We are IntechOpen,
the world's leading publisher of Open Access books
Built by scientists, for scientists

6,700
Open access books available

180,000
International authors and editors

195M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS
BOOK CITATION INDEX
INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Chapter

Recent Trends in Deep Learning for Conversational AI

Jyotsna Talreja Wassan and Veena Ghuriani

Abstract

Conversational AI has seen unprecedented growth in recent years due to which Chatbots have been made available. Conversational AI primarily focuses on text or speech inputs, identifying the intention behind them, and responding to users with relevant information. Natural Language Processing (NLP), Natural Language Understanding (NLU), Machine Learning (ML), and speech recognition offer a personalized experience that mimics human-like engagement in conversational AI systems. Conversational AI systems like Google Meena, Amazon's Alexa, Facebook's BlenderBot, and OpenAI's GPT-3 are trained using Deep Learning (DL) techniques that mimic a human brain-like structure and are trained on huge amounts of text data to provide open-domain conversations. The aim of this chapter is to highlight Conversational AI and NLP techniques behind it. The chapter focuses on DL architectures useful in building Conversational AI systems. The chapter discusses what are the recent advances in Conversational AI and how they are useful, what are the challenges, and what is the scope and future of conversational AI. This will help researchers to understand state-of-the-art frameworks and how they are useful in building Conversational AI models.

Keywords: conversational AI, deep learning, convolution neural networks, natural language processing (NLP), recurrent neural networks

1. Introduction

Conversational AI is a sub-domain of Artificial Intelligence (AI) that dwells primarily on speech-based or text-based AI agents that tend to simulate and automate conversations and interactions [1]. The use of Conversational AI agents like chatbots and voice assistants has proliferated in today's world [2]. The tremendous growth in the area of Conversational AI has revolutionized the way in which humans interact with machines. A conversational agent is built upon DL architectures making use of multiple-layer neural networks to learn from data and make decisions or predictions. The generative deep models are capable of learning the underlying distribution of the training data and then generating new samples that share similar characteristics. Generative models in deep learning are useful in tasks such as image synthesis, text generation, and audio generation [3]. They learn the statistical properties of the training data and use that knowledge to generate new samples that are not explicitly present in the training set. The generative models used in

conversational agents are primarily based on recurrent neural networks (RNNs) and transformer architectures [4]. The chapter aims to cover in detail the working of these architectures in conversational AI leading readers to understand large multimodal models that are working behind the scenes for the latest chatbot "ChatGPT" developed by OpenAI [5]. The advent of deep learning (DL) has led to substantial developments in this area of Conversational AI and the goal of this chapter is to familiarize the readers and research community with the recent advances in Natural Language Processing (NLP) techniques with DL aiding conversational AI. Due to NLP and DL and their design architecture, conversational agents have progressed in varied applications like healthcare, customer care, education, etc. This rise in the practical implementation and their demand has in turn made Conversational AI a ripe area for innovation and novel research. The chapter provides the research background, details on NLP and DL technologies for Conversational AI, available resources, and key insights to the application of NLP and DL in conversational AI systems. Finally, future work, outstanding challenges, and current applications are presented in this chapter.

2. What is natural language processing (NLP)?

Natural language processing (NLP) is the field of "artificial intelligence" (AI) that is concerned with providing computers the capacity to comprehend written and spoken words in a manner similar to that of humans [6]. Computational linguistics, or the rule-based modeling of human language, is combined with statistical, machine learning, and deep learning models to form NLP. These technologies work together to provide computers the ability to comprehend human language in the form of text or speech data and to "understand" its full meaning, including the speaker's or writer's intention and sentiment.

Computer programs that translate text between languages, reply to spoken commands, and quickly summarize vast amounts of text—even in real-time—are all powered by NLP. NLP has applications in the form of voice-activated GPS devices, digital assistants, speech-to-text dictation programs, customer service chatbots, and other consumer conveniences. The use of NLP in corporate solutions, however, is expanding as a means of streamlining business operations, boosting worker productivity, and streamlining mission-critical business procedures. NLP has two overlapping subfields: Natural Language Understanding (NLU), which focuses on semantic analysis or determining the meaning of text, and Natural Language Generation (NLG), which handles text generation by a machine [6].

2.1 How does natural language processing (NLP) work?

NLP models search for relationships between the letters, words, and sentences present in the text. NLP architectures employ a variety of methods for data preprocessing, feature extraction, and modeling. Some of these methods are as follows.

1. **Data Pre-processing**: It is frequently necessary to preprocess text before a model process it for a particular job in order to enhance model performance or to convert words and characters into a format the model can comprehend. Data preparation is prioritized by the emerging field of data-centric AI. Some of the techniques used for pre-processing are:

- Stemming and Lemmatization: Both are pre-processing techniques to convert inflicted words to their base word.

- Stop word removal: It is a technique of removing the most commonly occurring words which do not add much meaning to the text, for example, "a", "an", "and", etc.

- Tokenization: It is a process that divides the text into words and word fragments. Typically, the outcome is a word index and tokenized text, where words can be represented as numerical tokens for use in different DL techniques.

2. **Feature extraction**: The machine learning algorithms learn from a predefined set of features from the training data. Features are typically numbers that describe a document in relation to the corpus that contains it. However, the primary issue with language processing is that machine learning techniques cannot be used to directly handle raw text. So, to turn text into a matrix (or vector) of features, we need certain feature extraction techniques. Some of the techniques for feature extraction are Bag-of-Words, TF-IDF, or generic feature engineering; these rely on document length, word polarity, and metadata.

   - **Bag-of-Words**: Bag-of-Words maintains a count of each word or n-gram(combination of n words) which appears in a document. It creates a numerical representation of the dataset based on how many times each word appears in the document.

   - **TF-IDF**: Term Frequency – Inverse document Frequency technique is frequently employed in information retrieval and natural language processing. It gauges a word's significance inside a document in relation to a corpus, or group of documents. A text vectorization procedure converts words in a text document into significance numbers.

Term Frequency (TF) of a term or word is the number of times the term appears in a document compared to the word corpus of the document.
TF (word in a document) = Number of occurrences of that word in document/Number of words in document
Inverse Document Frequency (IDF) of a term reflects the proportion of documents in the corpus that contain the term.
IDF (word in a corpus) = log(number of documents in the corpus/number of documents that include the word)
The *TF-IDF* score of a term is the product of TF and *IDF*.
Numerous applications of natural language processing benefit from TF-IDF. For instance, search Engines rank the relevancy of a document for a query using TF-IDF. Text classification, text summarization, and topic modeling are more applications for TF-IDF.

   - **Word2Vec**: It takes raw text and turns it into high-dimensional word embeddings using a standard neural network. It has two variations: *Continuous Bag-of-Words* (CBOW), which tries to predict the target word from surrounding words, and *Skip-Gram*, in which we attempt to predict surrounding words given a target word. These models accept a word as input and output a word

embedding that can be used as an input to numerous NLP tasks after deleting the final layer after training. Word2Vec embeddings capture context. Words will have comparable embeddings if they arise in related contexts.

3. **Modeling**: After data is preprocessed, it is fed into an NLP architecture that accomplishes a variety of tasks.

- The feature extracted using the aforementioned methods can be input into a variety of models. The output of the TF-IDF vectorizer, for instance, could be fed to logistic regression, naive Bayes, decision trees, or gradient boosted trees for classification.

- Deep Neural Networks/ Deep Learning models can be employed.

2.2 ML-based natural language processing (NLP) modelling techniques

2.2.1 Logistic regression

It is a statistical analysis that builds a statistical model to explain the link between a group of independent predictor or explanatory factors and a binary or dichotomous (yes/no type) result (dependent or response variable) [7].

In order to explore the effects of predictor variables on categorical outcomes, logistic regression models are utilized. When the outcome is binary, such as the presence or absence of a disease (such as non-Hodgkin's lymphoma), the model is referred to as a binary logistic model. The model is known as a multiple or multivariable logistic regression model when there are several variables (for example, risk factors and treatments), and it is one of the most often used statistical models in medical publications. This chapter looks at categorical, continuous, and multiple binary logistic regression models, as well as interaction, quality of fit, categorical predictor variables, and multiple predictor variables [8].

2.2.2 Naive bayes

Naive Bayes is a supervised machine learning algorithm. It is a simple probability classifier, which determines a set of probabilities by counting the frequency and combinations of values in a given data set. It is based on the value of the class variable and the Bayes' theorem [9], the algorithm assumes that all variables are independent. Since conditional independence is rarely true in practical applications, it is considered naive. The method has been shown to learn quickly in a variety of controlled classification challenges. The theorem allows us to "invert" conditional probabilities as shown in Eq. (1).

$$P(A|B) = P(B|A)^* P(A) / P(B) \qquad (1)$$

Here;
P(A|B) is the probability of the occurrence of event A when event B occurs,
P(A) is the probability of the occurrence of A,
P(B|A) is the probability of the occurrence of event B when event A occurs,
P(B) is the probability of the occurrence of B.

It also belongs to the family of generative learning algorithms [9, 10], which model the input distribution of a certain class or category. In a Naive Bayes model, it is assumed that predictors are conditionally independent of one another or disconnected to any other model features. It also presumes that each attribute affects the outcome equally. Although these presumptions are generally not true in real-world situations, they make categorization problems more manageable from a computational standpoint. In other words, each variable will now only need one probability, simplifying the computation of the model. The classification method performs effectively despite this irrational independence assumption, especially with small data sets.

2.2.3 Decision trees

It is a supervised ML method used for classification and regression. It creates a model that predicts the value of a target variable with the help of simple decision rules which are inferred from the data features. It displays the predictions that come from a sequence of feature-based splits using a flowchart that resembles a tree structure [11]. They are used in different fields such as image processing, and identification of patterns. They are based on the idea of "divide and conquer" to develop learning machines that learn from prior knowledge to make a vast number of decisions. These decisions help in predicting outcomes for problems for predictive modeling.

2.2.4 Hidden markov model

Hidden Markov Models (HMMs) [12, 13] are an extension of Markov chains. A Markov chain is a stochastic model used to describe a system that transitions from one state to another over discrete time steps. The key idea is that the future state of the system depends only on its current state and is independent of previous states. This property is called the Markov property. An HMM has two sets of components [13]:

- Hidden States (X): These are the unobservable, underlying states of the system. Think of them as the "true" states that you are interested in modeling. For example, in speech recognition, hidden states might represent phonemes or words.

- Observations (Y): These are the observable data or measurements associated with each hidden state. Observations are what you can directly measure or record. In speech recognition, observations could be acoustic features like spectrogram frames.

HMMs incorporate probabilities of transitioning from one hidden state to another at each time step. These probabilities are represented by a transition matrix. It defines the likelihood of moving from one state to another. Each hidden state is associated with a probability distribution over possible observation. This distribution is known as the emission probability distribution. It describes how likely a particular observation is given the current hidden state. HMMs also have an initial state distribution, which describes the probability of starting in each possible hidden state.

Given an HMM and a sequence of observations, you can perform various tasks, filtering, prediction, smoothing and parameter estimation. HMMs are applied in a wide range of domains, including Speech Recognition, Natural Language Processing, Bioinformatics, Finance, Robotics, etc. Hidden Markov Models are versatile tools for

modeling sequential data with hidden structures, making them valuable in numerous fields where understanding temporal dependencies is crucial.

2.2.5 Convolutional neural networks (CNN)

Convolutional Neural Networks (CNN) have contributed a lot in the field of computer vision and image analysis, it minimizes human effort by automatically detecting the features [14]. CNNs are a class of Deep Neural Networks that can recognize and classify particular features. Convolution is a mathematical process that involves multiplying two functions to create a third function that expresses how the form of one function is altered by the other. The term "convolution" is used in CNN to refer to this mathematical activity [15].

CNN uses convolution layers to apply convolution operations to the input data to extract features. After convolution, pooling layers are used to reduce the spatial dimensions of the data. It involves down sampling the feature maps created by the convolutional layers, by taking max-pooling and average-pooling. It helps to reduce the computational complexity. After convolutional and pooling layers, activation functions are applied to introduce non-linearity into the network [16]. One or more fully connected layers are present at the end of the network. These layers take the high-level features learned by the previous layers and use them for tasks like classification or regression. While RNNs and transformer models have gained more prominence in recent years for NLP problems due to their ability to capture sequential dependencies. CNNs are useful for certain NLP applications dealing with smaller datasets or when a simpler yet efficient architecture is desired [17].

3. Conversational AI

Conversational AI refers to a technology that enables computers or machines to engage in natural, human-like conversations with users. It is a subdomain of artificial intelligence that enables computers to understand, process, and generate human language. It combines the principles of NLP and ML [6].

Large volumes of text and speech are used to train conversational AI systems. The machine is taught how to comprehend and process human language using this data. The technology then uses this information to communicate with people in a natural way. Through repeated learning from its interactions, it gradually raises the quality of its responses. There are numerous examples of conversational AI applications that showcase its versatility and utility across various domains. Listed below are a few notable examples for reference and applicability [18, 19].

- **Siri**: Apple's virtual assistant, Siri, is one of the pioneering examples of conversational AI. Siri can answer questions, set reminders, send messages, and perform a range of tasks using natural language interactions.

- **Google Assistant**: Google's counterpart to Siri, Google Assistant, is integrated into Android devices and other Google products. It can provide information, control smart devices, set alarms, and even engage in light-hearted conversations.

- **Amazon Alexa**: Alexa is Amazon's virtual assistant, found in devices like Echo speakers. It can play music, provide weather updates, order products

from Amazon, and integrate with various third-party services through voice commands.

- **IBM Watson Assistant**: Watson Assistant is a platform that allows businesses to create and deploy chatbots and virtual assistants tailored to their needs. It's widely used for customer support, helping users find information, and handling routine tasks.

- **Microsoft Bot Framework**: Microsoft's Bot Framework enables developers to build conversational AI bots for various platforms like Microsoft Teams, Skype, and more. These bots can provide customer support, automate tasks, and enhance user engagement.

- **ChatGPT**: ChatGPT, which you are currently interacting with, is an example of conversational AI developed by OpenAI. It can engage in text-based conversations on a wide range of topics, answer questions, and generate human-like responses.

- **Duolingo Chatbots**: Language learning app Duolingo uses conversational AI chatbots to help users practice speaking in different languages. The chatbots simulate real conversations to improve language skills.

- **Woebot**: Woebot is a mental health chatbot designed to provide emotional support and cognitive behavioral therapy techniques to users. It engages in conversations to monitor users' mental well-being and offer appropriate interventions.

- **Replika**: Replika is an AI chatbot designed to engage users in meaningful conversations and provide emotional support. It aims to create a virtual companion that users can interact with on a personal level.

- **LivePerson**: LivePerson provides conversational AI solutions for businesses to engage with customers through chatbots and messaging apps. It's often used for online sales, customer support, and lead generation.

3.1 Components of conversational AI

Conversational AI has the following key components [20, 21].

1. **Machine learning**: The first is machine learning (ML), a subfield of artificial intelligence that makes predictions by mining huge data sets for patterns using a variety of intricate statistical models and algorithms. Any conversation AI engine's performance depends on machine learning (ML), which gives the system the ability to continuously learn from the data it collects and improve its understanding of and replies to human language.

2. **Natural Language Processing**: The second element is Natural Language Processing (NLP), which entails converting unstructured input into a machine-readable format and processing it to generate the right response. Conversation AI engines depend on NLP since it enables the system to understand human input and generate pertinent responses. Input generation, input analysis, syntactic and semantic analysis, and output transformation make up its five key activities. NLP includes the following main stages:

a. Lexical analysis: The process of separating a text into tokens, or individual words or phrases.

b. Syntactic and semantic analysis: Understanding the relationships between words and phrases in a sentence and deciphering the text's meaning.

c. Sentiment Analysis: It is the analysis of a text's emotional tone, such as whether it is favorable, negative, or neutral.

d. Output transformation: Creating human-like language from a given input or environment.

3. **Data Mining**: Data mining, the third component, is utilized in conversation AI engines to extract patterns and insights from conversational data that programmers can use to improve the operation of the system. In contrast to machine learning, which focuses on making predictions based on current data, it is a technique for finding unidentified traits.

4. **Automatic Speech Recognition (ASR)**: The last part of AI for voice-based dialogs is ASR. It makes it possible for conversation AI engines to recognize human voice inputs, remove background noise, derive the question using speech-to-text, and mimic a human-like response. Natural language conversations and guided dialog are the two categories of ASR software.

4. Deep learning models for conversational AI

Conversational AI is a subfield of Artificial Intelligence driven by speech and text-based agents that automate verbal communication. Agents like chatbots and voice assistants have gained popularity due to tremendous advancements in Machine learning methods such as Deep Learning (DL) and the availability of higher computing power hardware such as GPUs and TPUs [22]. NLP, with DL, has made possible the applicability of conversational AI in a variety of fields such as education, online management of businesses, healthcare, customer care, etc. Natural language understanding (NLU) is a sub-field of NLP and is useful in understanding input made in the form of unstructured text or speech. NLU mainly consists of two tasks – Named Entity Recognition (NER) and Intent Classification (IC) [23]. Natural language Generation (NLG unit) is also one of the major components of conversational agent architecture that uses advanced DL techniques to transform data insights into automated informative narratives. Newer Conversational AI architectures involving DL are progressing at a very high rate [24]. Pre-trained deep ML models have been increasingly used in conversational agents [25]. This section is intended to highlight the latest research in Conversational AI architecture developments.

Conversational agents are primarily based on deep learning techniques. These methods provide responses to user queries based on syntax, structure, and words (Vocabulary) in the input along with understanding the contextual information. Conversational agents are built upon DL methods of neural networks involving Recurrent Neural networks (RNN) [26], Bi-LSTM [4], and pre-trained models like BERT [27] and (**Figure 1**) GPT [3, 4, 20].

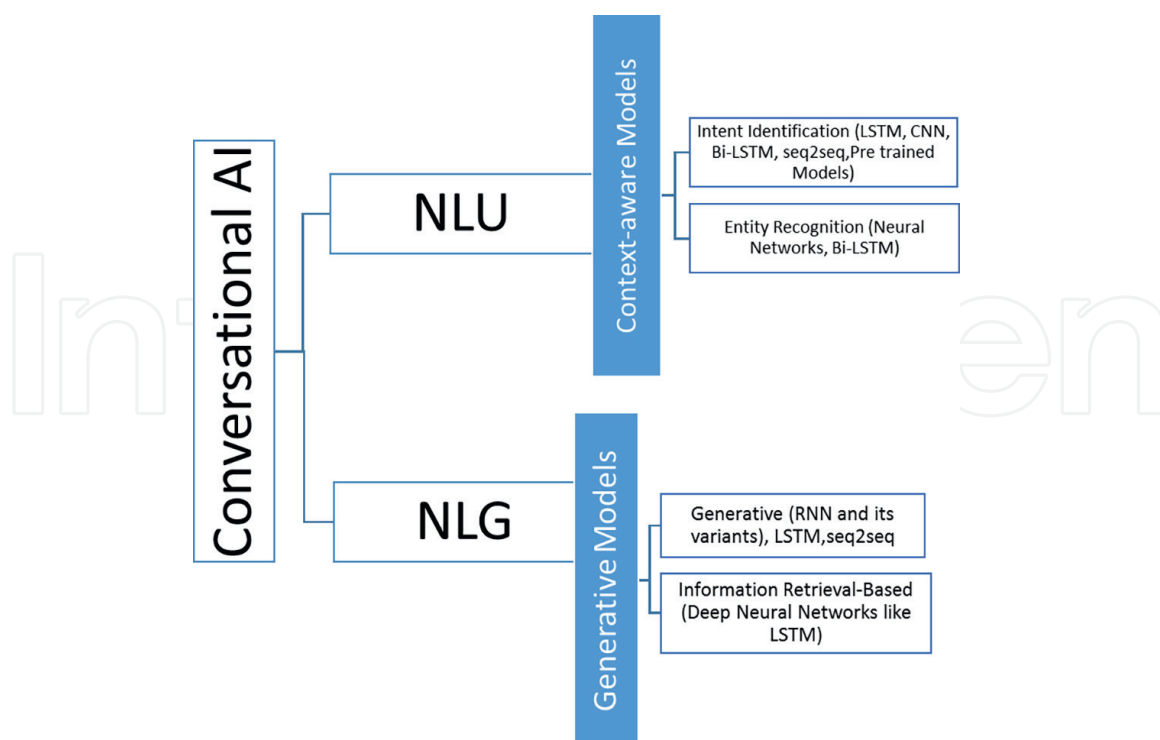The most prevalent DL architectures in Conversational AI are listed below [28].

Figure 1.
DL methods in conversational AI.

4.1 Recurrent neural networks (RNNs)

Recurrent neural networks (RNNs) are a type of neural networks suitable for processing sequential data, such as natural language text or time series data using feedback from previous iterations [26]. RNNs have the form of a chain of repeating modules of neural networks. The repeating module has a simple structure, such as a single tanh layer [29]. RNNs remember the previous computations and use this understanding of previous information in current processing. They can be used to handle a variety of tasks, including customer service, information retrieval, and language translation. RNNs also consist of the nodes representing the "Neurons" of the network. The neurons are spread over the temporal scale (i.e. sequence) and separated into three layers – (i) input layer indicating input to be processed; a hidden layer representing the algorithm used for problem-solving and the output layer show-ing the result of the operation. The hidden layer contains a temporal feedback loop as shown in **Figure 2** [29]. Particularly in conversation AI, sound waves are recognized into phonetic segments and subsequently joined together into words via the RNN application.

4.2 Long short-term memory networks (LSTM)

LSTM is a type of RNN that is designed by Hochreiter and Schmid Huber [30] to capture the memories of previous inputs of the input sequence. LSTM are useful in processing large sequences. Since it captures long-term dependencies, it facilitates the sequence prediction tasks. It is applied in conversational AI such as in tasks to predict the next word in the input sequence. This was the stepping stone to imbibe the functionality of remembering the conversation from the previous inputs in a dialog system [31].
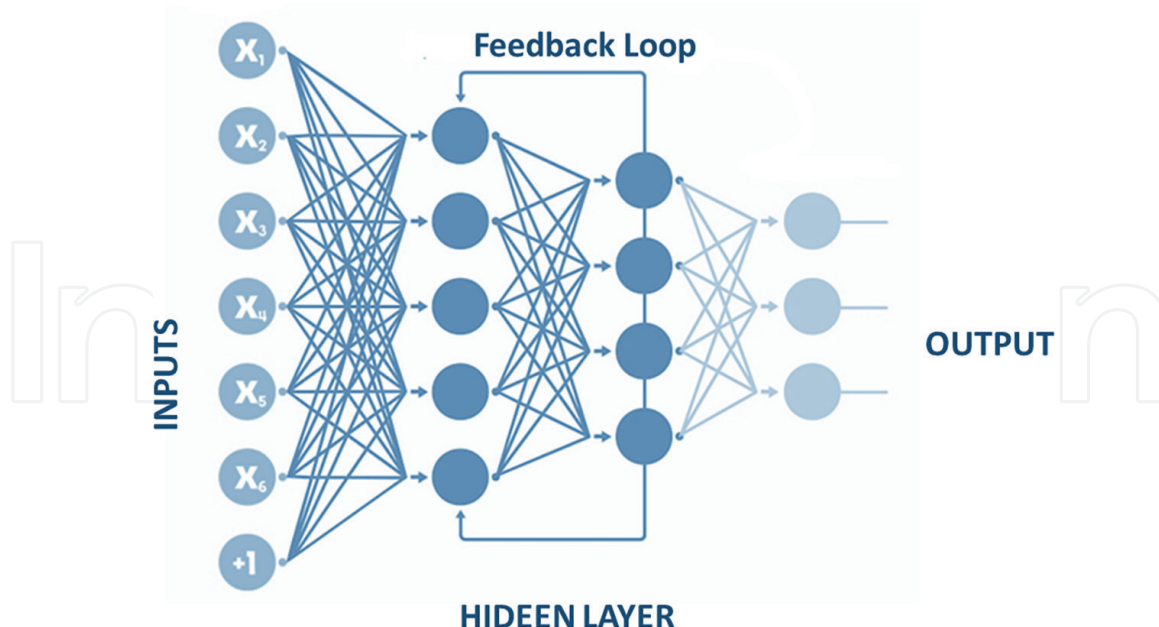
Figure 2.
The RNN architecture.

It works by controlling the Input Gate, Output Gate, Memory Cell, and Forget Gate in the system to successfully process and predict a significant sequence of events and any delays involved. LSTMs address long-term dependencies by introducing a memory cell, which contains information for an extended period of time. The first step in an LSTM model is to decide what information needs to get stored or thrown away in the memory cell state. A sigmoid layer (σ) called the "input gate layer" decides which values need to be stored and updated in a memory cell [30]. Next, a tanh layer creates a vector of new candidate values. The input gate controls what information is added to the memory cell. The forget gate controls what information is removed from the memory cell [30]. And the output gate controls what information is output from the memory cell. This facilitates LSTM networks to selectively retain or discard information as it flows through the network. The basic structure of LSTM is shown in **Figure 3** [31].

LSTMs have facilitated the generation of coherent and grammatically correct sentences in NLP models by learning the dependencies between words in a sentence are also able to recognize patterns in speech, time predictability, detect anomalies in varied fields, to analyze video data, and extract useful information and can aid in learning patterns in user behavior and use them to make personalized recommendations.

4.3 Sequence to sequence (Seq2Seq) neural models

The sequence-to-sequence (Seq2Seq) model was proposed by Sutskever, Le et al. [32] This model utilizes the LSTM structure and encoder-decoder structure (**Figure 4**) to process the conversation data. Seq2Seq is primarily used for tasks such as machine translation, text summarization, and image captioning in conversational AI [33]. The model consists of two main components: an encoder and a decoder (**Figure 4**) [32]. The encoder and decoder are typically implemented as RNNs. Seq2Seq models are trained on input sequences (words, letters, time series, etc.) to maximize the likelihood of the prediction of the correct output sequence. The encoder uses a hidden state/context vector of the input sequence and sends it
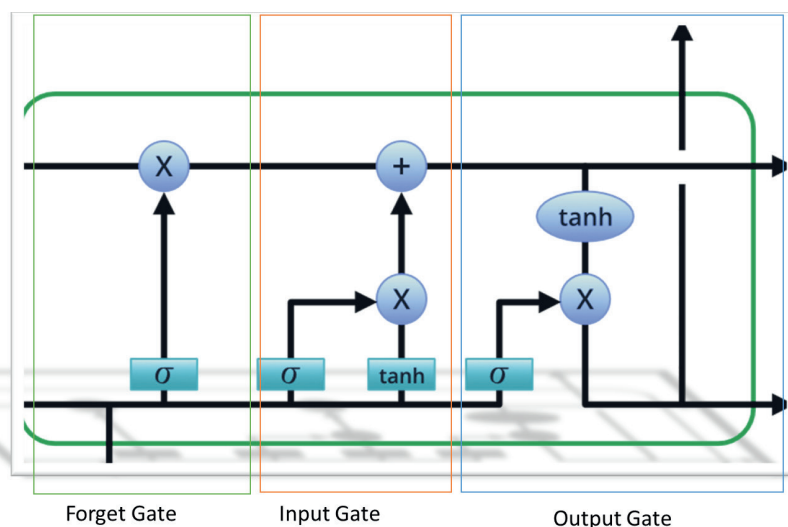
Figure 3.
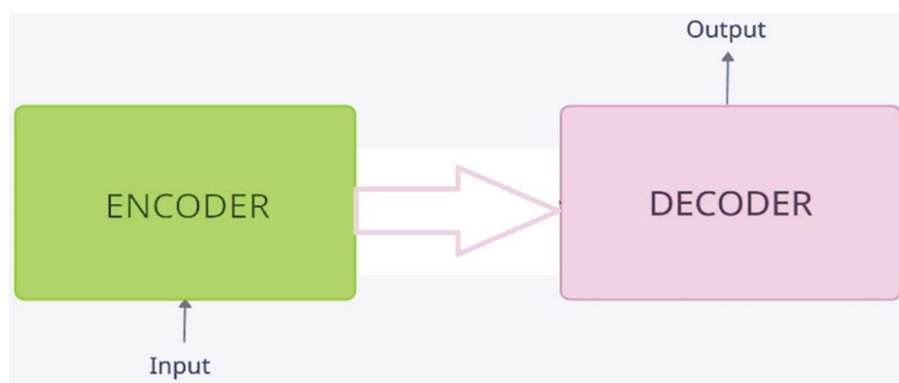*The basic structure of an LSTM model.*

Figure 4.
The basic model behind Seq2Seq model.

to the decoder, which then produces the output sequence. The output at time step *t* depends on the current input as well as the input at time *t-1* suggesting the suitability of sequenced tasks [32]. The sequential information is preserved in a hidden state of the network and used in the next instance. Seq2Seq models can generate responses in natural language based on user inputs. Seq2Seq models have evolved over time, and different variations, such as attention mechanisms and Transformer-based architectures, have been developed to address some of their limitations. These improvements have led to significant advancements in the quality and performance of sequence-to-sequence tasks.

4.4 Reinforcement learning

Reinforcement Learning is a branch of ML that learns by trial and error and is driven by human intelligence to give machines knowledge and improve by rewarding or punishing [34]. There are four main components of Reinforcement Learning, which are as follows: 1. Policy 2. Reward Signal 3. Value Function 4. Model of the environment [35]. Reinforcement learning uses a framework driven by the interaction between a learning agent and its environment in terms of states, actions, and rewards. This framework is intended to be a simple way of representing essential features of

the artificial intelligence problem popular example of reinforcement learning is an online chess platform that decides a series of chess moves depending on the current state of the chess board. Win or lose will be identified as the reward in the use case. The use of reinforcement learning in Conversational AI is to learn patterns in text, speech, etc. via the trial-and-error method. This is possible through a dialog set predefined in machines. The reinforcement environment is majorly implemented as a Markov Decision Process (MDP) which relies on learning agents taking actions in a given environment and moving from one problem state to another [13, 34]. Each action is associated with a reward. The aim of a reinforcement learning agent is to collect as many rewards as possible by performing actions.

4.5 Generative pre-trained transformer (GPT)

Generative Pre-trained Transformer, often abbreviated as GPT, refers to a class of NLP models developed by OpenAI designed for varied tasks including text generation, text completion, translation, question answering, and more [36]. The "transformer" part of the name refers to the underlying architecture used in GPT models and has been introduced in the paper "Attention is All You Need" by Vaswani et al. [37]. The architecture uses a self-attention mechanism for process input sequences, enabling the model to capture long-range dependencies and contextual information effectively.

4.6 BERT model

Google's AI division recently developed a "Bidirectional Encoder Representations" model called BERT, based on NLP impacting various applications significantly. It has the ability to understand bidirectional context modeling aiding rich contextual information. BERT is pre-trained on a huge corpus of textual data and learning the
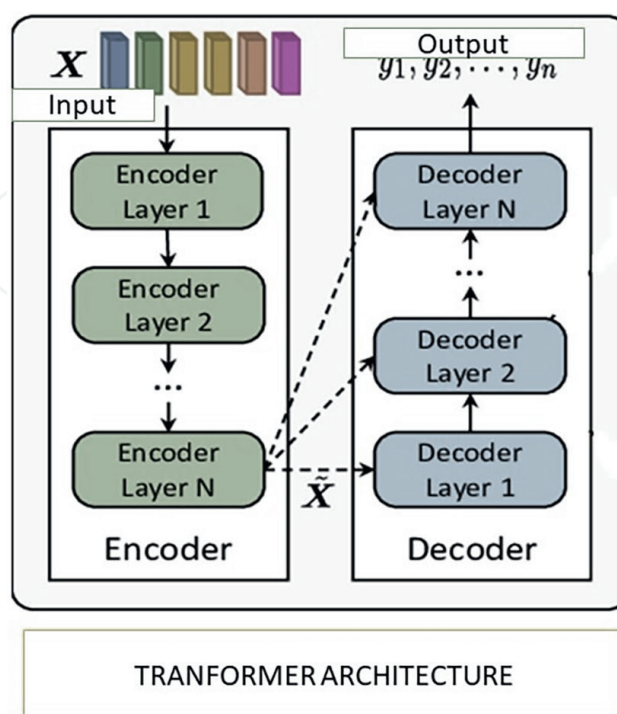

Figure 5.
BERT architecture.

relationships and meanings of words in context. After pretraining, BERT can be fine-tuned to perform text classification, named entity recognition, sentiment analysis, and question answering. BERT is a stacked Transformer's Encoder model as shown in **Figure 5**. A transformer architecture is an encoder-decoder network that uses self-attention on the encoder end. GPT (as discussed above) is not that different from BERT but is only a stacked Transformer's decoder model.

4.7 ChatGPT: a use case

ChatGPT is a natural language chatbot developed by OpenAI in 2022 [38]. It is an Artificial Intelligence tool with conversational capabilities. It can answer questions and help you compose emails, articles, notes, summarize books, etc. for various purposes. ChatGPT is based on a Generative Pre-trained Transformer (GPT), an AI language model developed by OpenAI [39]. The generative language models are trained on information available from the net, news articles, books, etc. The model is fine-tuned using supervised and reinforcement learning, this makes the model unique. Human feedback is incorporated in reinforcement learning so that the ChatGPT is able to provide responses to questions on the basis of context.

ChatGPT works in two phases, the first phase is called the pre-training phase and the second phase is the inference phase [40]. In the pre-training phase, the ChatGPT model is trained using supervised and unsupervised learning. In supervised learning, the overall model uses a labeled dataset and the inputs are accurately mapped to outputs. The techniques used in supervised learning are classification, regression, etc. There is a limitation here that the trainers cannot anticipate all the inputs (user questions) and outputs (responses), also there is a limit to the subject expertise. Hence, training would take longer. So, training ChatGPT using a supervised model is not feasible. ChatGPT there forth uses an unsupervised pre-training method [5]. In unsupervised pre-training, the model is trained on inputs that are not mapped to specific outputs. Here the model learns the inherent structure and pattern in the inputs, this helps the model to understand the syntax and semantics of the natural language to produce the meaningful responses in a conversational style. The model does not need to know the outputs associated with the inputs; it just uses these inputs for the sake of the pre-training phase. This is how the vast knowledge of ChatGPT is possible. This technique is called transformer-based language modeling [41].

5. Survey of NLP techniques in recent years

In early retrieval systems, TF-Id [42], bags of words, etc. were used as score functions in feature extraction. In recent years, deep learning has dominated a wide range of application fields. Deep learning methodologies are included in unsupervised learning. But it is also prevalent in supervised or semi-supervised learning. Deep learning classifiers enhance accuracy and performance by automatically learning and extracting information.

In Natural language understanding (NLU), entity identification and intent classification are two important phases. Named entity identification generally utilizes CNNs for fewer data pre-processing and for finding long-term dependencies, predicting entities, and generating feature matrix LSTM is preferred [30]. CNNs are used for modeling sentences [43], as they are good at extracting abstract and robust features from input. Conventional intent classification methods primarily employ supervised

machine learning algorithms such as Support Vector Machine [44], Decision Trees [45], and Hidden Markov Models [46]. Deep learning techniques as per [43, 47] used CNN, and RNN like LSTM to detect the intents from dialogs. Bi-LSTM [48], a variation of LSTM, is popular as it is able to process the input bidirectionally i.e., forward and backward directions. RNN and its different variants have some limitations. Long sequences of input tend to lose significant information since these models encode input into vectors of fixed length. It causes conversational agents to perform poorly. Transformer-based language models such as BERT [27]and GPT [36], overcome fixed-length limitations to utilize sentence-level recurrence and longer-term dependency. For natural language generation (NLG), the use of recurrent neural networks is common [33]. LSTM is used in Seq2Seq models for mapping input to a feature matrix followed by predicting tokens [49]. Seq2seq models have been combined with reinforcement learning to summarize text [34].

Machine learning for human-aligned conversational AI has been rigorously used. Bharti et al. [50] developed a Medbot for delivering telehealth after COVID-19 using NLP to provide free primary healthcare education and advice to chronic patients. The study introduces a novel computer application that acts as a personal virtual doctor using Natural Language Understanding (NLU) to understand the patient's query and response. NLP facilitated the study by reading, decoding, understanding, and making sense of human languages. Ashvini et al. [51] recently developed a dynamic NLP-enabled Chatbot for Rural Health Care in India. In another interesting study by Schlippe et al. [52] a multilingual interactive conversational artificial intelligence tutoring system is developed for exam preparation learning processes with NLP models. Conversational AI bots powered by NLP are also assisting farmers regarding all the intricacies of farming reducing costs significantly and increasing revenues. Reddy et al. [53] have developed Farmers Friend – A Conversational AI BoT for smart agriculture making use of NLP. NLP has the capability to automate the responses to customer queries in businesses. Olujimi [54] has shown NLP-based enhancement and AI in business ecosystems in their research. This indicates NLP is one of the fast-growing research domains in AI, with a variety of applications.

6. Advantages of conversational AI

Conversational AI has found its applications in various domains, chatbots, virtual assistants, and NLP systems offer several advantages. The key advantages are as follows.

i. **Improved Customer Service**: AI-driven chatbots provide quick and accurate answers to frequently asked questions, solving common issues without human intervention. This frees up employees to focus on more complex and value-added tasks. Also, the chatbots can operate round the clock, providing 24/7 support beyond business hours. This enhances customer satisfaction.

ii. **Scalability:** These systems can handle a large volume of conversations simultaneously, making them highly scalable without significant increases in costs.

iii. **Enhanced Productivity:** Conversational AI tools can assist in automating tasks like data retrieval, appointment scheduling, and information lookup, thereby boosting overall productivity. These tools are consistent in their responses and

information to the user, hence eliminating human error and uniform customer experience.

iv. **Personalization:** The conversational AI tools can analyze user data and preferences to deliver personalized recommendations, services, and content, providing an enhanced user experience. These tools can communicate in multiple languages, making them suitable for global markets and a diverse customer base.

v. **Accessibility:** Conversational AI can make services more accessible to individuals with disabilities, as they can interact using text or speech, and the system can be designed to accommodate various accessibility needs.

While conversational AI offers several advantages, human intervention may still be necessary for complex or sensitive issues. It may not be suitable for all situations and may require careful design and monitoring to ensure a positive user experience.

7. Applications of conversational AI

Conversational AI has primarily taken the form of advanced chatbots. The conventional chatbots have limited capabilities to handle limited functionality. They are featured on company websites and mobile applications to handle a predefined set of questions. Conversational AI chatbots are capable of answering frequently asked questions, troubleshooting issues and even conversing naturally [23]. They combine different AI technologies for more advanced capabilities. These technologies used in AI chatbots can also be used to enhance conventional voice assistants and virtual agents. Smart personal assistants (such as Amazon's Alexa, Apple's Siri, Google's Google Assistant, and Microsoft's Cortana) [55] are in trend as conversational agents that communicate with users through integrated voice in devices such as sound systems (speakers), smart devices like phones, watches, etc. and vehicles such as cars. Bai Li et al. [56] proposed a frame-based dialog management system, to search for and book hotels through text messaging. Their task-oriented chatbot system aided in searching for and booking hotels through text messaging. The demonstrative study indicated that the chatbot is a viable alternative to traditional mobile and web applications for commerce. Various studies [57–59] have indicated the usefulness of conversation AI in the educational domain. Conversational AI has been used in e-commerce websites to enhance businesses [60, 61]. **Table 1** indicates some of the prominent studies making use of chatbots in day-to-day life.

S. No.	Application	Reference
1	Education	[62, 63]
2	Healthcare	[64–67]
3	Food Dairy	[68]
4	Location Finder	[69]
5	Customer, business and Public Administration	[70, 71]

Table 1.
Some recent applications of conversational AI.

8. Challenges in conversational AI

One of the main challenges is "ML model selection that best suits the design algorithms" in Conversational AI. The unprecedented growth in NLP however has laid the path of choosing the most powerful NLP model with pretrained models such as BERT [27] or GPT [72]. In addition to this selection, data preprocessing demands a lot of pre-work in dialog systems to respond to a user's request in an efficient manner. It is also important to provide accurate service and information in dialog systems. For this purpose, data augmentation is used for translating raw data into the desired language. It faces the challenge of human language understanding and its Integration with media applications. It is important to choose the right chatbot development and deployment. When using chatbots in different domains, different challenges are being faced; Such as in the health domain chatbots need to respond without delays; in the educational domain, the chatbots should detect learners' progress effectively and should adjust to different difficulty levels. The chatbot's profile should dynamically be updated. Another challenge in conversational AI is the extraction and classification of useful data removing noisy patterns. This may affect the dialog systems as it may cause interference in different slots. Additionally, the challenges are being faced when the chatbot faces unscripted questions that it does not know how to answer and when it receives new and unplanned responses from the customers. To summarize, Conversational AI faces challenges as shown in **Figure 6**.

9. Future scope

LLM (Large Language Model) based chatbots like ChatGPT are the future and provide full open-ended support for human-like conversations and can perform varied tasks such as text summarization, paragraph writing, etc. LLMs can help in
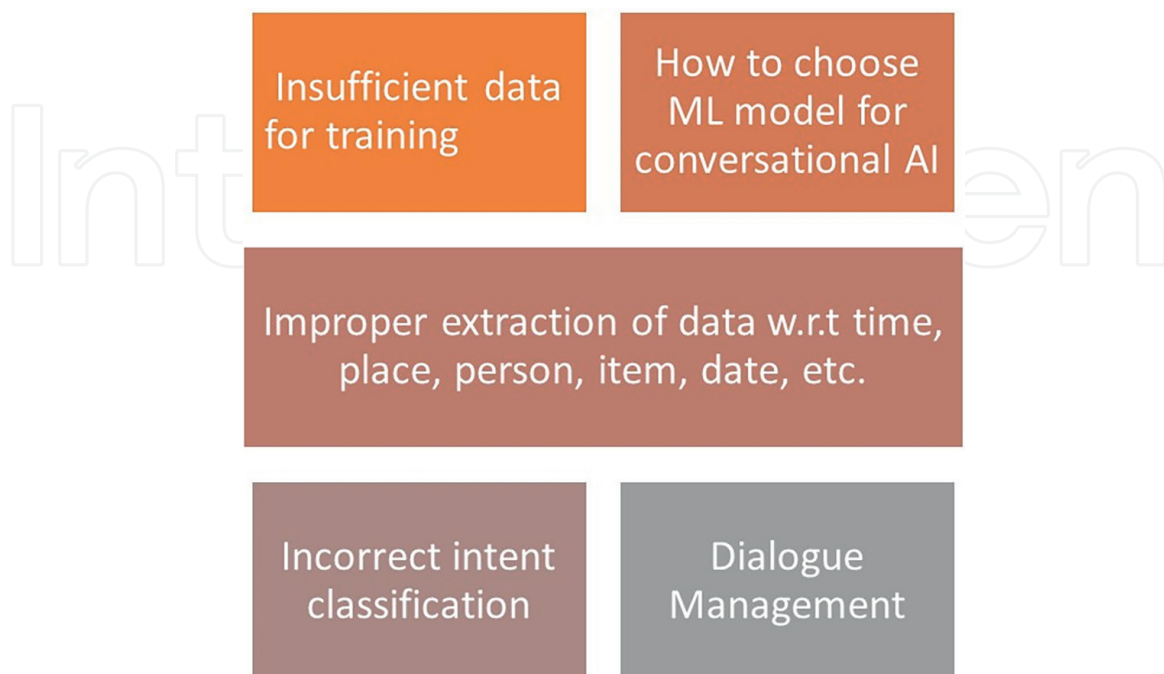

Figure 6.
Challenges in conversational AI.

understanding intents better and henceforth making better responses. One prominent example of a Large Language Model is GPT-3 developed recently, which stands for "Generative Pre-trained Transformer 3." GPT-3 was developed by OpenAI [73] and is known for its ability to perform a wide range of natural language processing tasks, such as text generation, language translation, text summarization, and more. With the latest advancements and continuous research in conversational AI, systems are getting better every day supporting personalized conversations and taking care of user engagement too. For example, if a bot finds the user is unhappy, it redirects the conversation to a real agent. The DL techniques like ChatGPT can automatically generate responses for queries using a knowledge base and are efficient. The future is the deeper integration of conversational AI with IoT devices and platforms. The rapid growth of the global conversational AI market size is projected to reach over $30 billion by 2030. Conversational AI serves as a cornerstone for growth in industries, and businesses, and also reshaping the way humans interact in the digital age.

Author details

Jyotsna Talreja Wassan and Veena Ghuriani*
Department of Computer Science, Maitreyi College, University of Delhi, Delhi, India

*Address all correspondence to: vghuriani@maitreyi.du.ac.in

IntechOpen

References

[1] Education, I. C. Conversational AI. Armonk: IBM Cloud Learn Hub [Internet]; 2017. August 31, 2020

[2] Chandra S, Shirish A, Srivastava SC. To be or not to be …human? Theorizing the role of human-like competencies in conversational artificial intelligence agents. Journal of Management Information Systems. 2022;**39**(4):969-1005. DOI: 10.1080/07421222.2022.2127441

[3] Johnson M. The primacy of data in deep learning NLP for conversational AI. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 2021. pp. 3-3

[4] Su PH, Mrkšić N, Casanueva I, Vulić I. Deep learning for conversational AI. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorial Abstracts. 2018. pp. 27-32

[5] Ray PP. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems. 2023;**3**:121-154. DOI: 10.1016/j.iotcps.2023.04.003

[6] Singh S, Mahmood A. The NLP cookbook: Modern recipes for transformer based deep learning architectures. IEEE Access. 2021;**9**:68675-68702. DOI: 10.1109/ACCESS.2021.3077350

[7] Sperandei S. Understanding logistic regression analysis. Biochemia Medica. 2014;**24**(1):12-18. DOI: 10.11613/BM.2014.003

[8] Zhang S, Zhang L, Qiu K, Lu Y, Cai B. Variable selection in logistic regression model. Chinese Journal of Electronics. 2015;**24**(4):813-817. DOI: 10.1049/cje.2015.10.025

[9] Bahri S, Saputra RA, Wajhillah R. Analisa sentimen berbasis natural languange processing (NLP) dengan naive bayes clasifier. Konferensi Nasional Ilmu Sosial & Teknologi. 2017;**1**(1)

[10] Okun O. Naïve bayes. In: Feature Selection and Ensemble Methods for Bioinformatics: Algorithmic Classification and Implementations. Information Science Reference-Imprint of: IGI Publishing; 2011

[11] Myles AJ, Feudale RN, Liu Y, Woody NA, Brown SD. An introduction to decision tree modeling. Journal of Chemometrics: A Journal of the Chemometrics Society. 2004;**18**(6):275-285

[12] Eddy SR. What is a hidden Markov model? Nature Biotechnology. 2004;**22**(10):1315-1316. DOI: 10.1038/nbt1004-1315

[13] Ghahramani Z. An introduction to hidden Markov models and Bayesian networks. International Journal of Pattern Recognition and Artificial Intelligence. 2001;**15**(1):9-42. DOI: 10.1142/S0218001401000836

[14] Wu J. Introduction to convolutional neural networks. In: National Key Lab for Novel Software Technology. Vol. 5, No. 23. China: Nanjing University; 2017. p. 495

[15] Brownlee J. A gentle introduction to pooling layers for convolutional neural networks. Machine Learning Mastery. 2019;**22**

[16] Basic Introduction to Convolutional Neural Network in Deep Learning. Analytics Vidhya; 2022. Available from: https://www.analyticsvidhya.com/blog/2022/03/basic-introduction-to-convolutional-neural-network-in-deep-learning/

[17] Kamath U, Liu J, Whitaker J. Deep Learning for NLP and Speech Recognition. Vol. 84. Cham, Switzerland: Springer; 2019

[18] McTear M. Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots. Springer Nature; 2022

[19] Caldarini G, Jaf S, McGarry K. A literature survey of recent advances in chatbots. Information. 2022;**13**(1):41. DOI: 10.3390/info13010041

[20] Lemon O. Conversational AI for multi-agent communication in natural language. AI Communications, (Preprint). 2022:1-14

[21] Ponnusamy P, Ghias AR, Yi Y, Yao B, Guo C, Sarikaya R. Feedback-based self-learning in large-scale conversational ai agents. AI Magazine;**42**(4):43-56

[22] Galitsky B, Galitsky B. Chatbot components and architectures. In: Developing Enterprise Chatbots: Learning Linguistic Structures. 2019. pp. 13-51

[23] Kulkarni P, Mahabaleshwarkar A, Kulkarni M, Sirsikar N, Gadgil K. Conversational AI: An overview of methodologies, applications & future scope. In: 2019 5th International Conference on Computing, Communication, Control and Automation (ICCUBEA). IEEE; 2019. pp. 1-7

[24] Samant RM, Bachute MR, Gite S, Kotecha K. Framework for deep learning-based language models using multi-task learning in natural language understanding: A systematic literature review and future directions. IEEE Access. 2022;**10**:17078-17097. DOI: 10.1109/ACCESS.2022.3149798

[25] Csaky R. Deep learning based chatbot models. arXiv preprint arXiv:1908.08835. 2019

[26] Banerjee I, Ling Y, Chen MC, Hasan SA, Langlotz CP, Moradzadeh N, et al. Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification. Artificial Intelligence in Medicine. 2019;**97**:79-88. DOI: 10.1016/j.artmed.2018.11.004

[27] Subakti A, Murfi H, Hariadi N. The performance of BERT as data representation of text clustering. Journal of Big Data. 2022;**9**(1):1-21. DOI: 10.1186/s40537-022-00564-9

[28] Hussain S, Ameri Sianaki O, Ababneh N. A survey on conversational agents/chatbots classification and design techniques. In: Web, Artificial Intelligence and Network Applications: Proceedings of the Workshops of the 33rd International Conference on Advanced Information Networking and Applications (WAINA-2019). Vol. 33. Springer International Publishing; 2019. pp. 946-956

[29] Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena. 2020;**404**:132306

[30] Gers FA, Schmidhuber J, Cummins F. Learning to forget: Continual prediction with LSTM. Neural Computation. 2000;**12**(10):2451-2471. DOI: 10.1162/089976600300015015

[31] Greff K, Srivastava RK, Koutnik J, Steunebrink BR, Schmidhuber J. LSTM: A search space odyssey. IEEE Transactions on Neural Networks and Learning System. 2016;**28**(10): 2222-2232. DOI: 10.1109/TNNLS. 2016.2582924

[32] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. Advances in Neural Information Processing Systems. 2014;**27**

[33] Gong G, An X, Mahato NK, Sun S, Chen S, Wen Y. Research on short-term load prediction based on Seq2seq model. Energies. 2019;**12**(16):3199. DOI: 10.3390/en12163199

[34] Matsuo Y, LeCun Y, Sahani M, Precup D, Silver D, Sugiyama M, et al. Deep learning, reinforcement learning, and world models. Neural Networks. 2022;**152**:267-275. DOI: 10.1016/j. neunet.2022.03.037

[35] Keneshloo Y, Shi T, Ramakrishnan N, Reddy CK. Deep reinforcement learning for sequence-to-sequence models. IEEE Transactions on Neural Networks and Learning System. 2019;**31**(7):2469-2489. DOI: 10.1109/TNNLS.2019.2929141

[36] Mhlanga D. The value of open AI and chat GPT for the current learning environments and the potential future uses. Available at SSRN 4439267

[37] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in Neural Information Processing Systems. Vol. 30. 2017

[38] Lambert J, Stevens M. ChatGPT and Generative AI Technology: A Mixed Bag of Concerns and New Opportunities. Computers in the Schools. 2023:1-25

[39] Stokel-Walker C, Van Noorden R. What ChatGPT and generative AI mean for science. Nature. 2023;**614**(7947):214-216. DOI: 10.1038/d41586-023-00340-6

[40] Wang FY, Miao Q, Li X, Wang X, Lin Y. What does ChatGPT say: The DAO from algorithmic intelligence to linguistic intelligence. IEEE/CAA Journal of Automatica Sinica. 2023;**10**(3):575-579. DOI: 10.1109/JAS.2023.123486

[41] Jalil S, Rafi S, Latoza TD, Moran K, Lam W. ChatGPT and software testing education: Promises & perils. In: 2023 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW). IEEE; 2023. pp. 4130-4137

[42] Towfighi S, Agarwal A, Mak DY, Verma A. Labelling chest x-ray reports using an open-source NLP and ML tool for text data binary classification. medRxiv. 2019:19012518

[43] Hu B, Lu Z, Li H, Chen Q. Convolutional neural network architectures for matching natural language sentences. Advances in Neural Information Processing Systems. 2014;**27**

[44] Rameshbhai CJ, Paulose J. Opinion mining on newspaper headlines using SVM and NLP. International Journal of Electrical and Computer Engineering (IJECE). 2019;**9**(3):2152-2163. DOI: 10.11591/ijece.v9i3.pp2152-2163

[45] Mendoza M, Zamora J. Building decision trees to identify the intent of a user query. In: International Conference on Knowledge-Based and Intelligent Information and Engineering Systems. Berlin, Heidelberg: Springer Berlin Heidelberg; 2009. pp. 285-292

[46] Cuayáhuitl H, Renals S, Lemon O, Shimodaira H. Human-computer dialogue simulation using hidden Markov models. In: IEEE Workshop on Automatic Speech Recognition and Understanding. IEEE; 2005. pp. 290-295

[47] Wen TH Gasic M, Kim D, Mrksic N, Su PH, Vandyke D, Young S. Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking. arXiv preprint arXiv:1508.01755

[48] Shafqat S, Majeed H, Javaid Q, Ahmad HF. Standard NER tagging scheme for big data healthcare analytics built on unified medical corpora. Journal of Artificial Intelligence and Technology. 2022;**2**(4):152-157. DOI: 10.37965/ jait.2022.0127

[49] Qiu M, Li FL, Wang S, Gao X, Chen Y, Zhao W, et al. AliMe chat: A sequence to sequence and rerank based chatbot engine. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2017. pp. 498-503

[50] Bharti U, Bajaj D, Batra H, Lalit S, Lalit S, Gangwani A. Medbot: Conversational artificial intelligence powered chatbot for delivering tele-health after COVID-19. In: 2020 5th International Conference on Communication and Electronics Systems (ICCES). IEEE; 2020. pp. 870-875

[51] Ashwini S, Rajalakshmi NR, Jayakumar L. Dynamic NLP enabled chatbot for rural health care in India. In: 2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA). IEEE; 2022. pp. 1-6

[52] Schlippe T, Sawatzki J. AI-based multilingual interactive exam preparation. In: Innovations in Learning and Technology for the Workplace and Higher Education: Proceedings of 'The Learning Ideas Conference'. 2021. Springer International Publishing; 2022. pp. 396-408

[53] Venkata Reddy PS, Nandini Prasad KS, Puttamadappa C. Farmer's friend: Conversational AI BoT for smart agriculture. Journal of Positive School Psychology. 2022;**6**(2):2541-2549

[54] Olujimi PA, Ade-Ibijola A. NLP techniques for automating responses to customer queries: a systematic review. Discover Artificial Intelligence. 2023;**3**(1):20. DOI: 10.1007/ s44163-023-00065-5

[55] Hoy MB. Alexa, Siri, Cortana, and more: An introduction to voice assistants. Medical Reference Services Quarterly. 2018;**37**(1):81-88. DOI: 10.1080/02763869.2018.1404391

[56] Li B, Jiang N, Sham J, Shi H, Fazal H. Real-world conversational AI for hotel bookings. In: 2019 Second International Conference on Artificial Intelligence for Industries (AI4I). IEEE; 2019. pp. 58-62

[57] Liu CC, Liao MG, Chang CH, Lin HM. An analysis of children' interaction with an AI chatbot and its impact on their interest in reading. Computers in Education. 2022;**189**:104576. DOI: 10.1016/j.compedu.2022.104576

[58] Hollander J, Sabatini J, Graesser A. How item and learner characteristics matter in intelligent tutoring systems data. In: International Conference on Artificial Intelligence in Education. Cham: Springer International Publishing; 2022. pp. 520-523

[59] Lin CJ, Mubarok H. Learning analytics for investigating the mind map-guided AI Chatbot approach in an EFL flipped speaking classroom. Educational Technology and Society. 2021;**24**(4):16-35

[60] Cui L, Huang S, Wei F, Tan C, Duan C, Zhou M. Superagent: A customer service chatbot for e-commerce websites. In: Proceedings of ACL 2017, System Demonstrations. 2017. pp. 97-102

[61] Pawlik Ł, Płaza M, Deniziak S, Boksa E. A method for improving bot effectiveness by recognising implicit customer intent in contact centre conversations. Speech Communication. 2022;**143**:33-45. DOI: 10.1016/j.specom.2022.07.003

[62] Dibitonto M, Leszczynska K, Tazzi F, Medaglia CM. Chatbot in a campus environment: Design of lisa, a virtual assistant to help students in their university life. In: Human-Computer Interaction. Interaction Technologies: 20th International Conference, HCI International 2018, Las Vegas, NV, USA, July 15-20, 2018, Proceedings, Part III 20. Springer International Publishing; 2018. pp. 103-116

[63] Georgescu AA. Chatbots for education–Trends, benefits and challenges. In: Conference Proceedings of "eLearning and Software for Education" (eLSE). Vol. 14, No. 02. Carol I National Defence University Publishing House; 2018. pp. 195-200

[64] Swain S, Naik S, Mhalsekar A, Gaonkar H, Kale D, Aswale S. Healthcare chatbot system: A survey. In: 2022 3rd International Conference on Intelligent Engineering and Management (ICIEM). IEEE; 2022. pp. 75-80

[65] Gupta J, Singh V, Kumar I. Florence-A health care chatbot. In: 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS). Vol. 1. IEEE; 2021. pp. 504-508

[66] Sharma B, Puri H, Rawat D. Digital psychiatry – Curbing depression using therapy chatbot and depression analysis. In: 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT). IEEE; 2018. pp. 627-631

[67] Aggarwal H, Kapur S, Bahuguna V, Nagrath P, Jain R. Chatbot to map medical prognosis and symptoms using machine learning. In: Cyber Security and Digital Forensics: Proceedings of ICCSDF 2021. Springer Singapore; 2022. pp. 75-85

[68] Casas J, Mugellini E, Khaled OA. Food diary coaching chatbot. In: Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers. 2018. pp. 1676-1680

[69] Mabunda K, Ade-Ibijola A. PathBot: An intelligent chatbot for guiding visitors and locating venues. In: 2019 6th International Conference on Soft Computing & Machine Intelligence (ISCMI). IEEE. pp. 2019, 160-2168

[70] Nirala KK, Singh NK, Purani VS. A survey on providing customer and public administration based services using AI: Chatbot. Multimedia Tools and Applications. 2022;**81**(16):22215-22246. DOI: 10.1007/s11042-021-11458-y

[71] Gabrielli S, Marie K, Della Corte C. SLOWBot (chatbot) lifestyle assistant. In: Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare. 2018. pp. 367-370

[72] Kurniadi D, Septiana Y, Sutedi A. Alternative text pre-processing using chat GPT Open AI. Jurnal Nasional Pendidikan Teknik Informatika: JANAPATI. 2023;**12**(1)

[73] George AS, George AH. A review of ChatGPT AI's impact on several business sectors. Partners Universal International Innovation Journal. 2023;**1**(1):9-23