Transformer vs. RASA Model: A Thorough Attempt To Develop Conversational Artificial Intelligence To Provide Automated Services To University Disciples

by

Shafkat Zahir
21101109
Prateeti Saha Roy
19141007
Humaira Tasnim Ridita
19141003
Tamanna Hossain
19101347

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
School of Data and Science
Brac University
January 2023

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Shafkat Zahir 21101109	Prateeti Saha Roy 19141007
Humaira Tasnim Ridita 19141003	Tamanna Hossain 19101347

Approval

The thesis/project titled "Transformer vs. RASA Model: A Thorough Attempt To Develop Conversational Artificial Intelligence To Provide Automated Services To University Disciples" submitted by

1. Shafkat Zahir (21101109)
2. Prateeti Saha Roy (19141007)
3. Humaira Tasnim Ridita (19141003)
4. Tamanna Hossain (19101347)

Of Fall, 2022 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on January 19, 2023.

Examining Committee:

Supervisor: (Member)

Dr. Md. Khalilur Rahman
Professor
Department of Computer Science and Engineering
Brac University

Program Coordinator: (Member)

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
Brac University

Head of Department: (Chair)

Dr. Sadia Hamid Kazi
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Ethics Statement

We pledge to uphold the legal requirements of our home nation, workplace, university, and community. We promise to behave honorably, faithfully, and with integrity. We will be willing to own up to our mistakes and only enter into agreements that we intend to keep.

Abstract

With the dynamic advancement of internet services over the past decade, chatbots, also recognized as conversational agents, have risen to prominence. They are significantly acclimated to develop a useful digital expert that can succumb to questions and provide comprehensive answers. The chatbots were designed to enhance community interaction in which they comprehend user inputs, get pertinent information depending on the inputs, and reply using a unified framework. Going to college or university to get necessary academic and supporting information like tuition fees and term schedules can be a hassle for students. It can take a lot of time and effort, especially if they have to visit multiple schools or departments. It can also be frustrating if they have to wait in line or if the information they need is not readily available. Additionally, the process of getting this information requires staff to be available to provide it, which can be costly and time-consuming for the school. So to address this problem, in this study, we ideate and attempt to implement our conceptualization to generate something that is interactive, easily accessible, and able to learn from its interactions with students. There have been many chatbot developments using various artificial intelligence models, but there are still many limitations in their functionality. After conducting extensive research, we have identified two classified models - Transformer and RASA model - to compare and evaluate their accuracy in order to build a more effective conversational artificial intelligence. We hope to gain a better understanding of their strengths and weaknesses by comparing these two models and determining which model is more suitable for chatbot development. This data will help to improve the overall performance and functioning of chatbots.

Keywords: Chatbots, Conversational agents, Artificial intelligence models, Transformer, RASA model, Accuracy, Chatbot development, Accessibility, Learning, Conceptualization

Dedication

We dedicate our thesis to our mentors, family and friends, who have inspired us to conclude our paper successfully amid tough times, and have supported us from beginning to end.

Acknowledgement

We would like to thank our supervisor, Dr.Khalilur Rahman, Phd, for providing us with the essential supervision and for his cooperation with the research.

We would also like to thank Sayantan Roy Arko, Research Assistant, for offering insight into the technical aspects of authoring the article and for his unwavering support in addressing our questions.

Finally, we would like to thank all of the faculty, tutors, and members of BRAC University's Computer Science and Engineering Department for their assistance in furthering our grasp of the principles involved in our study over the past three years.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

AI Artificial Intelligence

$AIML$ Artificial Intelligence Modelling Language

ANN Artificial Neural Network

$DLTL$ Digital Learning Technologies

$GloVe$ Global Vector

IoT Internet of Things

ML Machine Learning

NLP Natural Language Processing

NLU Natural Language Understanding

POS Parts of Speech

RNN Recurrent Neural Network

$Seq2seq$ Sequence To Sequence

$TF-IDF$ Frequency Inverse Document Frequency

Chapter 1

Introduction

1.1 Background and Motivation

Chatbots, also known as conversational agents, have been a popular topic of discussion for quite some time now, as their innovation has made problem-solving more convenient and efficient. The concept of chatbot is to provide instant and accurate responses to users' queries, which helps to save time and effort. With the advancement of technology, chatbots have become more sophisticated and human-like in their interactions. There are various methods for creating chatbots such as rule-based, template-based, and AI-based. But, among all the approaches, artificial intelligence (AI) has proven to be the most effective method for producing the best outcomes. AI-based chatbots can understand natural language, and adapt to the user's behaviour and preferences, which allows them to provide personalised and accurate responses. They can also learn from their interactions with users and improve their performance over time.
Increasing productivity is the primary motivation for chatbot users, followed by entertainment, social connections, and interest in new experiences. To effectively appeal to these incentives, a chatbot should be created to serve as both a helpful tool, a playful toy, and a friendly companion [3].According to research, user inquiries for technical assistance with chatbots often involve both emotional and informative elements, with the majority (over 40 %) being emotional in nature and not seeking specific information [6]. It is through the use of machine learning that customer service chatbots are able to detect and respond to emotions in a manner similar to human operators [7].

Natural Language Processing (NLP) is a highly advanced method for understanding and representing human languages [2]. It is a significant area of artificial intelligence and is used in a number of applications, including speech recognition, text mining, and machine translation. The fundamental components of NLP include phonetics, morphology, syntax, semantics, and pragmatics. Machines need to be able to break down text into paragraphs, phrases, and words, as well as learn to recognize word connections, deduce meaning from text, grasp phrases in various situations, and analyse the speech environment [5]. A feed-forward neural network language model was put forth in the early twenty-first century, and the word2vec implementation's use of word embeddings makes it possible to capture particular associations between words. Recurrent Neural Networks (RNN) and Long Short-Term Memory Networks

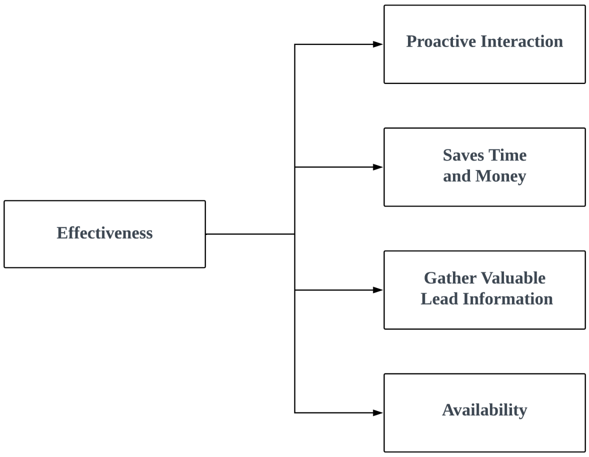

Figure 1.1: The Incentives of Using Chatbots

(LSTMs) have largely replaced feed-forward neural networks for language modelling. However, after thorough research, it has been determined that the Transformer and RASA models would be the best fit to deliver the optimal results.

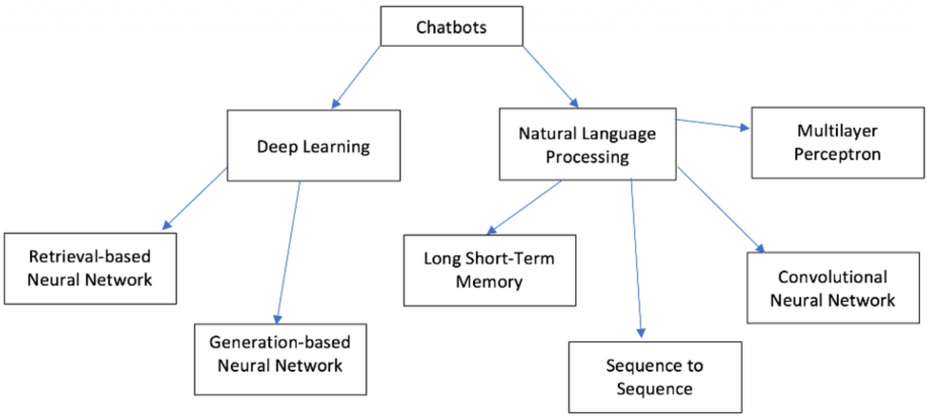

Figure 1.2: The Incentives of Using Chatbots

1.2 Problem Statement

Prior endeavours, such as [16], [17], employed neural network-based models to discern emotions in speech, while [10]–[12] made attempts to achieve similar outcomes through the utilisation of lexical characteristics. These techniques are unimodal in nature, as they focus primarily on a single mode of expression. However, this overlooks the significance of these attributes when employed in conjunction. To elaborate, the thoughts conveyed in text can be perceived in a plethora of ways based on the tone and voice characteristics. For example, depending on the tone of voice, the words "How could you do that?" might convey a variety of emotions. It may convey wrath if articulated loudly and aggressively, or grief if articulated slowly and quietly. Similarly, analogous speech qualities may be found in distinct emotional expressions and can only be differentiated through the examination of the spoken content, i.e. the lexical data. To generate outputs, chatbots employ the encoder decoder paradigm. There are abundance of methods that may be utilised to guide the process, such as rule-based or flow-based chatbots. While rule or flow-based chatbots are considered to be the most basic principle for constructing chatbots, they are not particularly effective in terms of providing unique and precise information. After years of extensive research and advancements in technology, scientists and engineers have devised the concept of utilising NLP to build chatbots. Chatbot NLP systems utilise sophisticated Machine Learning (ML) techniques to comprehend the user's intention and match it to a set of actions that the chatbot is capable of performing. The NLP engine, which is tailored to the specific business application, employs either finite state automata models or deep learning algorithms to interpret the user's inputs.

While many universities have begun to explore the potential of chatbots, few have successfully implemented the technology in a way that fully leverages its capabilities for personalization, cost-effectiveness, and improved student satisfaction. The development of chatbots utilising NLP models has been met with a surfeit of challenges, as finding the appropriate model remains a formidable task. The reality is that challenges are an inherent aspect of any sort of innovation, and chatbots are no exception to this rule. Throughout our research, we have encountered a surplus of difficulties such as the proper integration of NLP models, the ability to personalise responses, and the ability to provide accurate and comprehensive information. These challenges are not unique to chatbots and are commonly faced by any sort of innovative technology. However, with continued research and development, it is possible to overcome these obstacles and achieve the desired results. Throughout our research, we have encountered many challenges such as -

- Accuracy : Short and inadequate responses, errors in grammar and semantics that are not typically made by humans, and inconsistencies in the conversation can all lead to problems with accuracy. Text vectorization is used to enhance the semantics and similarity of the text, resulting in representations known as "static word embeddings." One drawback of static models is that they are prone to polysemy, meaning that the same words may be embedded in different context [14].

- Human Interactions : In spoken language, foreign words and phrases are frequently utilised. As a result, multilingualism may be one of the challenges in designing chatbots. Deciphering messages, interpreting the user's intent, and generating consistent replies are all tough challenges. When bots are abused, whether purposely or accidentally, issues may arise, with psychological, legal, economic, social, and democratic repercussions. In research, bot bans and unambiguous bot usage disclosure to end users were evaluated as measures to avoiding such abuses. Some of the technological and ethical challenges are being addressed through crowdsourcing, natural language processing, and machine learning [14].

- Data Collection : Large amounts of data must be collected, with audience behaviour and preferences taken into account [14].It's also worth thinking about ethical issues and removing offensive or obscene language from the data [20].

- Natural Language Processing's Limit : NLP has certain restrictions as well. It might get puzzling. ML is another alternative, but it requires a well-defined set of rules to be effective. It will be a tragedy otherwise. It does, however, make modification easier and significantly enhances the ability to identify relevant solutions to user inquiries [20].

1.3 Research Objective

AI has been changing various industries for a while now, and the education sector is no exception. The use of AI-based learning and communication tools is becoming more and more common. In the United States, most university students prefer using digital learning technologies (DLT) as an academic tool. A survey found that 79% of students believe that DLT helps teachers to be more efficient during classes and 81% think that it helps to improve grades. In order to enhance their educational systems, schools and universities are willing to take benefit from AI. Universities are using educational chatbots powered by AI more frequently to streamline interactions with prospective, current, and returning students. This study aims to develop an artificial intelligence-based bot by comparing two classified models, Transformer and RASA model, with the objective of identifying the superior model that can accept user input through speech or text, process and understand the messages, and provide text-based answers or responses. The use of this bot can significantly simplify the process of finding answers to frequently asked questions related to the university's curriculum, policies, regulations, and other information that students may need.

It is important to note that the Transformer model is a type of neural network architecture that is widely used in natural language processing tasks such as language translation, text summarization, and question answering. It is known for its ability to handle long-term dependencies and has shown to be effective in various NLP tasks. RASA, however, is an open-source framework for creating conversational AI. To design, train, and deploy chatbots, it offers a set of libraries and tools. As the utilisation of bots has gained traction in recent years, attention has been directed towards the development of a bot that can cater to the needs of an entire organization. However, an effective solution has yet to be identified that can provide comprehensive and accurate information to students, and also be cost-effective for universities.

This study aims to fill that gap by comparing these two models and finding the one that can be used to build a chatbot that can meet the needs of universities. It is expected that the results of this study will help universities understand the strengths and weaknesses of these models and determine which one is more suitable for chatbot development. This data will help to improve the overall performance and functioning of chatbots and make it more efficient for students to access information they need. The following are the study's objectives:

- Empowering students with convenient access to academic program specifics, campus layouts, financial aid details, and the wide array of information found on a website
- To offer answers to the users in a timely or effective manner
- To compare and analyse the models and how they work
- Help universities make a good first impression on website visitors while also making the most of the staff's time.
- Creating an easily accessible tool to gather contact information, such as names and email addresses, so that visitors can receive the information they desire immediately without the need to wait for a response over an extended period of time.

1.4 Research Overview

In our research, we have determined that comparing the NLP-based models - Transformer and RASA model, with distinct frameworks - and identifying the most suitable match in constructing an interactive conversational AI to assist universities and colleges is the optimal solution. After analysing the problem, we have proposed the design of this supportive mechanism. Our research commences with the collection of our dataset, which was procured through various websites, surveys among students, and resources from the academic office and faculties at BRAC University. We fed an adequate dataset to our models after gathering and developing it. Additionally, we employed appropriate cleaning procedures to eliminate any abnormalities from our dataset. Following the cleaning process, we partitioned our dataset for training and testing. Finally, we produced our results, determining accuracy and concluding our investigation for future improvements. The flowchart below provides a high-level summary of our research process.

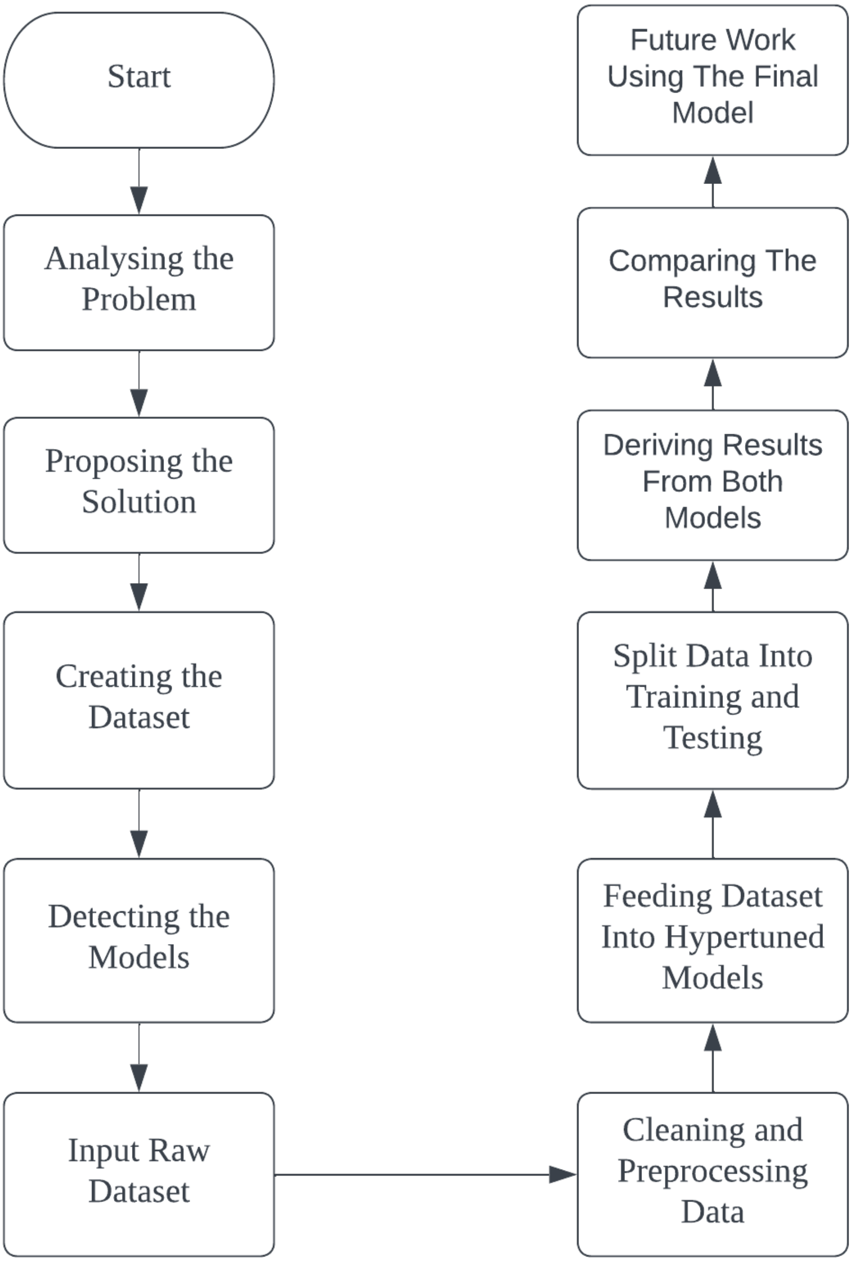

Figure 1.3: Workflow of the Process

Chapter 2

Related Work

Humans employ human language to converse with one another, whereas chatbots utilise natural language to interact with their users. Chatbots are computer programs that provide communication services through natural human language. With the swift growth of speech recognition technology and the advancement of computers, chatbots have become a prevalent means of engaging with users. Voice controlled interfaces have also surfaced and their popularity is rapidly increasing, with people incorporating them into various aspects of their lives. Chatbots are useful not only for human engagement but also in fields such as education, business, health, commerce and entertainment. Furthermore, chatbots have been integrated into various social media platforms and others. Alan Turing posed the question "Can machines think?" in 1950, wondering if a computer program could converse with humans without them realising they were talking to a machine. The Turing Test, also known as the "Imitation Game", was devised by Alan Turing, which eventually led to the creation of chatbots. The first chatbot, ELIZA, was developed by Joseph Weizenbaum in the year 1966 [13]. However, due to its limited capabilities, it could only discuss certain topics and couldn't maintain long conversations. In 1972, a more advanced version of ELIZA called Parray was introduced, but it had a slow response time and couldn't learn from the conversation. Later, in 1995, the first online chatbot, ALICE, was created, taking inspiration from ELIZA. The Artificial Linguistic Internet Computer Entities (ALICE) was considered remarkable due to its use of Artificial Linguistic Internet Computer Entities (AIML), which was also one of the main differences between ALICE and ELIZA.

Chatbots have greatly evolved in their capabilities with the advancements in contemporary AI systems. Intelligent devices that can handle voice commands have been integrated into home automation, smartphones and other devices. Apple developed Siri in 2020 which responds to user commands, requests for information and makes suggestions using various online services. Siri requires internet access to function and may not support all languages. IBM released Watson in 2011, a chatbot that can understand natural human language but only supports the English language. Google Now was released in 2012 and Google Assistant in 2016, which forms the next generation of Google Now. Microsoft designed Cortana, a personal assistant and Amazon designed Alexa in 2014. The incorporation of Alexa into household devices and appliances, such as those for home automation and entertainment, has increased the ease of access to the Internet of Things for individuals [4]. The pattern based method was employed by ELIZA and ALICE. The pattern-based approach stores a set of finite responses, the chatbot only compares the user input with a predetermined rule pattern and selects an answer accordingly, mostly using pattern matching algorithms. Since there is already a collection of responses, the reaction time is reduced, but the responses are repetitive and lack creativity. Languages such as Artificial Intelligence Markup Language (AIML), Rivercript and Chatscript are mostly used in this approach. In contrast, the machine learning approach implements Artificial Neural Networks (ANNs). The advent of machine learning has made the Internet of Things (IoT) more user-friendly by eliminating the need for predefined sets. In the creation of chatbots, there are two methods that can be employed: pattern recognition and machine learning.

NLP and machine learning (ML) are two AI-related terms. A chatbot, powered by NLP and ML, communicates with users using natural language. NLP is used by the chatbot to match the user's purpose with the previously classified intent. NLP does not employ keywords; instead, it uses its understanding of pattern recognition, phrase structure, and idioms. There are three main modules included in the modern dialogue system architecture: NLP, dialogue manager and NLG (natural language generation). NLP is a branch of computer science that focuses on computational techniques. Computational techniques are used to learn, understand, and create human language material. ML methods are frequently employed to improve NLP. ML approaches allow machines to understand part of the knowledge of the actual world. ML may be defined as a computer software that automatically learns and improves its performance via experience and is able to extract patterns from data instead of depending on rules. A chatbot was engineered by Rahman (2012) to assist university students with information. Prior studies have demonstrated that chatbots can act as a domain-specific information system and have explored methods to enhance their precision in a specific area [1]. Research by Gomathy (2021) focuses on the creation of a voice-controlled chatbot system for colleges to increase efficiency. The study illustrates the use of speech recognition technology to operate the chatbot remotely and make choices on behalf of the user. The system employs artificial intelligence and is designed to be able to monitor and manage the environment in real-time. The AI takes speech as input and delivers results to the entire institution in a shorter time frame [19]. Ranoliya et al. [4]developed a chatbot to provide an efficient and accurate answer for university frequently asked questions using artificial Intelligence Markup Language and Latent Semantic Analysis. Then, the chatbot is integrated into a particular website. Recent work done in [13] investigated educational chatbots on Facebook Messenger to recommend personalised learning content.

In order to gain a better understanding of previous studies on chatbots, we examined research articles which employed the use of a client-server architecture and a Sequence To Sequence (seq2seq) model of Recurrent Neural Network (RNN), along with an Android GUI, to create a chatbot named HappySoul. The chatbot was designed to alleviate the psychological stress of adolescents and provide a safe space for them to discuss their mental health issues. We also reviewed a paper [8] that developed a chatbot for a college that students and staff could use to easily access resources and information about the institution. The chatbot used AIML, lemmatization and parts of speech (POS) tagging with Wordnet, semantic sentence similarity, and log file, by using libraries from Tensorflow. The paper discussed using AIML files to check if the user's input pattern was available before returning a response. If the input pattern was not available, they proposed using keywords from the input and finding the appropriate lemmas for the keywords through lemmatization and POS tagging to group the different inflected forms of the words together. After that, the similarity score was calculated by averaging the similarity of each key word in two phrases. The implementation also included a log file to record inputs that the chatbot was unable to respond to. Rasa NLU is an open-source library for natural language understanding that is employed for identifying user intent and extracting specific information, such as entities, from text inputs.It has been designed to be very flexible and modular, allowing developers to customise the models according to their specific use case. According to a review of Rasa NLU, it was found to be highly accurate in intent classification and entity extraction when compared to other open-source libraries. The library is also praised for its user-friendly design, with a simple and easy-to-use API that has clear and well-documented instructions. However, some drawbacks have also been identified, such as limited support for certain languages and the requirement of more training data for optimal results. Overall, RASA NLU is a valuable tool for natural language processing tasks and is particularly useful for developers who are looking for an open-source solution. To gain a deeper understanding of RASA NLU, we examined research article [21] which provides a comprehensive examination of the models through the development of a chatbot named "Farmers' Assistant" that offers professional advice on managing crops and vegetation. The authors chose RASA because it is an open-source library that is compatible with NLU and allows developers to set up, install, and use NLU on local servers.

This improves processing performance by reducing network time when compared to cloud-based platforms. When creating this chatbot, which ultimately showed promising results, the authors of the study reported using RASA Core v2.0, the latest version of RASA. Additionally, the paper [15] developed a chatbot that responds to inquiries about finance by utilising RASA NLU and neural network methods for entity extraction. The study found that RASA NLU is preferable as it is easy to classify intent and predict entities with an intent accuracy of 0.99375. Additionally, the research reveals that when using Tensorflow technology, a neural network model with an additional representation layer similar to word2vec can be used as a NER model. However, the paper concludes that in a single experiment, the RASA NLU technique outperforms the neural network method in terms of accuracy. Furthermore, another research presents a novel approach to improve the user experience of conversational chatbots by designing a Question Answering Model (QAM) that can evaluate large comprehensive datasets to provide accurate and context-aware answers. The proposed QAM uses the state-of-the-art Bidirectional Encoder Representations from Transformers (BERT) model and Google Dialog Flow to simulate human-like conversation and improve reading comprehension tasks. The integration of BERT model and Dialog Flow is achieved through webhook and API. Additionally, the Chatbot Interaction with Artificial Intelligence (CIAI) framework is an innovative way to train a chatbot-like architecture that is focused on natural human interaction, rather than relying on interfaces, code, or specific commands. The system prompts human users to rephrase commands and questions for task identification, which are then divided into training and validation sets. The authors of the paper found that using an ensemble of the five best-performing transformer models through Logistic Regression of output label predictions led to an accuracy of 99.59% on the dataset of human responses. This highly-performing model allows the system to interpret human commands at the social-interaction level through a chatbot-like interface, making AI more accessible for non-technical users.

The paper [18] suggests using a chatbot, also known as a retrieval-based user interface, to extract specific information from construction specifications as per the user's request. The paper assesses the viability of using a question answering approach that is based on Bidirectional Encoder Representations from Transformers (BERT) to develop the information retrieval chatbot. The proposed system uses artificial paraphrasing to augment human-sourced data, creating a large set of training data for various natural language processing (NLP) techniques, including classical, attention, and language transformation-based learning approaches. Human users are asked to paraphrase commands and questions for task identification, which are then split into training and validation sets. The training set is then augmented by the T5 model to generate additional data. The paper benchmarks seven state-of-the-art transformer-based text classification algorithms (BERT, DistilBERT, RoBERTa, DistilRoBERTa, XLM, XLM-RoBERTa, and XLNet) on the augmented and non-augmented sets after fine-tuning for two epochs. The results show that all models improved when training data was augmented by the T5 model, with an average increase in classification accuracy of 4.01%. The best result was obtained by the RoBERTa model trained on T5 augmented data, achieving 98.96% classification accuracy. Finally, the paper found that an ensemble of the five best-performing transformer models through Logistic Regression of output label predictions led to an accuracy of 99.59% on the dataset of human responses. This highly-performing model allows the system to interpret human commands at the social-interaction level through a chatbot-like interface, and provides better accessibility to AI for non-technical users.

Ensuring proper construction quality and avoiding contractual problems requires monitoring construction specifications during every phase of construction. However, manual review is inefficient, costly and prone to errors. While there have been attempts to automate the review process, they have limited practical application. This research [22] proposes using a retrieval-based user interface, also known as a chatbot, to extract specific information from construction specifications as requested by the user. To develop this information retrieval chatbot, the paper evaluates the feasibility of using a question answering methodology based on BERT. This approach leverages pre-trained BERT models to successfully extract user-wanted information from construction specifications. This method is flexible and can respond to a variety of questions, eliminating the need for time-consuming manual tasks such as labelling. This research [26] focuses on the ChatGPT model of OpenAI, which is an artificial intelligence that is trained through supervised and reinforced machine learning to answer questions from users. The answers provided by the algorithm are based on the input it receives from users, as well as from the content it has been trained on. The paper examines how the answers given by the model for definitions of crowdfunding, alternative finance, and community finance compare to those given by real human experts in academic scholarship. These terms were chosen because academic literature does not provide consistent definitions for each of them, but some definitions are accepted by more scholars. The study aims to fill the research gap concerning the accuracy of answers generated by an artificial intelligence, and thereby contribute to the growing literature on the implications of textual artificial intelligence on academia.

This study investigates the potential of artificial intelligence, particularly natural language processing, in improving academic performance, using economics and finance as a specific example. By using a case study approach and focusing on ChatGPT as an NLP tool, the study aims to examine the ways in which it can enhance academic research. The analysis of ChatGPT's applications, capabilities, and limitations revealed that it has the potential to significantly improve research in general and in the field of economics and finance in particular. It can aid researchers in tasks such as data analysis and interpretation, scenario generation, and communication of findings. However, there are some limitations to consider when using chatbots or similar tools in research, such as generalizability, reliance on data quality and diversity, lack of domain expertise, limited ability to understand context, ethical considerations and limited ability to generate original insights. Therefore, it is important to be aware of these limitations and to use ChatGPT in conjunction with human analysis and interpretation when conducting research [23]. This study [25] examines the potential of using chatbots to improve public services through human-AI interactions in a Norwegian public service organization. Researchers and practitioners work together to understand how chatbots can assist human service agents, and the potential for creating hybrid human-AI teams. The study emphasises the importance of allowing service agents to discover the potential uses of chatbots based on their actual practices, and provides insight on how chatbots can work in partnership with human service agents to better serve citizens' needs. The authors offer key takeaways on human-AI partnerships, theory-based interventions, and institutionalised collaborative research for researchers and organisations looking to engage with technology and innovation, as well as engage with research.

Chapter 3

Dataset and Preprocessing

3.1 Data Collection

For the fabrication of a proficient dialogic model, a copious training corpus is imperative, one that encompasses a plethora of verbiage. In this study, we procured a grand total of 1000 inquiry and response pairs through a canvassing of individuals from diverse universities. Additionally, we also procured queries through sundry social media congregations established and managed by students of Brac University, as well as scouring other web portals to gather more general inquiries made by university students. The categories of queries we received were mainly germane to entrance prerequisites, mandatory documentation, requisite insurance, enrollment quotas for each subject in a particular year, admission standards for each major, scholarship programs, and course curriculum. To furnish accurate answers, we also procured information from various sources such as the Brac University website, mentors and advisors, the registrar's office, and our own expertise.

Question	Answer
I want to get admission into Brac. What are the BRACU admission requirements?	Minimum GPA of 3.5 in SSC/Equivalent and HSC/Equivalent separately (including additional subject). Minimum CGPA of 2.5 in O-Levels in five subjects and A-Levels in two subjects separately, according to the scale (A=5, B=4, C=3 & D=2). Subjects with E grade will not be considered.
What is the process of getting admitted to Brac?	To get admitted into BracU, applicants must qualify in the admission test (written and an interview) after submitting their online application form. As the admission test is a highly competitive one, applicants are required to perform satisfactorily in the examination in order to secure their seats.
When is the application deadline for admission ?	If you are willing to seek for admission deadlines you can check out the 'Announcement' segment on the Bracu website. The link has been provided - https://www.bracu.ac.bd/

Question	Answer
What are the core courses in CS?	Here are the courses required to complete the dregree in Computer Science - https://www.bracu.ac.bd/academics/departments/computer-science-and-engineering/bachelor-science-computer-science/cs
What are the all course codes for CSE?	To know more about it, please follow the link provided below - https://www.bracu.ac.bd/academics/departments/computer-science-and-engineering/bachelor-science-computer-science-and/cse-0
I am new to CSE. What courses should I start with?	These are the courses mostly done during first semester - CSE110, ENG101,MAT110, & PHY111. Incase you are still worried on how to progress further, you can take advise from the alumnis, advisor or faculties.

Figure 3.1: Fragments of the Dataset

3.2 Preprocessing

3.2.1 RASA

Our RASA paradigm is a composite and sophisticated edifice, constructed upon both NLU and Core pipelines. The elements, as per Rasa's classification, constitute the NLU pipeline and operate in a consecutive manner to convert user input into a structured output. The NLU pipeline comprises of various features such as Entity extraction, which extracts relevant entities from the user input, intent categorization, which assigns predefined intent categories to the input, answer selection, which selects the most appropriate answer from the predefined list of answers, preprocessing, which includes various preprocessing steps such as lowercasing, stemming, etc. and other features that are essential for the proper functioning of the model.

Components in Rasa open source framework are classified as following:

- Tokenizers
- Featurizers
- Intent Classifiers
- Entity Extractors
- Selectors

Tokenizers

Tokenization is the process of segmenting the original text into smaller bits. known as tokens, and returning a list of words or tokens. this is the initial step in any natural language understanding pipeline and must be done before the text is transformed for machine learning.

Featurizers

Featurization is the process of converting raw input data into a processed form known as a Feature Vector, which is a machine-readable representation used as input for the machine learning model. In RASA, text features are divided into two categories:

- Sparse featurizers
- Dense featurizers

Sparse featurizers return feature vectors with a high number of missing values, such as zeros, and are stored as sparse features as they would otherwise take up a lot of memory. They only store non-zero values and their positions in the vector. Dense feature vectorizers, on the other hand, return feature vectors with pre-trained embeddings. The length of these vectors ranges from 50 to 300, and they perform better than sparse vectors in all NLP problems. As a result, words with similar meanings have similar symbols. In sparse vector representations, terms like "home" and "house" denote distinct things, whereas in dense feature representations, the similarity between these phrases is captured.

-name	WhiteSpaceTokenizer
-name	RegExFeaturizer
-name	LexicalSyntacticFeaturizer
-name	CountVectorsFeaturizer

Table 3.1: Pipeline Elements

It's possible to extract various types of characteristics from text and combine them together to serve as input for a machine learning model. In the pipeline above, three separate featurizers are used in sequence, with one featurizer relying on the output of the previous featurizer. The RegexFeaturizer uses regular expressions to generate a sparse feature vector representation of raw text input, the LexicalSyntacticFeaturizer generates lexical and syntactic characteristics for raw text input to aid in entity extraction, and the CountVectorsFeaturizer generates a bag-of-words representation of raw text input, intentions, and answers.

Intent Classifiers

Intent classification is a crucial part of RASA, an open-source framework for building conversational AI. RASA utilises machine learning to train an intent classifier that can recognize and classify a user's intent based on their text input. To perform intent classification in RASA, a training dataset that includes examples of user inputs (called "utterances") and the corresponding intent labels is needed. Rasa uses this dataset to train a machine learning model that can predict the intent of new user inputs. RASA offers several pre-built models for intent classification like 'DIETClassifier', 'EmbeddingIntentClassifier', 'KeywordIntentClassifier', and 'SklearnIntentClassifier'. Custom models can also be used by integrating with other libraries like TensorFlow and scikit-learn.

Once the model is trained, it can be used to classify new user inputs by matching them against the training data. The classifier will return the intent label that best matches the input, along with a confidence score indicating how confident the model is in its prediction. After generating features for all the tokens and the entire sentence, it can be passed to an intent classification model. Intent classifiers assign one of the intents defined in the domain file to incoming user messages. The use of RASA's DIET model is recommended as it can handle both intent classification and entity extraction. It can also learn from both token- and sentence-level features.

In the pipeline, we use the DIET classifier, which extracts entities, intents, and outputs entities, intents, and intent rankings.

Entity Extractors

Entity extractors are used to extract entities from text data, such as human names or places. While the DIET model is capable of identifying entities, it is not necessary to use it for all forms of entities. Entities that follow a regular pattern, such as phone numbers, do not require an algorithm to recognize them. In the pipeline, the "EntitySynonymMapper" is used for entity extraction. It maps synonymous entity values to the same value. If the training data has defined synonyms for an entity, this component will ensure that detected entity values are mapped to the same value. For example, if the training data contains examples of similar entities such as "Bangladesh" and "BD", this component will allow you to map the entities "Bangladesh" and "BD" to bd. The entity extraction will return 'uk' even if the message contains "Bangladesh". When this component changes an existing entity, it adds itself to the processor list of that entity.

Selectors

After extracting intents and entities, a response may need to be constructed from the supplied input text message, this is where selectors come in. Selectors anticipate a response from a collection of predetermined replies based on the confidence of the intentions. In the pipeline, the "Response Selector" component is used to return a dictionary with the key as the answer selector's retrieval intent, and the value including anticipated replies, confidence, and the response key under the retrieval intent.

Figure 3.2: RASA Pre-processing Pipeline

3.2.2 Transformer

Transformer architectures, like other neural networks, are unable to handle unprocessed text inputs. Therefore, the first step in the pipeline is to convert the text inputs into numerical encodings that the model can understand and process. To accomplish this, we use a tokenizer which is responsible for -

- Tokens are created by dividing the input into words, subwords, or symbols
- Each token is assigned an integer.
- Including new inputs that may be relevant to the model

To effectuate this preprocessing appropriately, we must initially procure the indispensable data from the Model Hub. This is because the preprocessing must be executed in the precise identical manner as when the paradigm was pre-trained. To achieve this, we availed ourselves of the AutoTokenizer class and its pretrained() method. By proffering the model's checkpoint name, it will automatically retrieve and store the data germane to the tokenizer.

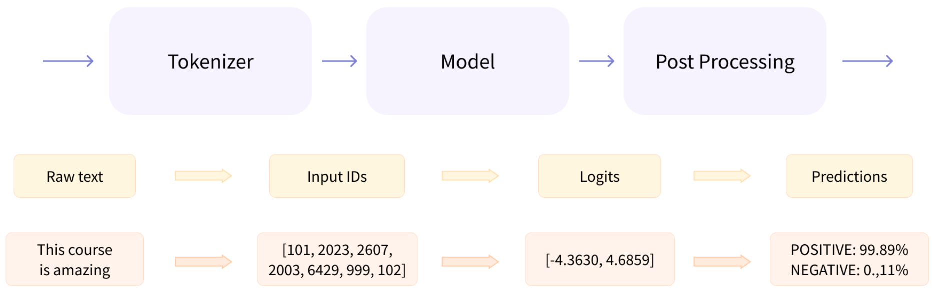

Figure 3.3: Transformer Preprocessing Pipeline

Chapter 4

Proposed Methodology

4.1 Background

We employed NLP frameworks to erect our chatbot as NLP is an essential element for chatbots as it empowers the chatbot to comprehend and respond to human input in a naturalistic manner. A chatbot can decipher the user's objective and reply accordingly by utilising NLP techniques such as sentiment analysis, named entity recognition, and language comprehension. Given the fact that both the RASA model and Transformers are a sub-discipline of NLP and are relatively effortless to implement, we are determined to employ both for our research to determine which framework-based chatbot is the most suitable.

4.1.1 RASA

We resolved to toil with the RASA framework rather than fabricate a chatbot from the ground up because RASA is an open-source edifice for creating conversational AI that offers a plethora of resources and libraries for creating chatbots and other conversational systems. RASA is crafted to be highly modular and adaptable, rendering it effortless to integrate into extant systems and workflows. Chatbots constructed with RASA can be deployed on various portals such as Facebook Messenger, Microsoft Bot, and Slack. RASA comprises of two chief components:

- Rasa NLU (Natural Language Understanding) : Rasa NLU is an open-source tool for natural language processing that is used for determining the user's intent, extracting structured data from user's input and allowing the chatbot to comprehend the user's statement.

- Rasa Core: A chatbot framework that utilises machine learning-based dialogue management which takes the structured input from NLU and predicts the most suitable action using a probabilistic model such as an LSTM neural network rather than using if/else statements.Additionally, it employs reinforcement learning to enhance the prediction of the next action.

In other words, RASA NLU's role is to understand and organise the user's input as structured information, while RASA Core is responsible for determining the next steps or actions that the chatbot should take. Both RASA Core and RASA NLU are separate components that can be used separately or together. RASA allows for training the NLU and dialogue management models through interactive learning by giving the model examples of the types of conversations that the bot should be able to handle. Additionally, RASA offers a customizable web-based interface called RASA X, which makes it easy to manage and test a trained model, as well as make improvements to it.

NLU in RASA stands for Natural Language Understanding. NLP is a field of artificial intelligence that deals with analysing, understanding, and generating human language. It involves techniques and algorithms for processing and analysing text and speech data. NLU is a subfield of NLP that focuses on extracting meaning from natural language text or speech. It involves techniques for understanding the intent of the text, identifying entities and concepts, and determining the relationships between them. Overall, RASA NLU is a powerful and flexible framework for building conversational AI that can be easily integrated into existing systems and workflows, it allows developers to build chatbots and other conversational systems that can understand and respond to natural language inputs. Hence, in our research, We're going to create a model for intent classification that can recognize user-provided intents and categorise phrases accordingly. The Rasa pipeline was primarily designed for chatbot understanding, but in this case we are modifying it to achieve the desired outcome.

Architecture

The RASA Framework is built on the foundation of natural language understanding (NLU) and dialogue management (DM) which is managed by the RASA Core. The RASA NLU component handles user input, and the RASA Core (Dialogue Management) makes decisions based on that input.

- **RASA NLU Architecture**

  However, the architecture of RASA NLU is composed of several components that work together to extract meaning from user inputs, allowing for easy customization and integration with other NLP and ML libraries such as scikit-learn, TensorFlow, and spaCy to build more powerful and accurate models [9], [24]. RASA NLU uses the component named Pipeline to extract intent and entities through tokenization of the text such as splitting each word as a token which is turned into a numeric feature through vectorization along with fine-tuning the model to improve its performance. One of the several main components include Intent Classification where RASA NLU uses the component named Pipeline to extract intent and entities through tokenization of the text, such as splitting each word as a token which is turned into a numeric feature through vectorization, along with fine-tuning the model to improve its performance. One of the main components includes Intent Classification where RASA NLU uses a supervised machine learning approach to classify intent along with a variety of classifiers such as logistic regression, decision tree, and neural networks to identify the user's intent based on user's input, through tasks like lowercasing, stemming, and removing stop words. Pre-processing of the input is completed to improve the accuracy of intent and entity recognition. Additionally, the Feature Extraction component of the RASA NLU model extracts features from the pre-processed data through word embeddings, such as TFIDF or GloVe, or other features like POS tags and dependency parse. Afterwards, lastly, the Language Support component of the RASA NLU model supports several languages, so that models for different languages could be built.

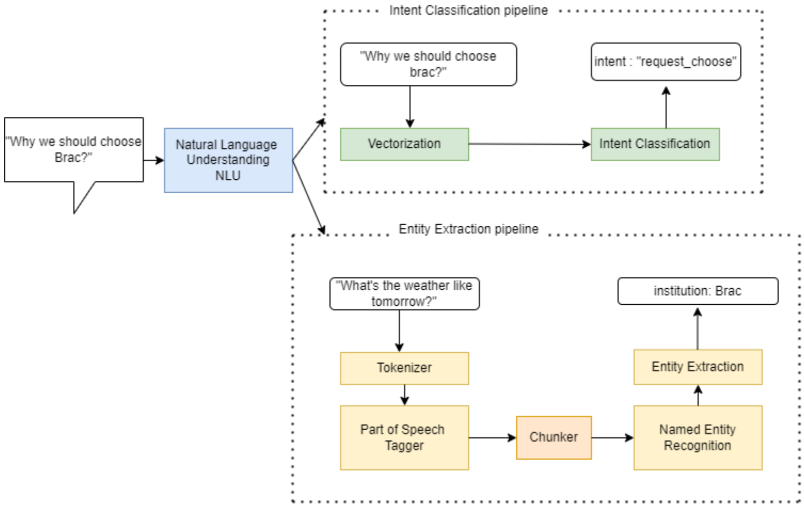

Figure 4.1: RASA NLU Architecture

- **RASA Core Architecture**

  The architecture of RASA Core is designed to manage the flow of a conversation and determine the next action to take based on the user's input and the current context[5] and it is done by one of the main components of RASA Core called the Dialogue Management that uses a policy based approach like rule-based, ML-based and hybrid policies to complete the task. RASA Core can decide what to do next by using this component since it stores all the conversational data, including the user's prior inputs and additional conversational context, in a database. This component is called Tracker Store. After RASA Core decides what action has to be taken, it sends an API request to an Action Server, which executes the action. Then, based on the results of NLU and dialogue management, the Action Selection component determines which action should be carried out next. A typical RASA Core architecture can be represented in a diagram as followss:

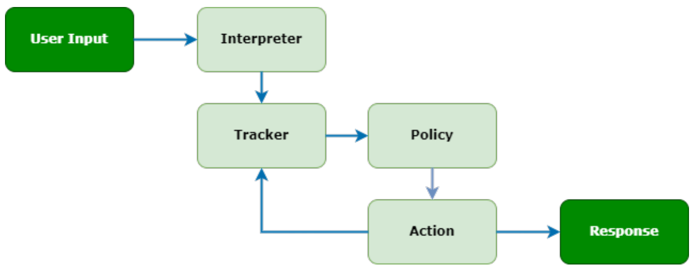

Figure 4.2: RASA Core Architecture

Work Plan

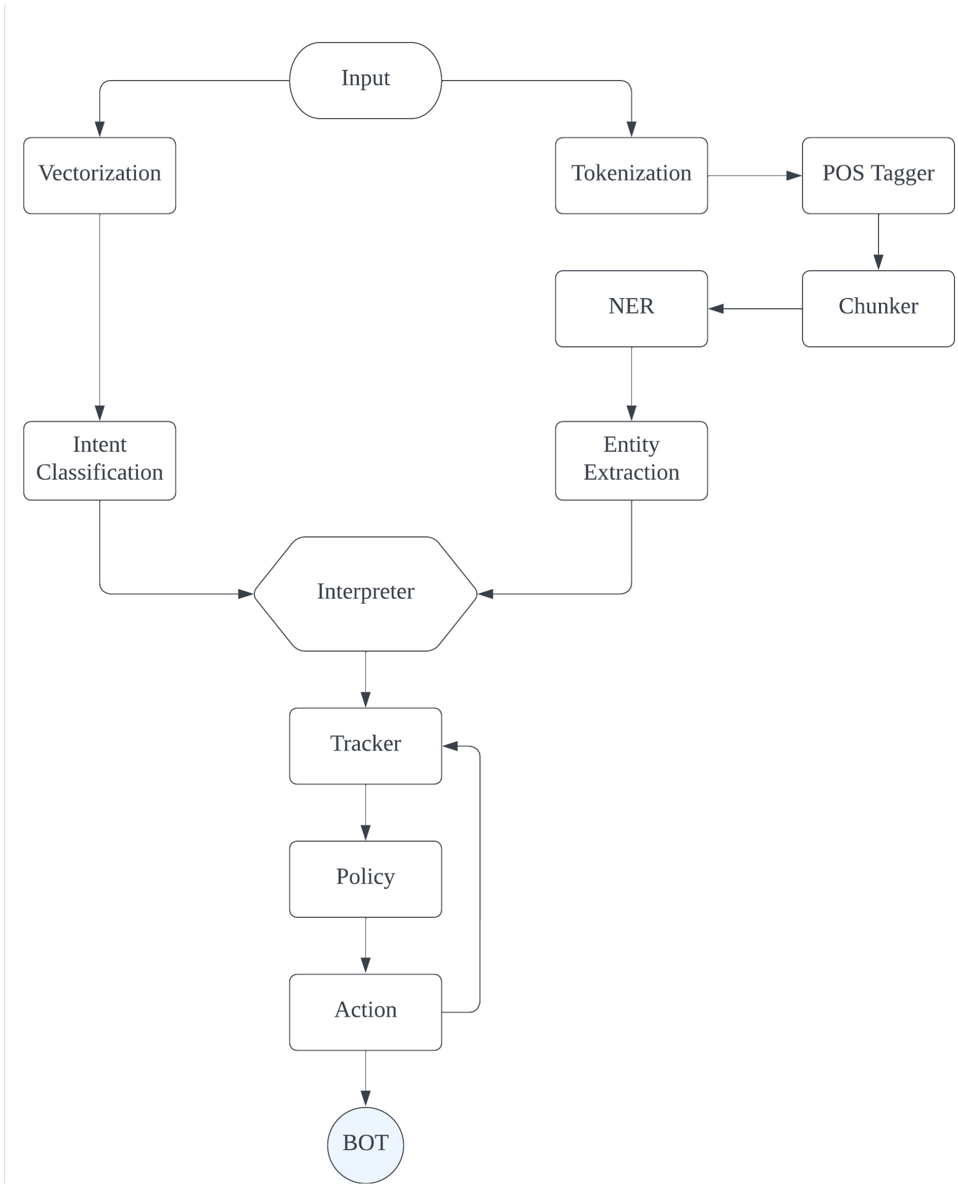

Figure 4.3: RASA Work Plan

4.1.2 Transformer

Background

If a company is able to make it simpler for individuals and organisations to access and utilise AI technology, especially in the field of conversational AI, there will certainly be more interest in it. This ease of access should be approached in two ways. Firstly, by providing cost-effective or free access to the software and tools needed to develop and implement AI systems. Secondly, by making the development process straightforward and accessible to people with a range of technical skills, from beginners to experts. Lowering the barrier to entry in this way will make it more accessible for businesses and individuals to adopt and integrate AI technology into their operations, leading to a wider adoption and utilisation of the technology. It's crucial to have the ability to use and test software through Jupyter Notebooks without any cost and without needing advanced technical skills, in order to increase the number of people who adopt it.

Hugging Face has been successful in providing an open-source library that is easy to access and use for a wide range of natural language processing tasks. This is why we have used its libraries and pipelines to build our transformer-based chatbot. The Hugging Face community and platform for AI and machine learning was founded in 2016 by Thomas Wolf, Julien Chaumond, and Cl´ement Delangue. The platform aims to make natural language processing (NLP) more accessible by giving data scientists, AI professionals, and engineers direct access to more than 20,000 pre-trained models built on the transformer architecture. These models can be used for various tasks like extracting information, text classification, question answering, text generation, translation in over one hundred languages, speech recognition, audio classification, object detection, image classification, tabular data analysis and reinforcement learning.

Hugging Face is a valuable resource for building chatbots as it offers a variety of advantageous features. The platform provides access to a plethora of pre-trained models that can be used for different natural language processing tasks, without the need to train a model from scratch. The platform also offers easy to use libraries and pipelines that simplify the development process and make it more efficient. Furthermore, the large community of developers and researchers on the platform are a great source of support and guidance. Additionally, the pre-trained models can be adapted to suit the specific needs of a chatbot, providing a greater degree of flexibility. Furthermore, the platform provides pre-trained models for over 100 languages, making it easier to build chatbots for a wide range of languages. Lastly, the pre-trained models available on the platform covers a wide range of tasks, from information extraction, text classification, question answering, text generation, translation, speech recognition, audio classification, object detection, image classification, tabular data analysis and reinforcement learning, this allows developers to choose the best model for the task they are working on.

How do pipelines function in Transformers?

Through the pipeline method, they offer a simple API for carrying out inference over various activities. The entire process of each Natural Language Processing activity, including text cleaning, tokenization, embedding, etc., is captured by them.

Architecture

The Transformer architecture is a neural network-based model that is widely used for natural language processing tasks such as machine translation and text summarization. The architecture consists of an encoder and a decoder, which are made up of multiple layers of self-attention and feed-forward layers. The encoder is responsible for processing the input and creating a representation of it and the decoder generates the output. One of the key components of the transformer architecture is Multi-Head Attention. This allows the model to focus on multiple parts of the input simultaneously, enhancing the model's ability to understand the context of the input. Multi-head attention is achieved by dividing the model's attention mechanism into multiple "heads", each of which can attend to a different aspect of the input. Additionally, the Position-wise Feed-Forward Network is used to further process the output of the attention mechanism. It is composed of two linear layers with a ReLU activation function in between, increasing the model's capacity and allowing it to learn more complex relationships between the input and output. To normalise the output of each layer, Layer Normalisation is used, which helps to improve the stability and performance of the model. Lastly, the Embedding component is used to represent words as vectors. By incorporating these components in the transformer architecture, the model can handle input sequences of varying lengths and produce outputs that are more contextually relevant, making it a powerful model for natural language processing tasks such as chatbot.

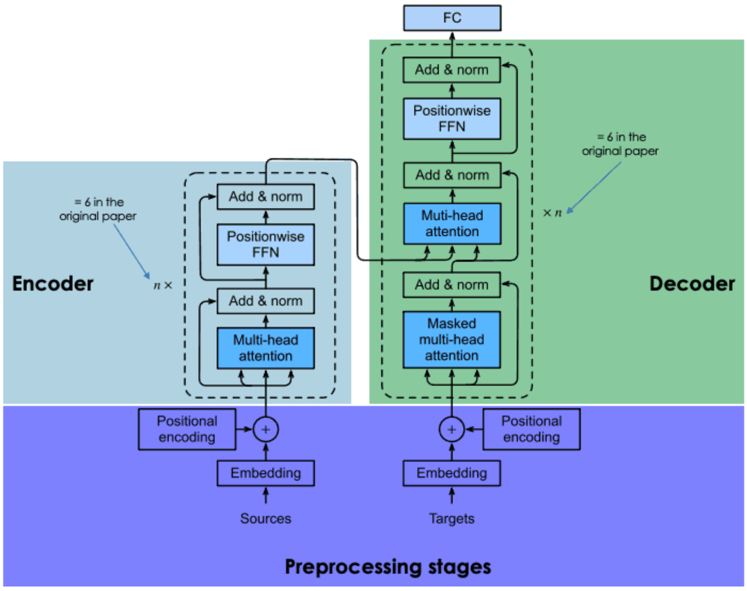

Figure 4.4: Transformer Architecture

Chapter 5

Experimental Setup

Hyperparameter optimization is the methodical exploration of different hyperparameter settings to identify the optimal configuration that leads to the best performance of a machine learning model. Hyperparameters are parameters that are set before training a model and control the learning process, such as the learning rate, the number of hidden layers, or the number of neurons in a layer. The goal of hyperparameter optimization is to find the combination of hyperparameters that results in the best performance on a validation set. This can be achieved through manual trial and error or through automated techniques such as grid search, random search, or Bayesian optimization. Hyperparameter optimization is a crucial step in the machine learning pipeline as it can have a significant impact on the model's performance. We applied hyperparameter optimization on our RASA model to enhance its performance, using a combination of methods to explore different parameters. After conducting multiple experiments and evaluating the model's performance on a validation set, we were able to identify the optimal combination of hyperparameters that resulted in the best performance. This process of hyperparameter optimization allowed us to fine-tune our RASA model and improve its ability to accurately classify and respond to user input. As a result, our model became more robust and better suited for our specific use case. The parameters we used are mentioned below -

Pipeline tuning

It is advised to use the default open source library spaCy in the RASA NLU pipeline because it supports word embeddings that have already been trained for the model. However, since we created our own dataset, we used a pipeline with a tokenizer called WhitespaceTokenizer, which also includes a count vectorizer for feature extraction, to train our dataset that has also helped us get more accurate results.

Multi intent Classifier

Next, we combined the DIETClassifier with a multi intent classifier since users frequently ask a variety of questions with the same general purpose that can have more than one intent. Therefore, by using the multi intent classifier, the model can combine the intents of these many question types to predict the correct answer as the output.

Handling class imbalance

Then to further improve the accuracy of our model we tried to keep a balance between all of the intent classes by using classification algorithms that solve class imbalance.

GPU

Another situation we had to deal with was that TensorFlow by default blocks all of the GPU memory that is available for the process that is executing, which might affect performance as memory distribution is hindered. To avoid this situation, we set the environment option TF FORCE GPU ALLOW GROWTH to True.

Chapter 6

Experimental Results and Discussion

6.1 Evaluation and Analysis of RASA and Transformer model

The results of the evaluation test are shown in the following table. We observed that Rasa had better accuracy and F1 scores than Transformer. Specifically, RASA NLU achieved 96.3% for accuracy and 97.7% for F1, while Transformer had scores of 79.07% and 55.3% respectively. This is because RASA NLU is a more specialised model specifically designed for intent classification and entity extraction, whereas transformer models are more general-purpose models that can be used for a wide range of NLP tasks, including language translation and text generation. Transformer models are often pre-trained on large datasets and may require more computational resources for fine-tuning. On the other hand, RASA NLU offers more flexibility in terms of customization and fine-tuning for specific use-cases. Additionally, transformer models extract answers based on the given context, which may lead to a lack of accuracy and F1 scores in certain test cases.

Model	Accuracy	F1
Rasa Nlu	96.3%	95.7%
Transformer	72.02%	67.3%

Table 6.1: Model evaluation results

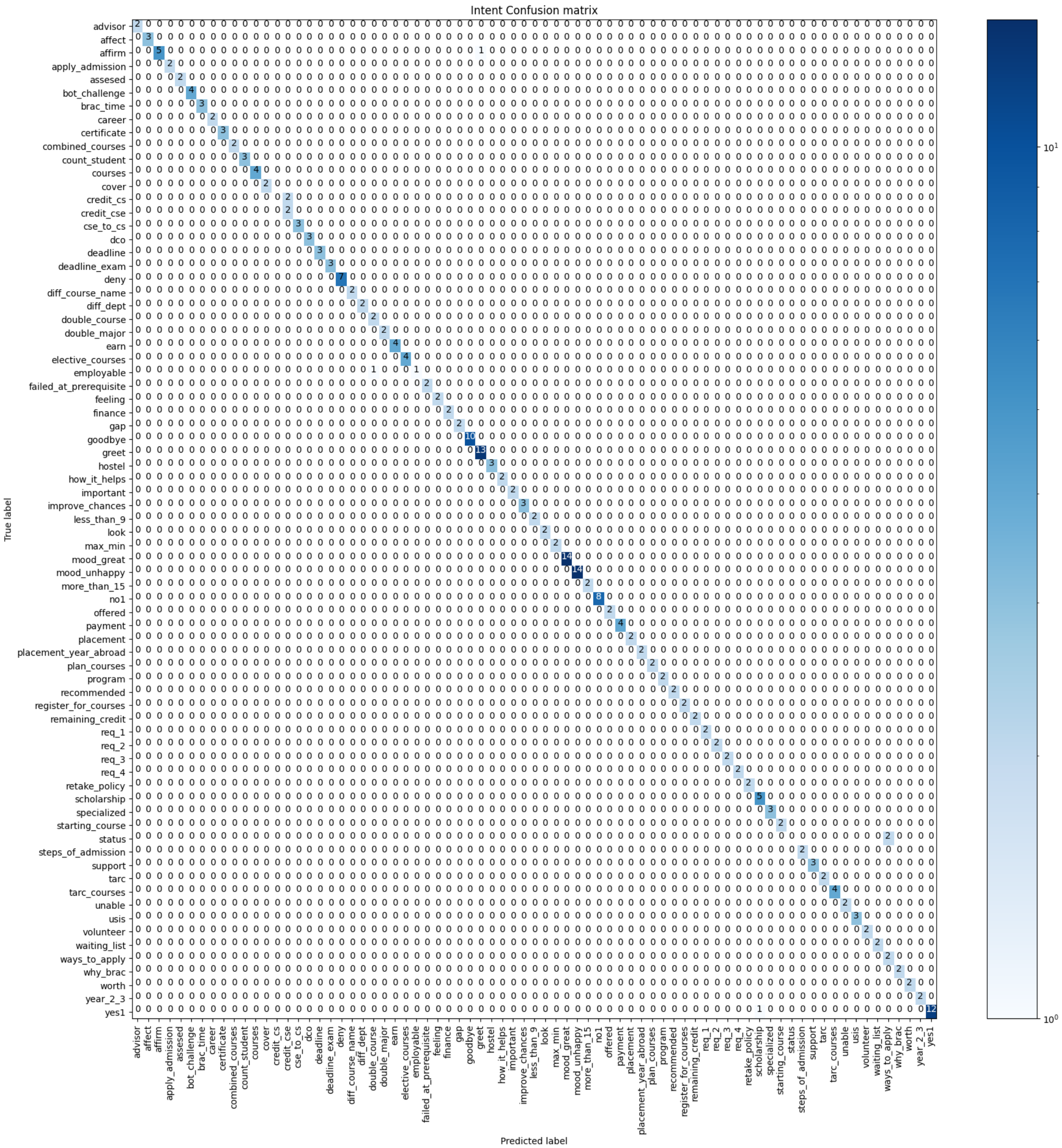

Figure 6.1: Intent Confusion Matrix of RASA

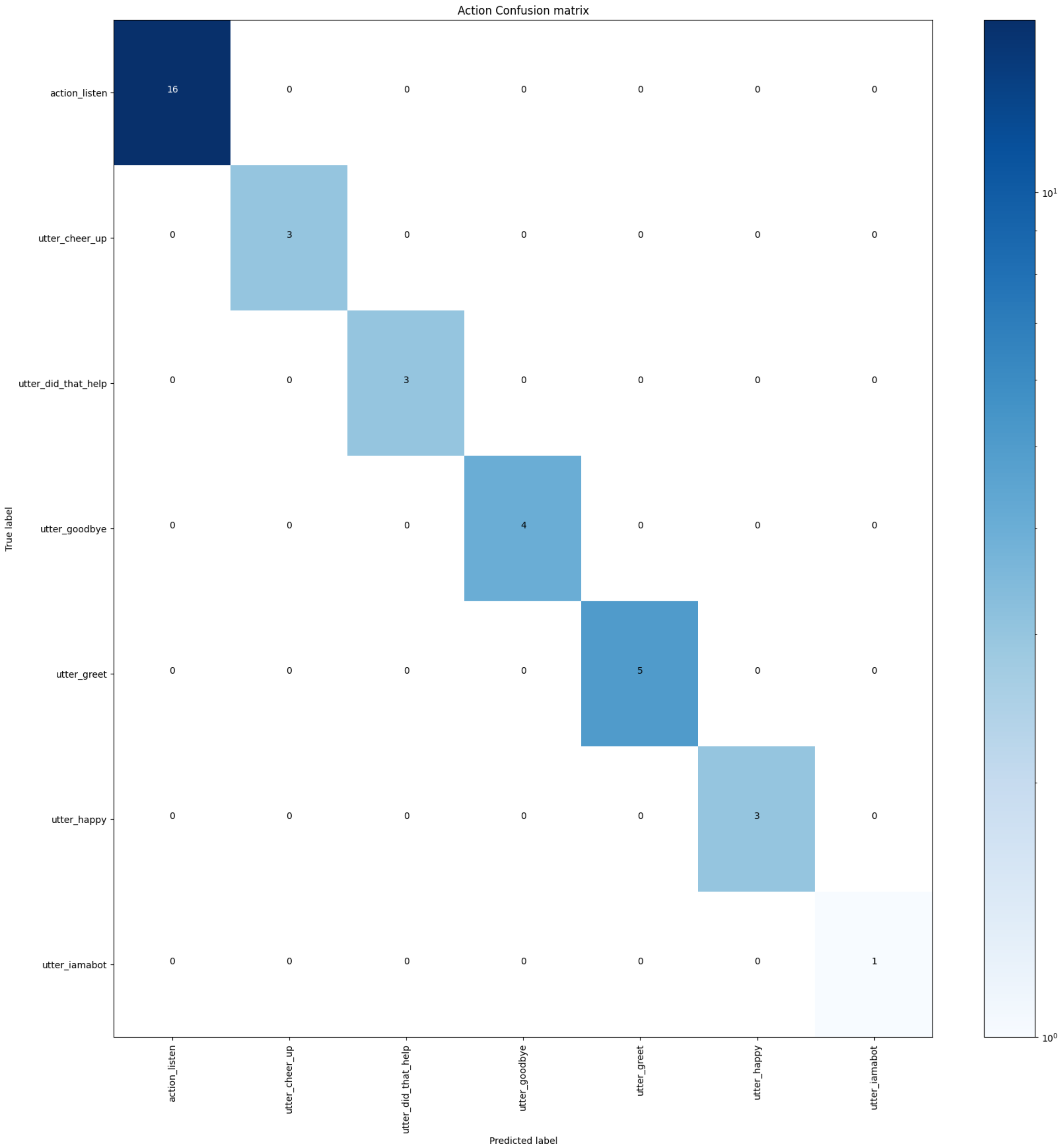

Figure 6.2: Action Confusion Matrix of RASA

We were able to analyse the RASA model's progress in training the data, classifying, and extracting intent and entity due to the intent and entity confusion matrix for RASA NLU model. Since the diagonal elements of the matrix are all coloured, with the majority of them having dark colours that represent a higher score for that particular word, it can be seen from the figures of the confusion matrix that the bot was able to predict almost all of the intent classes and entities correctly. Hence, the accuracy score and F1 score obtained as a result was 96.3% and 95.7% respectively.

6.2 Results

Our bot was created with the capability to understand the user's queries and recognize the specific entities and intents for which it was programmed. The bot was able to furnish the appropriate response to the user by utilising data that had been correctly trained and pre-processed by the models. Even though there were some cases where the bot experienced difficulty in providing accurate answers, on the whole, the results were highly satisfactory. Despite this, we remain committed to enhancing the bot and preparing it to tackle future challenges. After conducting our research and experiments, it was concluded that RASA Model works better than Transformers. By using advanced machine learning techniques, such as natural language processing and deep learning, we aim to improve the bot's ability to understand the nuances of human language and respond accordingly. Additionally, we are also working on expanding the bot's knowledge base to ensure it can handle a wider range of queries and provide more accurate responses.

6.3 Limitations

6.3.1 RASA

- Generating NLU Data - The Rasa framework includes NLU , which conducts intent detection, entity extraction, and answer-retrieval. NLU will accept a user inquiry.like "Please transfer 500$ to Mike" and return structured data like:

  Creating ML/AI solutions for any problem can be challenging, and NLU problems are no exception. The primary challenge for building NLU models is acquiring and using sufficient data, and developing models that are suitable for use in real-world applications is even more difficult.

- Conversation Driven Development - It refers to using actual user conversations as the primary guide for developing a chatbot. This approach allows the bot to learn how people naturally speak, including commonly used slang, synonyms, and abbreviations that may be difficult for a bot developer to anticipate and include in the training data.

- Capture Real Data - When it comes to training RASA NLU models, a large amount of data is required for efficient training. As a result, bot builders often rely on text generation tools and programs to quickly create a sufficient number of training examples. These tools are effective for generating large amounts of training data, but they can also lead to overfitting, where the model is trained on information that is not representative of what a real user would say. To avoid this problem, it is important to use as much real user interactions or conversations as possible for training data. While the model may make mistakes initially, training and evaluating on actual user interactions will help the model to perform better in real-world situations.

- Avoiding Intent Mismatch: Depending on the featureizers that have been integrated into the NLU pipeline, character- and word-level features generated from the training samples are used to identify intents. The intent classifier may become confused if several intentions contain the same terms in the same order.

- Extracting entities - An NLU model needs a lot of training data on entities like names, addresses, and cities in order to generalise well. RASA offers a variety of options for entity extraction, but it can be difficult to determine which method to use. For pre-trained extraction models, RASA provides two practical options: the DucklingEntityExtractor and SpacyEntityExtractor. These extraction models don't require you to label the entities they support in your training data because they have already been trained on a sizable set of data. In addition to these pre-trained extractions, RASA also provides tools for other types of entity extraction

6.3.2 Transformer

- Memory constraints: The model requires a significant amount of memory to run, which can be a limitation for some devices or systems.
- Computational resources: Training large transformer models requires a significant amount of computational resources, such as powerful GPUs.
- Overfitting: Transformer models have a large number of parameters, which can lead to overfitting if not properly regularised.
- Training data bias: The calibre and variety of the training data have a significant impact on the performance of the model. The performance of the model may be impacted by biassed or undiversified training data.
- Limitations with understanding certain tasks: Transformer models have been shown to be less effective in certain tasks such as logical reasoning, common sense and understanding physical interactions.

Chapter 7

Conclusion and Future Work

Natural language processing plays a crucial role in the development of sophisticated chatbot systems. With the goal of building an effective FAQ bot system for our university students, we conducted a study comparing the two NLP framework based models, RASA and Transformers. After thorough research and experimentation using a dataset we created, we reached the conclusion that RASA's framework offers a wealth of features but may have a steeper learning curve for understanding how the features work together. Some developers have noted that RASA may not be the most user-friendly option for beginners or those without previous experience with chatbot development frameworks. In contrast, the Transformer framework is highly complex in terms of interpretation, which can make processing more challenging. However, through our investigation, we found that the RASA framework performed better than the Transformer framework and also provides more room for future enhancements to our chatbot that we plan to work on in the future.

From our perspective, to make a RASA chatbot better, we would ensure that the training data is high-quality, diverse and representative of the use cases and user inputs that the chatbot will encounter in real-world scenarios, incorporate external knowledge bases such as Wikipedia or DBpedia to enhance the chatbot's ability to provide more accurate and complete responses, consider adding more functionalities such as sentiment analysis or language translation to make the chatbot more versatile and useful for users, investigate the use of advanced NLP techniques such as transformer models or pre-trained language models to improve the chatbot's ability to understand and respond to user inputs and regularly monitor the chatbot's performance and gather feedback from users to identify areas for improvement and make necessary adjustments. By implementing these strategies, we believe we can foster the RASA chatbot's precision and effectiveness, as well as its ability to understand and respond to user input information, and provide the academics and interested parties of Brac University with a highly integrated virtual assistant that will substantially reduce all aggravation and ensure accountability.

Bibliography

[1] J. Rahman, "Implementation of alice chatbot as domain specific knowledge bot for brac u (faq bot)," Ph.D. dissertation, BRAC University, 2012.

[2] E. Cambria and B. White, "Jumping nlp curves: A review of natural language processing research," *IEEE Computational intelligence magazine*, vol. 9, no. 2, pp. 48–57, 2014.

[3] P. B. Brandtzaeg and A. Følstad, "Why people use chatbots," in *International conference on internet science*, Springer, 2017, pp. 377–392.

[4] B. R. Ranoliya, N. Raghuwanshi, and S. Singh, "Chatbot for university related faqs," in *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, IEEE, 2017, pp. 1525–1530.

[5] S. Tembhekar and M. Kanojiya, "A survey paper on approaches of natural language processing (nlp) j," *International Journal of Advance Research, Ideas and Innovations in Technology*, vol. 3, no. 3, pp. 1496–98, 2017.

[6] A. Xu, Z. Liu, Y. Guo, V. Sinha, and R. Akkiraju, "A new chatbot for customer service on social media," in *Proceedings of the 2017 CHI conference on human factors in computing systems*, 2017, pp. 3506–3510.

[7] A. Følstad, C. B. Nordheim, and C. A. Bjørkli, "What makes users trust a chatbot for customer service? an exploratory interview study," in *International conference on internet science*, Springer, 2018, pp. 194–208.

[8] T. Lalwani, S. Bhalotia, A. Pal, V. Rathod, and S. Bisen, "Implementation of a chatbot system using ai and nlp," *International Journal of Innovative Research in Computer Science & Technology (IJIRCST) Volume-6, Issue-3*, 2018.

[9] S. Ruder, "A review of the neural history of natural language processing," *AYLIEN, October*, vol. 1, 2018.

[10] M.-H. Su, C.-H. Wu, K.-Y. Huang, and Q.-B. Hong, "Lstm-based text emotion recognition using semantic and emotional word vectors," in *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, IEEE, 2018, pp. 1–6.

[11] E. Batbaatar, M. Li, and K. H. Ryu, "Semantic-emotion neural network for emotion recognition from text," *IEEE Access*, vol. 7, pp. 111 866–111 878, 2019.

[12] D. Haryadi and G. P. Kusuma, "Emotion detection in text using nested long short-term memory," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 6, 2019.

[13] S. Natale, "If software is narrative: Joseph weizenbaum, artificial intelligence and the biographies of eliza," *new media & society*, vol. 21, no. 3, pp. 712–728, 2019.

[14] A. Chizhik and Y. Zherebtsova, "Challenges of building an intelligent chatbot.," in *IMS*, 2020, pp. 277–287.

[15] A. Jiao, "An intelligent chatbot system based on entity extraction using rasa nlu and neural network," in *Journal of Physics: Conference Series*, IOP Publishing, vol. 1487, 2020, p. 012 014.

[16] J. Joy, A. Kannan, S. Ram, and S. Rama, "Speech emotion recognition using neural network and mlp classifier," *IJESC*, vol. 2020, pp. 25 170–25 172, 2020.

[17] M. Sajjad, S. Kwon, *et al.*, "Clustering-based speech emotion recognition by incorporating learned features and deep bilstm," *IEEE Access*, vol. 8, pp. 79 861–79 875, 2020.

[18] J. J. Bird, A. Ekárt, and D. R. Faria, "Chatbot interaction with artificial intelligence: Human data augmentation with t5 and language transformer ensemble for text classification," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–16, 2021.

[19] C. Gomathy, P. Meghana, and P. V. V. Reddy, "Voice based university information chatbot system," 2021.

[20] G.-J. Hwang and C.-Y. Chang, "A review of opportunities and challenges of chatbots in education," *Interactive Learning Environments*, pp. 1–14, 2021.

[21] N. Darapaneni, S. Raj, V. Sivaraman, S. Mohan, A. R. Paduri, *et al.*, "Lstm-rasa based agri farm assistant for farmers," *arXiv preprint arXiv:2204.09717*, 2022.

[22] J. Kim, S. Chung, S. Moon, and S. Chi, "Feasibility study of a bert-based question answering chatbot for information retrieval from construction specifications," in *2022 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, IEEE, 2022, pp. 0970–0974.

[23] M. M Alshater, "Exploring the role of artificial intelligence in enhancing academic performance: A case study of chatgpt," *Available at SSRN*, 2022.

[24] D. S. Mishra, A. Agarwal, B. Swathi, and K. C. Akshay, "Natural language query formalization to sparql for querying knowledge bases using rasa," *Progress in Artificial Intelligence*, vol. 11, no. 3, pp. 193–206, 2022.

[25] P. Vassilakopoulou, A. Haug, L. M. Salvesen, and I. O. Pappas, "Developing human/ai interactions for chat-based customer services: Lessons learned from the norwegian government," *European Journal of Information Systems*, pp. 1–13, 2022.

[26] K. Wenzlaff and S. Spaeth, "Smarter than humans? validating how openai's chatgpt model explains crowdfunding, alternative finance and community finance.," *Validating how OpenAI's ChatGPT model explains Crowdfunding, Alternative Finance and Community Finance.(December 22, 2022)*, 2022.