] H. Sharma, ”Understanding Encoders-Decoders with an Attention-based mechanism,” https://medium.com/data-science-community-srm/understanding-encoders-decoders-with-attention-based-mechanism-c1eb7164c581 (accessed June. 5, 2022).

[43] Haşim Sak, Andrew Senior and Françoise Beaufays, ”Long short-term memory recurrent neural network architectures for large scale acoustic modeling” in Fifteenth annual conference of the international speech communication association, 2014.