Developing a Mental Health Virtual Assistance (Chatbot) for Healthcare Workers and their Families

by

Ali Zamani

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

University of Alberta

Abstract

Approximately 1 in 3 Canadians experiences addiction or mental health challenges at some point in their lifetime. Unfortunately, there are multiple barriers to accessing mental healthcare, including system fragmentation, episodic care, long wait times, and insufficient support for health system navigation. In addition, stigma may further reduce an individual's likelihood of seeking support. Digital technologies present new and exciting opportunities to bridge significant gaps in mental healthcare service provision, reduce barriers pertaining to stigma, and improve health outcomes for patients and mental health system integration and efficiency. Chatbots (ie, computer programs designed to simulate conversation with human users, especially over the Internet) may be explored to support those in need of information or access to services and present the opportunity to address gaps in traditional, fragmented, or episodic mental health system structures on demand with personalized attention. The recent COVID-19 pandemic has exacerbated even further the need for mental health support among Canadians and called attention to the inefficiencies of the system. As healthcare workers and their families are at an even greater risk of mental illness and psychological distress during the COVID-19 pandemic, this technology is first piloted to support this vulnerable group.

In this project, we developed mental health software Mira Chatbot to support healthcare workers and their families in the Canadian provinces of Alberta and Nova Scotia. The software features four major elements: the Mira Chatbot, Mira Resource Portal, Mira Dataset, and Mira Interface. Mira Chatbot's

primary purpose is to provide strategic resources for users responding to the custom needs that they share. Users provide their unique information through two main tasks defined within the Mira Chatbot: intent detection and entity extraction. Intent detection is a process where Mira Chatbot identifies the needs of the user based on a chat experience with an accuracy rate of 99.1%. Through the task of entity extraction, Mira Chatbot recognizes important keywords from a sentence with an accuracy rate of 95.4%.

Preface

Ethics approval was granted on August 12, 2021, by the University of Alberta Health Research Board (case Pro00109148) and on April 21, 2022, by the Nova Scotia Health Authority Research Ethics Board (case 1027474). All data and computer code will be password-protected and stored on a secure server at the University of Alberta in Canada.

Our paper "Developing, Implementing, and Evaluating an Artificial Intelligence Guided Mental Health Resource Navigation Chatbot for Healthcare Workers and Their Families During and Following the COVID-19 Pandemic: Protocol for a Cross-sectional Study" is accepted for publication in the Journal of Medical Internet Research (JMIR-Research protocols). The authors of the paper are Jasmine M. Noble, Ali Zamani, MohamadAli Gharaat, Dylan Merrick, Nathanial Maeda, Alex Foster, Isabella Nikolaidis, Rachel Goud, Eleni Stroulia, Vincent Agyapong, Andrew J. Greenshaw, Simon Lambert, Dave Gallson, Ken Porter, Deb Turner, Osmar Zaiane.

Acknowledgements

Words cannot express my gratitude to my professor and my supervisor for their invaluable patience and feedback. I also could not have undertaken this journey without my defense committee, who generously provided knowledge and expertise. Additionally, this endeavor would not have been possible without the generous support from the Mood Disorder Society of Canada, who financed my research.

I would be remiss in not mentioning my family, especially my parents. Their belief in me has kept my spirits and motivation high during this process.

This project received funding to support student involvement through the Mood Disorders Society of Canada and the Mathematics of Information Technology and Complex Systems Accelerate grant. This study includes multiple partnerships, including support from a committee of experts entitled the Expert Advisory Committee, as well as Mood Disorders Society of Canada volunteers. The author would like to give thanks to these members for volunteering their time to validate resources that the chatbot will ultimately draw from as well as for the insights provided in the beta testing of the chatbot itself.

Contents

List of Tables

List of Figures

Glossary

Artificial Intelligence Markup Language (AIML)

Artificial Intelligence (AI)

Artificial Linguistic Internet Computer Entity (ALICE)

Average Intent and Entity (AIE)

Bilingual Evaluation Understudy Score (BLEU)

Coached Conversational Preference Elicitation (CCPE)

Cognitive Behavior Therapy (CBT)

Conditional Random Field (CRF)

Conversational Agents (CAs)

Dual Intent and Entity Transformer (DIET)

Easy Data Augmentation (EDA)

Embodied Conversational Agents (ECA)

Long Short Term Memory Network (LSTM)

Mean Reciprocal Rank (MRR)

Mental Health Resource Assistance (MIRA)

Natural Language Understanding (NLU)

Natural Language Processing (NLP)

Question Answer (QA)

Radial Basis Function (RBF)

Recurrent Neural Network (RNN)

Sequence to sequence (Seq2Seq)

Social Network Platform for Physicians (SERMO)

Support Vector Machine (SVM)

Term Frequency-Inverse Document Frequency (TF-IDF)

Ubuntu Dialogue Corpus (UDC)

Chapter 1

Introduction

In recent years chatbots, such as Amazon Alexa, Google Assistant, and Apple Siri have become popular among people [45]. According to the Oxford English Dictionary, a chatbot is defined as follows:

chatbot (n.):
A computer program designed to simulate conversation with human users, especially over the Internet.

Chatbots are more technically known as Conversational Agents (CAs) in the scientific literature. The phrases "chatbot" and "conversational agent" are used interchangeably throughout this thesis.
CAs are designed either for casual chatting (in which case the system is referred to as an open-ended agent) or to provide the user with relevant information for a specific task (such as a flight reservation), in which case the system is referred to as a task-oriented agent. Some CAs simply employ text-based input and output, whereas others use a variety of input and output modalities (for instance, speech). The more sophisticated versions of CAs are Embodied Conversational Agents (ECAs), which have an animated visual representation (face or body) on-screen. A good example of ECA is Ellie [28] , a computer program that diagnoses depression and post-traumatic stress disorder.

The way CAs display and manage communication is arguably the essential aspect [140]. There are several options, ranging from simple rule base systems

like Eliza [148], Parry [31], and Artificial Linguistic Internet Computer Entity (ALICE) [142] to very complicated representations based on deep neural networks [140]. For example, a health chatbot that helps patients must address at least two tasks [18]. One task is detecting the requests of the patients with regards to problems that they are facing, i.e., whether the patient needs to connect to a doctor or only needs a general guide. There are a lot of natural languages understanding techniques [119] for detecting the user's intent from spoken or written text. Another task is to generate a written or spoken response according to the detected intent.

In this research, we investigate strategies for creating a healthcare conversational agent by looking into medical applications that offer relevant resources to healthcare workers and their families with various mental health issues such as depression and anxiety. As a result, the users can speak with the chatbot about their mental health problems and obtain relevant resources.

1.1 Motivation

Mental disorders are the leading cause of disability in Canada; approximately 1 in 3 Canadians experience substance use or mental health disorders in their lifetime [86], [98]. Unfortunately, there are also significant gaps in care. According to a 2018 study, 5.3 million Canadians expressed a need for mental health services in 12 months [129]. Of these, 48.8% reported that their mental health needs were not being met [129]. Of those reporting unmet or only partially met needs, 78.2% identified personally circumstances, including affordability and not knowing where to receive help, as barriers to care [129].

Barriers to seeking support include stigma, denial, concerns over privacy, and difficulty connecting effectively with a care provider [90], [92], [126], [151]. In addition, prominent access issues include fragmented or episodic care, lack of support for navigating the healthcare system and connecting with an appropriate provider or specialist, and long wait times to access services [90], [92], [97], [118], [126], [131], [151].

Canada's publicly funded healthcare system is administered and delivered

by the provinces and territories through public health authorities or entities operating on a nonprofit basis. Hospitals and other healthcare services deemed medically necessary must be insured by provincial and territorial plans. Many citizens acquire additional private insurance to pay for unfunded services [130]. Mental healthcare coverage across Canada varies widely, and many available services are not deemed medically necessary despite mental health being increasingly recognized as fundamental to health. Only mental health services received in hospital settings are covered universally by Canada's public health system. Mental healthcare in Canada is unique as it is provided by a "meshwork" of local hospitals, community programs, residential care centers, private practices, and more [141]. Adding to this complexity, many organizations are particular to one jurisdiction or specific to a certain type of mental health concern.

Canada's healthcare system has been described as a "labyrinth" where individuals may even resort to paying private sector agents to act on their behalf to find and connect with services, further exacerbating socioeconomic inequalities in access to care [130]. Many Canadians who have received unsatisfactory help for their mental health needs reported "not knowing where to go" as a primary barrier to care [59]. Testimonies of Ontario-based patients and caregivers highlight feelings of confusion in having to navigate this system on their own, resulting in longer delays in care access [59]. Wait times have been as long as two and a half years [102], with many individuals receiving no documented care [10]. The Wait Time Alliance 2014 Report Card highlights the lack of system coordination and insufficient staff and resources as determinants of long wait times to access mental health services in Canada [59]. Heightened demands for care and lack of navigation toward community services contribute to overcrowding within emergency departments, with a 75% increase in mental health-related visits for patients aged 5 to 24 years since 2006 [67]. System integration and system navigation support services between community-based health and social services and formal healthcare providers have been identified as a key policy issue in Canada and other jurisdictions such as the United Kingdom [14], [36], [55], [67], [68], [96], [101], [110], where lack of knowledge

of service options often poses a barrier to referrals from healthcare providers to community-based services [14], [39], [79], [89], [101].

In 2019, an outbreak of COVID-19 (SARS-CoV-2) resulted in a global pandemic. By early 2022, COVID-19 had spread worldwide, with 334 million known cases and 5.5 million deaths [150]. In anticipation of a high volume of severe hospitalizations with technical respiratory needs, Canadians were asked to self-quarantine or practice social distancing to reduce the burden on health systems [78]. This intensified the mental health crisis within Canada; according to an Angus Reid Institute poll, 50% of Canadian respondents indicated that their mental health had worsened over the COVID-19 pandemic, with 10% indicating that it had worsened "a lot" [111]. Multiple public surveys deployed during the pandemic reported respondents' experiences of multiple mental health stressors such as economic instability, fear of becoming sick, and life disruption as a cause of the COVID-19 pandemic, resulting in stress, anxiety, and depression [87]. A recent Ontario survey revealed that approximately 25% of respondents reported unmet mental health needs due to the pandemic, moderate to severe anxiety, and symptoms of loneliness and depression [25].

Following the adverse mental health outcomes observed in this and previous epidemics and pandemics [2], [82], it is widely agreed by the international medical community that a wave of the widespread need for mental health-related services will result from the pandemic that will persist beyond the acute phase [87]. Within the Canadian context, in consideration of the pre-pandemic prevalence of mental illnesses such as depression (lifetime prevalence of 5% in Canadian men and 10% in Canadian women [104]) and insomnia (12-month prevalence ranging from 9.5% to 24% [21], [88], [134]) and existing gaps in service delivery, public health practitioners and policy leaders must urgently consider innovative ways to connect a large portion of the Canadian public with appropriate services in an efficient manner.

In addition to the negative impact on Canadians' mental health, many services have faced disruptions because of adjusting to social distancing and capacity restrictions, often eliminating face-to-face instead of remote service settings [114]. Many countries have developed new web-based mental health

information sites or phone lines to provide coping support [85]. For those facing modest mental health burdens, connection with these web-based resources can aid in self-management and may provide a bridge before professional support is available [114]. With these changes in offered services and increased web-based application use, navigation to individual personalized, timely, and relevant resources is increasingly important.

healthcare workers and their families are particularly vulnerable during pandemics and, in reflection of anticipated needs, are the target participant group for this project. healthcare workers face an increase in mental health risk factors, including anxiety, burnout, and depression, because of factors such as increased exposure and risk of disease transmission to themselves or others (e.g., family and friends) and unsafe (e.g., personal protective gear shortages) or stressful working conditions [21]. Of concern is the trauma that healthcare workers witness within the workplace, how their ongoing work limits their ability to address their mental health concerns, and how they may be processing these experiences when they are outside the workplace with more time to reprocess what they see. For example, a recent umbrella review of meta-analyses found that the prevalence of anxiety and depression among healthcare workers was relatively high at 24.94% [114]. A recent survey by the Canadian Centre for Addiction and Mental Health (January 2022) documented an increase in self-reported symptoms of severe anxiety (37% compared with 23.5% in summer 2021) and depression (35.7% compared with 24.8% in summer 2021) among healthcare workers and other frontline workers [2], suggesting that mental health problems are being exacerbated with time. Together, these risk factors may lead to healthcare workers resigning from their positions, increasing staff shortages and, in turn, pressures on the remaining employees [85].

Digital technologies provide an opportunity to bridge service gaps, increase points of access to and knowledge of the mental healthcare system and existing services, enhance mental health literacy, and permit more excellent health system and social system integration, which could improve health and social system coordination, efficiency, patient navigation, satisfaction, and overall health outcomes [85]. In addition, efficiencies realized through new technology

may lower healthcare costs, enabling resources to be redirected to other areas of priority. Artificial intelligence (AI) presents the opportunity to bypass barriers inherent to traditional brick-and-mortar health system structures, meet individuals in need in a discrete and personalized way, and connect them with services on time regardless of where they are. For example, commonly cited factors identified for why individuals choose to access web-based services include 24-hour accessibility, ease of accessibility despite geographic location, anonymity, and privacy [53], [63], [70], [71]. Although further analysis is required in the context of mental healthcare, research suggests that patients report greater comfort or preference in disclosing sensitive health information to a computer or technological device than to a human [16], [76]. AI then presents the opportunity to address social stigma as a barrier to care, which may hinder an individual's drive or motivation to seek access to care.

Existing evidence supports the use of health chatbots for empowering users to engage in physical activity and consumption of nutritious food and increasing patient access to health information, among other benefits [12], [13], [40], [146]. Although human-computer interaction technology itself is not new as a concept, evaluative research on the use of applied AI as a tool for bridging gaps in mental healthcare is limited. More specifically, although chatbots currently show promise in a variety of healthcare settings [3], [66], [137], there is limited information on their effectiveness in supporting mental health system navigation [66], [103]. As such, the use of a conversational chatbot for this general purpose is novel. In addition, existing chatbots are commonly tailored to address one or a limited range of mental health issues [24]. Our conversational chatbot, the Mental Health Intelligent Information Resource Assistant (MIRA), seeks to support various mental health disorders and considerations.

Fortunately for the emergence of digital health intervention options, technology uptake among the general public has been substantive. There are 3.96 billion internet users internationally. In Canada, 91% of the population is estimated to be actively using the Internet, and 85% have a cell phone (65% have a smartphone specifically) [1]. As such, there remain significant opportunities to use existing and widely adopted technological infrastructure to bridge

significant gaps in care and improve health outcomes for Canadians.

1.2 Thesis Contributions

In this project, our pan-Canadian, multidisciplinary team of subject matter experts, including individuals with lived experience, members of the Indigenous community, clinicians, and psychiatry and computing science experts, report on the design, implementation, and anticipated evaluation of MIRA, a domain-specific AI-enabled chatbot able to understand standard taxonomies in the mental health domain and respond with relevant, appropriate resources aligned with the clients' intents and needs. The MIRA Chatbot is informational only and does not provide medical advice (i.e., it does not diagnose or provide treatment recommendations), nor does it replace a counselor or mental health professional. This project population group of interest was healthcare workers and their families in the Canadian provinces of Alberta and Nova Scotia.

An additional component has been developed to complement the chatbot's functionality (a resource management portal the MIRA Resource Portal). MIRA does not search for resource recommendations extracted from the open Internet. Instead, MIRA draws recommendations from the MIRA Resource Portal, which not only facilitates the input and an expert validation of mental health resources for use by the chatbot but also automatically monitors validated resources for any changes after approval and subsequently reports them to the editors. The MIRA resource portal was developed by other MIRA team members, so it is not part of the contributions of this thesis. However, I actively participated in the discussions regarding the design of the portal and particularly the makeup of the interface for communication between the portal and the chatbot, which is not explicitly highlighted in this manuscript. Also, we built our dataset from scratch to train the chatbot, in which one of the MIRA team members manually wrote some examples, and then by applying some data augmentation techniques, the dataset was expanded. The dataset can also be used for other mental health chatbots.

To summarize, in this thesis, our accomplishments include:

1. Developed, implemented, and launched a mental health chatbot named MIRA Chatbot to support healthcare workers and their families. This was done in collaboration within a team, each member of which was responsible for different components but interconnected within the chatbot system.

2. Constructed a dataset named MIRA Dataset for training the machine learning model that the MIRA Chatbot is using to automatically identify intent. The dataset was initially manually curated and then systematically augmented with synthetic samples using data augmentation techniques.

3. Compared and contrasted different machine learning models for intent detection and entity extraction tasks in order to select the most appropriate one to integrate within the chatbot.

1.3 Thesis Organization

The rest of the thesis is organized as follows. Chapter 2 (Conversational Agents) provides a detailed review of existing chatbot frameworks as well as chatbot design techniques. Furthermore, we define several evaluation metrics for evaluating the CAs. We conclude this chapter by, describing some well-known datasets, and presenting a comprehensive overview of CAs in the literature and industry. Chapter 3 (Method) describes the machine learning models used for the MIRA Chatbot ,the MIRA dataset, and the MIRA design and implementation. Chapter 4 (Evaluation of the MIRA Chatbot) presents the performance of the machine learning model, that we used for the MIRA Chatbot. Chapter 5 (Conclusion and Future Work) discusses future work and summarizes the thesis and conclusions of the techniques used to improve the MIRA Chatbot.

Chapter 2

Conversational Agents

2.1 Introduction

Conversational agents employ the communication channel of the web and computational linguistics methods to interpret and react to user statements in ordinary natural language. The user asks a simple question, and the chatbot intelligently responds in a human-like manner, bridging the communication gap. The chatbot deciphers the user's query, selects a suitable response, and responds seemingly in a natural way. This complicated fusion technology has yet to be implemented since it is based on the theoretical concept of near-human thinking [127], which can accomplish a wide range of activities while facing a wide range of problems. Online customer service with frequently asked questions capabilities, employment processes, patient queries with trusted responses, and chatbots like Amazon Alexa, Siri, and Google Home are just a few examples [127]. Conversational agents participate in the contextual discussion using Natural Language Processing (NLP) and other artificial intelligence approaches. Natural Language Understanding (NLU) recognizes the intent and extracts information from users' words, regardless of how they are expressed, including errors in grammar, punctuation, and other areas. The chatbot must then use NLP typically with machine learning to select the correct response from a set of response variations [100] (Fig. 2.1).

In this chapter, some of the existing chatbot design techniques are provided, which is followed by an introduction to the most influential chatbots and chatbot frameworks. The evaluation metrics for assessing chatbots are

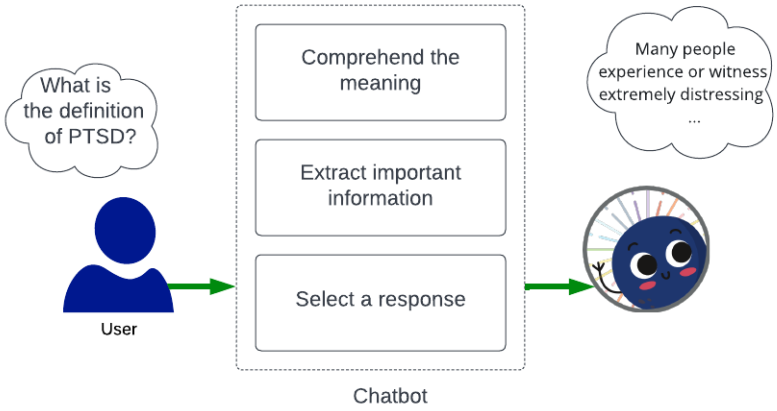

Figure 2.1: General view behind a conversational agent.

then described. Chapter two concludes with a review of training datasets and the current research analyzing their impacts and implementation methods.

2.2 Chatbots Design Techniques

In this section, a classification of chatbots based on the tasks they perform or approaches they employ is presented. A simple categorization of chatbots can be done as follows [127]:

- **Rule-based**: The rule-based chatbots are programmed to follow a set of rules and patterns. Then, a heuristic algorithm is utilized to select the most likely response. Due to the limited number of rules, the capabilities of the rule-based chatbots are limited, and there is a risk of low accuracy responses, resulting in an unpleasant user experience [6], [41].

- **Retrieval-based**: The retrieval-based chatbots are designed to follow graphs or directed flows. To offer the best possible response, the retrieval-based chatbots are trained to select a response from a database of predefined responses. To find the most appropriate response, the chatbots employ keyword matching, machine learning, and deep learning techniques. It is worth mentioning that, regardless of the technology, these chatbots deliver predefined responses and do not produce a new

output [5], [6], [132]. However, since a trained machine learning model selects the responses, more accurate responses are chosen compared to the rule-based chatbots.

- **Generative-based**: The generative-based chatbots rely on advanced NLP algorithms to decipher users' utterances, detect the users' intents, and respond without using human intervention. Since the generative-based chatbots use data patterns to generate more human-like responses and are not based on predefined rules, users experience more user-friendly chatbots, resulting in a pleasant user experience [5], [6], [154]. However, the training of the advanced NLP algorithms requires a large amount of training data which is not accessible for all applications.

The MIRA Chatbot is based on a combination of rule-based and retrieval-based approaches. The main reason for not employing generative-based models is that the chatbot must respond appropriately to acute emergencies such as suicidal ideation in the case of mental health applications. Should the chatbot respond inappropriately or harshly, the consequences to a client could be incredibly detrimental. Hence, we wrote a set of predefined responses and verified them with psychiatric experts and individuals with lived experience to avoid any kind of improper response from the Chatbot, and then programmed the Chatbot to employ the predefined responses.

The following are some methods that can be used alongside the mentioned chatbot design techniques (i.e., rule-based, retrieval-based, and generative-based) for building chatbots.

- **Parsing**: The parsing process uses text as an input and breaks it down into a set of simpler words that can be more readily stored and processed. Moreover, the parsing process establishes how a sentence is grammatically structured [148]. This approach is used within an early chatbot called ELIZA [149]. Semantic parsing is a more complex technique that transforms a user's input sentence into a machine-understandable representation of its meaning. This approach is employed in some of the recent commercial chatbots like Dialogflow [127] to detect a user's intent.

- **Pattern Matching**: The pattern matching method categorizes user input as $< pattern >$ and generates an appropriate answer from a $< template >$. The pattern matching technique is beneficial for starting conversations. However, it has several scalability problems since all the patterns must be constructed manually, which is not a trivial task. Because of the scale problem, chatbots' ability to extract information is constrained, and their responses may be predictable and repetitious [142]. Some chatbots like ALICE [43] employ more complex pattern matching rules, while others like ELIZA use simple pattern matching rules.

- **AIML**: Artificial Intelligence Markup Language(AIML) is a programming language for creating conversational flows of chatbots and it consists of two main units, topics, and categories. The topic is a top-level element with a name attribute and a collection of associated categories. Each category represents a rule for matching an input to output and must have at least two extra elements: pattern and template. The pattern compares user input to the template, which is used to construct the Chatbot's response [125]. AIML is a strong tool for defining the conversational flow of chatbots and is versatile and simple to use. However, it does require some natural language processing and programming skills.

- **Chatscript**: Chatscript is a free and open-source tool for creating chatbots, which uses a natural language engine and dialogue management system that allows users to have interactive discussions while retaining their user state [54].

- **Markov Chain Model**: The Markov Chain is a probabilistic and mathematical model that attempts to explain the probabilities of state transitions across time. The essential concept behind the Markov chain model is that given the present state, there is a fixed probability that it will proceed to one or more subsequent states. The Markov Chain enables the chatbot to build sentences for more probabilistically appropriate re-

sponses while being more or less intelligible. Markov Chains are a common way of creating chatbots that mimic simple human interaction for amusement. Because the Markov chain model is a reduced representation of a complex decision-making process, it is ineffective for simulating rich and complex conversations [109]. Programming a simple Markov chains model is straightforward, and the entire model may be summarised in a matrix.

- **Artificial Neural Networks Models**: Artificial neural networks have been used to solve various problems, including computer vision, decision-making, speech recognition, machine translation, social network filtering, and medical diagnosis. Artificial Neural Network-based chatbots can use both retrieval-based and generative-based approaches in response generation, although the research trend is moving into the generative-based approach [32], [94]. Different Artificial neural networks including Recurrent Neural Network (RNN), sequence to sequence, and Long Short Term Memory Networks (LSTM) can be utilized for NLP [33].
  - **RNN**: An RNN is a type of artificial neural network that feeds a layer's output into a new input layer to predict the next outcome. More specifically, a recurrent neural network can remember prior computations and apply this knowledge to future processing. The landscape of chatbots has evolved as a result of this simple technique [157]. Also, RNN can capture natural language's inherent sequential nature, in which words gain semantical meaning based on the previous words in the sentence. This permits the RNN to maintain context and create a result based on the sentence's previous words. Hence, RNN is useful for chatbots since for a chatbot comprehending the conversational environment to interpret the user's inputs and deliver contextually correct responses is necessary [22]. Moreover, Chatbots tend to produce more contextually correct responses if they are fed with information from previous conversations [22].
  - **Seq2Seq Neural Models**: The Sequence to Sequence (Seq2Seq)

model is based on the RNN architecture and consists of two recurrent neural networks, an encoder that processes the input, and a decoder that creates the output [22]. The Seq2Seq model is the most extensively utilized in industry best practice for response generation [29]. The model can be fed varying input sentences through the encoder and decoder. The encoder encrypts input text and decodes it to produce the desired output. The model is most commonly employed in language translation, where the input sentence is in one language and the output sentence is in another. This approach can also be used to convert between input and output in chatbots [29].

- **LSTM**: LSTMs are a type of RNN [109], designed to prevent the dependency concern of RNNs. Memory cells and gates are two main components of LSTMs, which allow the LSTMs to recall past information for extended periods. Memory cells are like a computer's memory that may store, write, and read information. The gates consist of input gates, forget gates, and output gates used to control the flow of information. Compared to a typical RNN, an LSTM network is better at learning from experience. Thus, RNNs have now been replaced as the standard by LSTM networks. Even when there is a long period of unknown size gaps between major events, a well-trained LSTM network may do superior categorization, processing, and prediction of time series. These features demonstrate LSTM's superior performance compared to other RNNs, hidden Markov models, and other sequence learning approaches utilized in various applications [115]. Because LSTMs can regularly refer to a piece of distant information in time, they are extremely valuable in creating chatbots [109], [115].
- **Transformers**: Transformer-based models in contrast to LSTMs do not focus on one token at a time, so transformer-based models do not suffer from the vanishing gradient problem [120]. Specifi-

cally, transformer-based models pay attention to words in a learned order and thus enable more parallelization while improving upon many NLP problems [138], [144]. Hence, transformer-based models, as opposed to sequential models such as the LSTM, are more analogous to human reading comprehension [136] and are forming state-of-the-art scores for many NLP tasks [136] such as Generation [106], question answering [77], [123], sentiment analysis [91], [122], translation [145], [158], and paraphrasing [26].

In order to develop the MIRA Chatbot, transformer-based models are applied strategically to achieve superior performance while using other approaches as a baseline. There is an overview of the most impactful chatbots in the following section.

2.3 History of Conversational Agents

Much work has been done in recent years to improve the classic chatbot system, from the basic scripted question-answering bots to self-learning bots [127]. This section contains a review and analysis of past research in this field conducted in earlier years.

- **ELIZA**: It was introduced by J. Weizenbaum [149]; the first bot that came close to competing with the Turing Test by effectively recording input, rephrasing it, and matching keywords with a predefined list of responses. Because ELIZA is a rule-based system, it cannot have meaningful conversations with humans, yet it manages to convey the impression that it is not a computer program [56].

- **PARRY**: It was introduced by K. Colvy (1972) and created to assist a person with paranoid schizophrenia. ELIZA and PARRY both use a rule-based approach, but ELIZA has an advantage over PARRY in terms of advanced control structure, language understanding, and a model that can imitate emotions such as anger and fear [30]. PARRY has enough

intelligence to respond with hostility if anger is high. PARRY demonstrated how technology could assist in replicating a person with mental health issues [30].

- **ALICE**: It was introduced by R. Wallace (1995). Artificial Linguistic Internet Computer Entity (ALICE) is a well-known AIML-based open-source bot. The Chatbot was inspired by ELIZA and won the Loebner prize in Jan. 2000. The supervised learning parts rely on the individual teaching and tracking of the chatbot's discussions, as well as suggesting additional AIML content to make the responses more relevant, rational, exact, and credible. Because ALICE is a preset set of questions and answers, it lacks the robustness to respond to all queries [43], [142].

- **Mitsuku**: It is introduced by S. Worswick (2013). The rule-based Chatbot is written in AIML, and the advanced bot won the Loebner Prize five times in 2013, 2016, 2017, 2018, and 2019 [15], [108]. Mitsuku's improvement includes holding long discussions, learning from chats, and recalling personal information about users such as age, location, and gender.

- **Watson**: It was developed by IBM (2006) and won the Jeopardy TV show in 2011. Watson is a retrieval-based chatbot based on the Hadoop-based machine learning system that uses advanced NLP technologies, including Information Retrieval, Knowledge Representation, and Automated Reasoning [38].

- **Siri**: It was released by Apple in 2011. With simple features like making a call, responding to messages, and managing alarms, Siri offers elements like genre, accent, and language that can be configured and customized. The Back-end of Siri uses Automatic Speech Recognition (to convert speech to text); Natural Language Processing (part of speech tagging, noun chunking, constitute parsing, and dependency parsing); and turning transcribed text into "parsed-text" [127].

Table 2.1: The comparison of well-known chatbots.

	Chatbot	Year	Developed by
1	ELIZA	1966	J. Weizenbaum
2	PARRY	1972	K. Colvy
3	ALICE	1995	R. Wallace
4	Watson	2006	IBM
5	Siri	2011	Apple
6	Mitsuku	2013	S. Worswick
7	Alexa	2014	Amazon
8	Cortana	2014	Microsoft
9	Google Assistant	2016	Google

- **Alexa**: It is developed by Amazon in 2014. Alexa has a natural voice and can speak with users. Receiving, recognizing, and naturally responding to voice commands are the foundations of the system. English, French, Spanish, German, and Japanese are among the languages supported. Alexa can perform various simple activities, such as providing weather, traffic, and other information, as well as reading books, podcasts, and audiobooks [127].

- **Cortana**: It is developed by Microsoft in 2014. With the help of Windows 10 and Windows Mobile, Cortana can perform the same duties as Siri and Alexa [127].

- **Google Assistant**: It is introduced by Google in 2016. The chatbot has a voice recognition feature and a user my talk and performs basic functions such as Question Answering (QA), providing information, and displaying the fastest trip to any location [127].

As illustrated in Table 2.1, significant progress has been made throughout the development of the aforementioned chatbots. This data indicates that people can be expected to feel comfortable using chatbots and that more technology companies will establish teams of chatbot developers. In the following chapter, existing chatbot frameworks for building a conversational agent are listed.

2.4 Existing Chatbot Frameworks

In the following, some publicly available chatbot frameworks are compared. The comparison is done according to three criteria: open-source, free, and self-host. Access to the source code of open-source frameworks makes any custom modifications possible. Free frameworks are available without any cost and self-hosted frameworks can be deployed on a personal server.

- **Dialogflow**: Dialogflow is a natural language understanding platform developed by Google that integrates twenty languages on platforms such as Facebook Messenger, Slack, and Skype for a monthly fee. Designing and implementing the speech systems of chatbots is made simple through the platform as it operationalizes the features of Google and Amazon Alexa. Dialogflow, a closed-source framework, uses Google Cloud Functions to connect with applications, which makes implementing the platform in a personal server impossible [127].

- **Microsoft Bot Framework**: The Microsoft bot framework is introduced by Microsoft which consists of two main components: bot builder and channel connectors. Bot builder is for creating business logic while channel connectors are for linking the chatbot to messaging systems like Slack. It is worth mentioning that the framework can connect to NLU services [127].

- **Amazon Lex**: Amazon Lex is introduced by Amazon and uses the same technology as Alexa, with the flexibility to scale up. Moreover, Amazon Lex builds chatbots for mobile applications using platforms such as Slack and Facebook Messenger. The platform is not free to use, and it is closed-source, and deploying it on a personal server is not allowed [127].

- **Pandorabots**: Pandorabots uses AIML for writing scripts, and the framework provides an online web service for generating and deploying the scripts. Pandorabots provides free open-source libraries (ALICE, BaseBot, and Rosie), as well as premium libraries and modules for a

monthly price citeSINGH2021. Also, the framework is closed-source and can be hosted on a personal server.

- **Chatterbot**: Chatterbot is an open-source framework with an active learning feature in which the performance of the chatbot improves over time with more users' data. It is simple to use and it has automatic responses [127]. However, it requires a monthly fee and cannot be hosted on a personal server.

- **RASA**: Rasa consists of two main components: RASA NLU and RASA core. RASA NLU is mostly for interpreting natural language, while Rasa core is for building a chatbot, allowing more sensible dialogue, and training using interactive and supervised machine learning. Rasa has the advantage of being able to host on a personal server. Rasa core and NLU are open-source and free, but there is a Rasa Enterprise platform that is paid and a more advanced version [127].

Table 2.2 summarises the frameworks. To develop the MIRA Chatbot, we decided to use the RASA framework and we deployed it on the University of Alberta servers. We did this for several reasons. We needed a framework that could be deployed on our server to ensure a high level of data security and protection. We also needed an open-source framework that could give us the ability to perform any kind of modifications required (including the addition of new features), as well as permit widespread scalability and translation. Additionally, in consideration that the MIRA Chatbot is being provided to healthcare workers free of cost, affordability was taken into account and free platforms were given preference.

2.5 Evaluation Metrics

Evaluation of conversational agents is a challenging task given the lack of a predetermined response to a question that a user asks. For example, consider the question "Who is the best soccer player?". "It is Lionel Messi" and "without any doubt, It is Lionel Messi" are two examples of acceptable responses.

Table 2.2: The comparison of chatbots frameworks.

Framework	Open source	Free	Self-host
Dialogflow	✗	✗	✗
Microsoft bot framework	✗	✗	✗
Amazon Lex	✗	✗	✗
RASA	✓	✓	✓
Pandorabots	✗	✗	✓
Chatterbot	✓	✗	✗

One can say if a response contains Lionel "Messi", it is correct. However, consider "although Lionel Messi is playing better now, Kylian Mbappe is the best player." although "Lionel Messi" is mentioned in the response, it is incorrect. Thus, advanced approaches need to be employed to evaluate chatbots. Some of the different evaluation metrics for assessing the chatbots are listed below:

- **Precision**: Precision is the quality of a positive production made by a machine learning model. To calculate precision, the number of true positives is divided by the total number of positive predictions (positive predictions include both true and false positive predictions). The calculation of precision is required, for example, to assess how many clients can be predicted to have depression. To find this, the model predicts depression in many clients, and precision is found by dividing the total number of true positive predictions by the total number of clients the model predicted to have depression [83]. Mathematically, precision is defined as follows:

$$Precision = \frac{TP}{TP + FP}$$

  Where TP is the number of samples that the machine learning model predicts as positive correctly and FP is the number of samples that the model predicts as positive incorrectly.

- **Recall**: The recall is the division of true positive samples by several samples that should have been classified as positive. For example, in

the cancer example, recall is defined as the number of samples truly predicted as positive divided by the number of patients who have cancer [83]. Mathematically, recall is defined as follows:

$$Recall = \frac{TP}{TP + FN}$$

Where FN is the number of samples that the model predicts as negative incorrectly.

- **F1-score**: The F1-score indicates the classifier's correlation between recall and precision when evaluating the model's performance. Higher F1-score is usually preferable, which can be achieved through increasing recall and precision [99], [153]. F1-score is defined as follows:

$$F1\text{-}score = 2.\frac{precision.recall}{precision + recall}$$

- **Perplexity**: Perplexity is a generative modeling metric used to evaluate a generative model based on how well the agent predicts what the person would say next or if a human or a program generated a phrase by selecting random words. In layman's terms, perplexity is a metric that measures how well a probability distribution or model predicts a sample. The optimal confusion score of 1 implies very human-like behavior. The better the performance, the lower the number; in other words, the model's confusion is represented by higher perplexity [6], [7].

- **MRR**: MRR (Mean Reciprocal Rank) is a statistic for evaluating any strategy that generates a list of likely answers to a set of questions, organized by likelihood of certainty [4].

- **BLEU**: BLEU (Bilingual Evaluation Understudy Score) is a score used to assess how successfully a translation from one language to another was completed. The perfect BLEU score is 1, while the mismatch score is 0. BLEU is independent of language and correlates with human evaluation [95].

- **Human Evaluations**: Some metrics can not be evaluated without the intervention of a human. Empathy, content evaluation, user happiness, functional aspect, relevance, fluency, and ambiguity are just a few evaluations that require human intervention.

2.6 Datasets for Training a Chatbot

Many datasets in numerous languages and domains are available, including movie reviews, chats, QA, subtitles, etc. In this section, we introduce some of the available datasets for training a chatbot.

OpenSubtitles[1]: This dataset is taken from movies and television subtitles (available in English and other languages) to test open-domain chatbots' robustness to varied personalities, and linguistic variances [72]. The datasets consist of 283,651,561 subtitles for training data and 33,240,156 subtitles for test data [27].

Reddit [1]: Reddit comments were gathered from 2015 to 2018, and the dataset is most commonly useful for short discussions and to give chatbots distinct personality responses. The dataset consist of 654,396,778 and 72,616,937 comments for training and test data, respectively [27].

Amazon QA[1]: The Amazon QA dataset is based on the corpus created by Amazon [81], [143]. The dataset can be used to answer costumer questions about items and provide assistance in customer service. The dataset consists of 3,316,905 QAs for training data and 373,007 QAs for test data [27].

Taskmaster-1: The Taskmaster-1 dataset contains 13,215 talks in English with 301,876 utterances to aid researchers in developing chatbots with booking capabilities. There are also 5,507 spoken and 7,708 written dialogues (created

[1]These fine-datasets are then filtered and processed version of the raw datasets provided by Chen et al [27]. These datasets are publicly available on https://github.com/PolyAI-LDN/conversational-datasets.

by workers playing both sides of the conversation). For ordering and making appointments, each interaction can be linked to one of six domains: pizza delivery, auto repair appointments, ride services, movie tickets, coffee delivery, and restaurant reservations [20].

Coached Conversational Preference Elicitation (CCPE): The dataset provides 502 English interactions with 12k annotated statements between a user and an agent, where the workers discussed film preferences. The assistant asks questions to reduce the user's bias and allows them to express their views spontaneously. Every entity has a dialogue, which includes preferences expressed regarding the entities, descriptions, and other utterances. This dataset can be used to put open-domain systems to the test in terms of context jumping and remembering past contexts [107].

Ubuntu Dialogue Corpus (UDC): The dialogues and words statistics in Table 2.3 are unique to the dataset presented by [75], making it the only database for research into building dialogue systems based on Term Frequency-Inverse Document Frequency (TF-IDF) [60], RNN, LSTM. Finally, UDC can assess personality differences and short conversation issues.

The Gutenberg Dialogue Dataset: Csaky and Recski [34] used open-domain dialogue datasets to bridge the gap between quality (as Daily Dialog) and size (as OpenSubtitles). As shown in Table 2.4, multiple high-quality datasets of many utterances in various languages are created. This is an essential dataset for evaluating the robustness of language variants and open-domain agents as social chatbots.

A thorough analysis of previous research did not provide any specific datasets for mental health chatbots, which motivated us to extend the project to include the establishment of "MIRA Dataset." Data augmentation techniques were employed to expand the examples of this unique dataset, which will be described thoughtfully in the following chapter.

Table 2.3: Properties of UDC dataset [75].

Dialogues (human-human)	930,000
Utterances	7,100,000
Words	100,000,000
Min. number of turns per dialogue	3
Avg. number of turns per dialogue	7.71
Avg. number of words per dialogue	10.34
Median conversation length (min)	6

Table 2.4: Properties of Gutenberg dataset [34].

Language	**Utterances**	**Avg. Utterance Length**	**Dialogues**	**Avg. Dialogue Length**
English	14,773,741	22.17	2,526,877	5.85
German	226,015	24.44	43,440	5.20
Dutch	129,471	24.26	23,541	5.50
Spanish	58,174	18.62	6,912	8.42
Italian	41,388	19.47	6,664	6.21
Hungarian	18,816	14.68	2,826	6.66
Portuguese	16,228	21.40	2,233	7.27

2.7 Related Work

In this section, a brief review of past research in the field of conversational agents is provided.

Bagchi [8] developed a novel library agent utilizing open-source RASA framework. The architecture of the library agent consists of several parts: connector modules, dialogue management, and output module. A user's query is processed by connector modules, and then input modules are used for intent detection and entity extraction. Afterward, the backend/database/API request is handled by dialogue management, which retrieves information from resources. The output module receives a response followed by a response from the connector module. In the MIRA Chatbot, similar architecture is utilized. The work can be improved by including evaluation metrics and descriptive data analysis on the dataset.

Hwerbi [56] suggested a COVID-19 assistant that would deliver up-to-date information within a day using a question–answering system that would provide accurate information. The systems are divided into seven parts: ontology (Corona Virus Infection Ontology), web scraping module (using selenium and beautiful soup from Wikipedia and Google for COVID-19), database (XML format), finite state machine (to show the entire architecture for all states and transitions), keyword extractor (keyword detection to extract a set of terms or phrases from an unstructured text), trained model (Chatterbot trained on English corpus), and the user interface. Future development could include a universal chatbot or open-domain bot that can support several languages, speech recognition techniques, more powerful and efficient NLP operations, and more informative use of enriched ontologies such as trends and advice.

Gillian Cameron [23] offered a mental health chatbot named iHelpr Chatbot to offer guided self-assessment on stress, anxiety, depression, sleep, and self-esteem. iHelpr allows the user to complete a self-assessment instrument

based on the option they have selected at first. Based on the self-assessment survey results, the user is subsequently given tailored counsel with evidence-based recommendations. Links to supplementary support books accessible on the Inspire Support Hub website and recommended e-learning programs are among the recommendations. The user is given helpline numbers and emergency contact information if there is a heightened risk. The Microsoft Bot Framework was used to create the iHelpr Chatbot. Language Understanding Intelligent Service from Microsoft was included to recognize users' intent and relate them to the appropriate resources.

Denecke [35] offered Social Network Platform for Physicians (SERMO), a smartphone software with an integrated chatbot that uses Cognitive Behavior Therapy (CBT) approaches to help mentally ill people regulate their emotions and deal with their thoughts and feelings. SERMO polls the user regularly about recent events and emotions. It uses NLP to automatically determine a user's primary emotion from natural language input. SERMO suggests relevant measurements based on detected emotions, such as activities or mindfulness exercises. Among the additional features are an emotion diary, a list of pleasurable activities, mindfulness exercises, and information on emotions and CBT.

Using a Convolutional Neural Network, a novel approach for automatic depression detection in speech is offered by Ellie chatbot [28]. A total of 2568 speech samples were collected from 77 non-depressed and 30 depressed people to test the model. The experimental findings revealed a baseline accuracy of 77%.

Although there is a considerable number of mental health chatbots, most are not free to use, require a monthly fee, are not customized for a specific region, or are for a specific domain like depression. For example, iHelpr, SERMO, and Ellie require a monthly fee, and their recommended resources are not customized for a specific region. Other types of mental health chatbots are not

publicly available. The newly developed MIRA Chatbot fills this gap as a mental health chatbot that is free to use and customized only for healthcare workers and their families.

Chapter 3

Building the MIRA Chatbot

In this chapter, the MIRA design and implementation are explained in depth. Following this, the word representation techniques are described. Afterward, the basic machine learning algorithms and a transformer-based machine learning model are introduced. Moreover, the MIRA Dataset is introduced by mentioning some data augmentation techniques used to expand the training data.

3.1 MIRA Design and Implementation

In this section, the architecture of the MIRA Chatbot is described. The architecture consists of three parts: MIRA Resource Portal, MIRA Backend, and MIRA Interface. All of these elements are demonstrated in Fig. 3.1 and are described in further detail.

3.1.1 MIRA Resource Portal

The MIRA chatbot is not actively searching the open Internet for resources to share with clients. Instead, to ensure the chatbot is offering evidence-based and verified resources, we have developed a unique portal where the resources can be stored and accessed by the chatbot. The definition of "resource" for this project is broad and may include websites, graphics, phone numbers, and more. We capture up to 25 data points for each resource. Our chatbot uses this information to find the best resources for client needs. Also, our intelligent portal user interface has features to reduce the burden on volunteers inputting

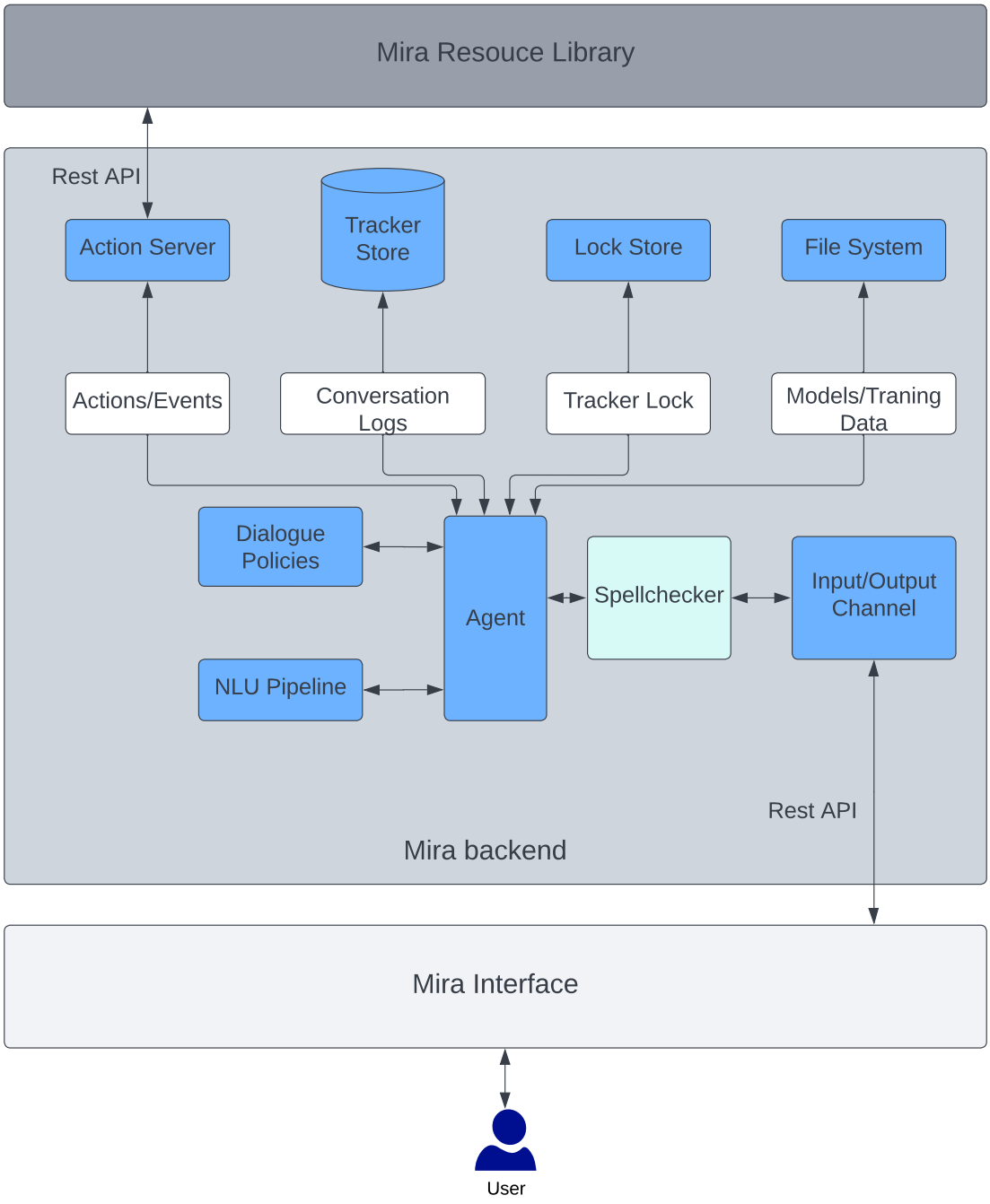

Figure 3.1: MIRA uses the Rasa framework for its backend, a resource portal for cataloging resources, and a MIRA Interface that allows users to chat with the MIRA chatbot.

resources, including search functions, survey logic, and information pop-ups. Our expert advisory committee members review all resources submitted to the portal. Committee members will also be helping us beta test the chatbot. They will also share information about our chatbot with their networks. Verification of resources is done via a review process reflecting that of peer-review academic journals. That is, there will be two reviewers per resource and an editor.

The portal resource quality assessment matrix was developed using items drawn from several existing validated tools for the quality assessment of online resources. The matrix includes the opportunity for reviewers to check the resource for errors and rate the resource based on variables, including: level of accessibility, evidence to support the resource, and risk of bias. The matrix tabulates the responses from the reviewer and automatically provides a recommendation based on the matrix for the reviewer to consider for their final decision.

3.1.2 MIRA Backend

MIRA Backend is responsible for receiving users' messages from different input channels such as Slack and Telegram. Other duties assigned to the MIRA Backend include intent detection; entity extraction; storing conversation logs; storing the trained model for intent detection and entity extraction; action selection; and spell-checking. These critical duties make MIRA Backend a core element of MIRA's architecture. A user's message comes from an input channel, and after going through the spell-checker, it will pass to an agent, which is responsible for connecting different parts of the Rasa Open Source. The Rasa Open Source architecture is depicted in the MIRA Backend part of Fig. 3.1 with blue components. Natural Language Understanding and dialogue management are the two main parts. Entity extraction, intent detection, and answer retrieval are all handled by the NLU component, which is based on the DIET model. Based on the context, the dialogue management component (dialogue policies) chooses the subsequent course of action in a discussion. In the MIRA Chatbot, all the actions were predetermined as shown in the MIRA

tree (Fig. 3.2) (more detail in A.3) in which a user should follow one of the paths. It is worth mentioning that, according to the detected intent by the NLU pipeline, having a jump in the tree is possible. For example, suppose the chatbot asks a user to mention their name but the user types, "what are the symptoms of anxiety?". In this case, the chatbot will jump to another path to reply to the user.

The conversation logs will be stored within a tracker store as shown in Fig. 3.1. Different implementations of the store types were provided by Rasa open source; for the MIRA project, we have used PostgreSQL for the tracker store.

After training the chatbot, the trained model can be stored in different places such as a local disk, HTTP server, or cloud storage. For the MIRA project, we have saved the trained model on the local disk.

Rasa locks conversations while messages are being processed and employs a ticket lock system to ensure incoming messages for a specific conversation ID are handled in the correct sequence. As a result, many Rasa servers can run simultaneously as replicated services, and clients are no longer required to transmit messages to the same node when using a specific conversation ID.

3.1.3 MIRA Interface

Once the client clicks on the MIRA URL[1], they are directed to the MIRA chatbot interface. MIRA begins by welcoming the client and asking them for their consent to use anonymized data from the conversation to evaluate and improve its services, with a link to a pop-up with consent information. If the client provides consent, the chatbot then asks a short series of demographic-related questions—employment type, location, and end-user (e.g., for the client or someone else; if someone else, then the end user's age is also collected). Following these preliminary questions, the chatbot asks in an open-ended manner (e.g., "How can I help you?") and provides some examples of questions that could be asked in the form of button options (e.g., "I want to find programs and services" or "I want to learn coping skills"). If a client asks to "chat" with

[1] mymia.ca/chatbot

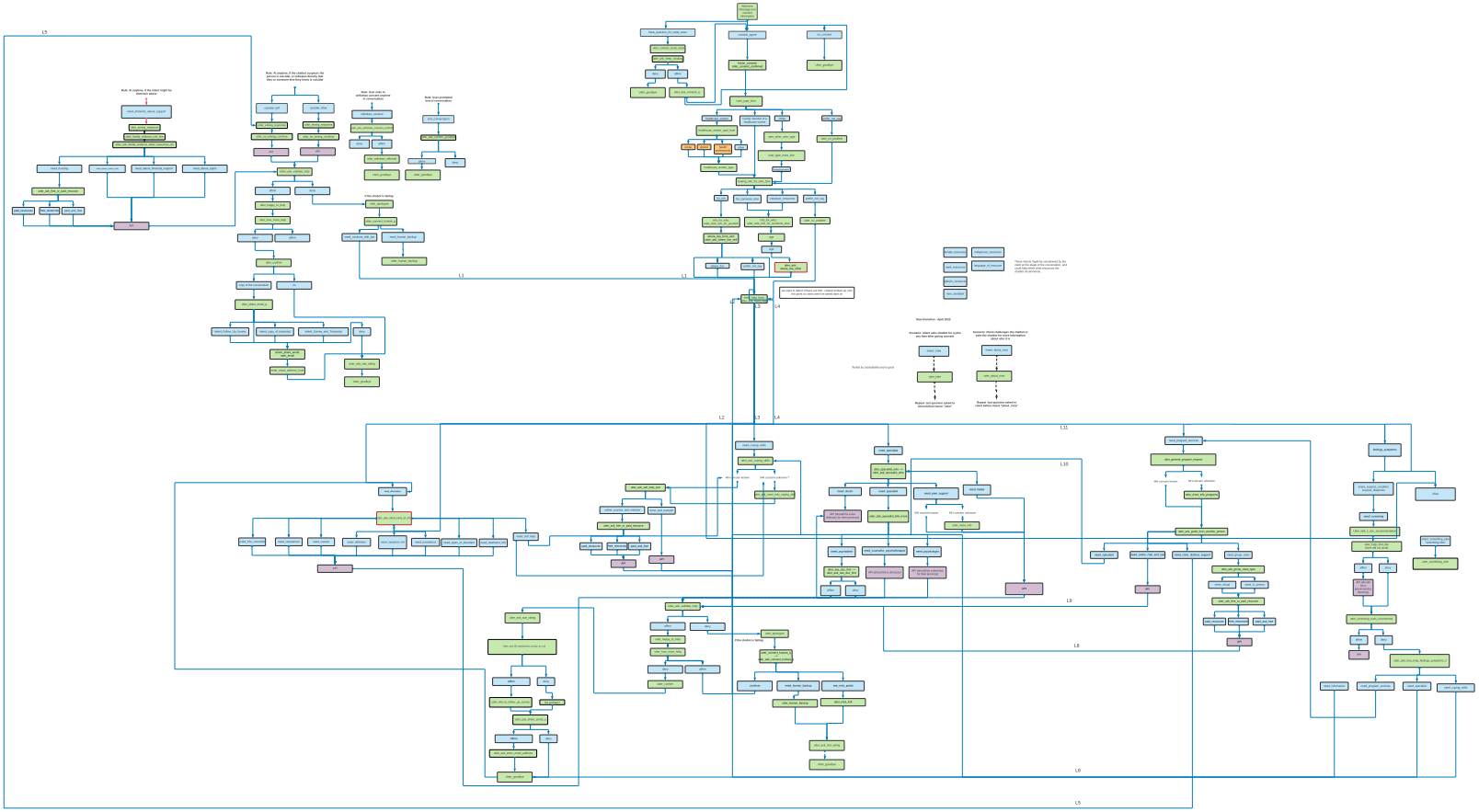

Figure 3.2: MIRA decision tree.

MIRA or begins expressing feelings as opposed to an apparent request for information, eEliza will be prompted, in which we have developed an enhanced version of Eliza to have an open conversation. The MIRA chatbot and the MIRA Resource Portal were tested for compatibility with multiple electronic devices, including tablets, smartphones, and desktop computers, as well as multiple browsers Safari and Chrome.

The MIRA Interface provides a convenient interface for users to chat with MIRA (Fig. 3.3). At the bottom of the interface, a user has two options for chatting: type or speak through the microphone by clicking on the microphone icon. At the top of the interface, the MIRA logo and the list of icons described in the following are placed. If the user clicks on the "Exit" button, the chatbot window will immediately close and redirect the client to the Weather Network to protect individuals who may be experiencing family violence. By clicking on the reset icon, the whole conversation will restart. The next button (bug icon) is for reporting bugs and issues. The client can see the consent information by clicking on the shield icon. Moreover, the play icon makes MIRA's conversation audible to the clients. The user's messages will be sent to the MIRA backend through REST API [42] and a response will return from the MIRA backend.

3.2 Word Representation

Natural language processing and several other machine learning tasks have attempted to understand how to create the most accurate representations of a word [11]. Text words cannot be processed or comprehended by the computer system directly, unlike pictures and audio, which analog or digital impulses may represent [58]. In this section, some methods for mathematically representing text words are introduced.

One straightforward but typical approach is to encode each text word as a single-hot vector [69]. To be more precise, given a text corpus, a vocabulary that includes all of the terms in the corpus or just the most crucial ones is developed. A binary vector whose dimension is equal to this vocabulary size may

Figure 3.3: MIRA Interface.

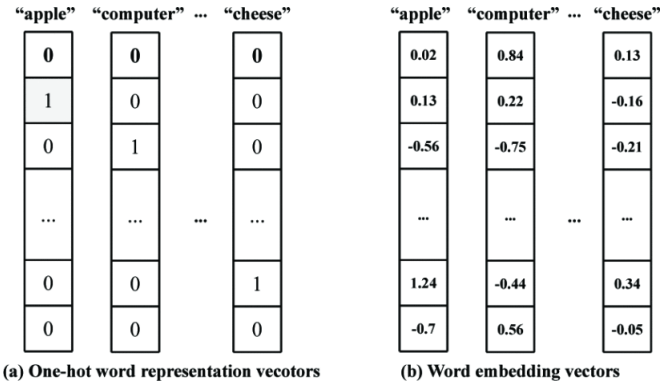

Figure 3.4: Word representation; Word embedding values are continuous, whereas one-hot vector values are binary [121].

then be used to represent each word. The word indices, as seen in Fig. 3.4.a, distinguish these vectors from one another. This approach, along with others like TF-IDF and bag of words [105], has been widely utilized in several NLP applications like document categorization [121] because of its straightforward development. We refer to this approach as "Count Vector" and we will use it for the basic machine learning models.

The one-hot vector representation, however, has obvious disadvantages. Above all, this approach frequently experiences the curse of dimensionality [11]. For instance, millions of text words may be involved in the language translation task [9]. As a result, the associated one-hot vector will grow very large and sparse, making language modeling ineffective. Furthermore, these discrete vectors cannot capture semantic connections between words.

Hence, using some techniques for learning low-dimension and continuous vector representations, also known as word embedding, is essential. The idea of word embedding can be credited to the work of Bengio et al. [11] and it is illustrated in Fig. 3.4.b. A real-value vector represents each word in this diagram, and words with similar meanings are located closer together in the embedding space. The distributed hypothesis, which asserts that words occurring in similar situations tend to have similar meanings, is the foundation for word embedding [74], [47]. In the following, we briefly describe the embedding methods used in this project.

ConveRT: A new sentence embedding model called ConveRT has been open-sourced by Henderson et al. [48]. The model, which has six layers of transformer modules, was developed using talks from Reddit to train on a response selection task. The power of sentence-level representations from ConveRT is that one can get top performance with a model that trains in a couple of minutes on a CPU. Note that pre-trained ConveRT is only available in English.

BERT: Another pre-training-oriented language model is BERT which uses a 12-layer Transformer network as its language model. BERT takes into account the bi-directional language modeling of context terms. Additionally, BERT incorporates new technical innovations like position embedding, which combines static word embeddings as model input. The outputs of the Transformer network can be utilized as contextualized word representations.

3.3 Machine Learning Models

Four algorithms were used to detect the intent of a sentence and to extract entities from it. Naive Bayes, logistic regression, whereas the support vector machine was the basic machine learning models employed, and Dual Intent and Entity Transformer (DIET) architecture was the transformer-based model used. The model performance was assessed through precision, recall, and the F1-score.

3.3.1 Basic Machine Learning Algorithms

Models that were utilized, including Naive Bayes, logistic regression, and support vector machine models, are further described in this subsection. The data were converted to the numeric version using the Count Vector featurizer.

Naive Bayes

Because Naive Bayes is quick and straightforward to use, Naive Bayes is frequently employed as the baseline in text classification. Although its strict

presumptions enable such efficiency, they also have a negative impact on the results of its output [112]. Under the scikit-learn[2] library, three different types of Naive Bayes models exist.

- Gaussian: It is employed in classification tasks and relies on the assumption that feature data is distributed normally.
- MultinomialNB: It implements the naive Bayes algorithm for multinomially distributed data and is used in text classification.

We have used Gaussian Naive Bayes for the MIRA project as a baseline.

Logistic Regression

A popular approach for binary classification is logistic regression. Given one or more independent variables as input, a generalized linear model is used to predict the likelihood of a dependent variable [51]. It calculates the likelihood that a particular instance belongs to a specific class. The model predicts that the instance belongs to the positive class if the expected probability is larger than 0.5. The instance belongs to the negative class if the model predicts a value less than 0.5. In logistic regression, the goal is to transfer the result of a linear equation between 0 and 1 to one of two labels. Since we are facing a multi-class dataset in the MIRA project, we have used a one-vs-all approach to be able to use logistic regression. One-vs-all is a heuristic technique for applying binary classification algorithms for multi-class classification in which the multi-class dataset is divided into various binary classification problems [113].

Support Vector Machine (SVM)

Support vector machine is one of the traditional machine learning methods that can still be used to help with big data categorization problems. In a big data setting, it can be beneficial for multidomain applications [93]. It manipulates the straightforward mathematical concept to enable linear domain

[2]https://scikit-learn.org/stable/

division. The model is known as a linear support vector machine if the classes in the original domain can be separated linearly (for example, along a straight line or hyperplane). Nonlinear support vector machines are used when the data domain cannot be divided linearly but can be transformed into a space known as the feature space. The data domain may be divided linearly in this space to separate the classes. Different kernel functions, including linear and Radial Basis Function (RBF) kernels, can be used for the transformation [133]. We have employed linear and nonlinear SVM using linear and RBF kernels for this study. It is worth mentioning that one-vs-all is used.

3.3.2 DIET Architecture

Separating the intent detection and entity extraction tasks yields error propagation [44], so using a single architecture, achieves a better result than two architectures one for the intent detection task and one for the entity extraction task. Hence, we used the DIET architecture [19] which has been open-sourced inside the Rasa framework for the MIRA chatbot. Consider a chatbot that asks a user, "What would you want to know?" and the user replies, "Definition of Depression," which determines the intent of the user (need_definition) and the entity "Depression." An intelligent chatbot should be able to detect the intent and extract the entity. In the following, two main questions for understanding the DIET architecture in the training phase are answered.

- How does DIET work?
- Why does DIET work?

To answer the first question, we are going to explain several key parts of the DIET classifier as shown in Fig. 3.5.

Featurization: Consider "Definition of Depression" as an example in our training data, so we know the intent is "need_definition" and the entity is "Depression," feed to the model as shown in Fig. 3.5. There are two separate paths (sparse feature and pre-trained) for converting the tokens to numeric vectors. The pre-trained path accepts a pre-trained neural network like CoverRT

[49], or BERT [37], while the sparse features are a concatenation of token level one-hot encodings and multi-hot encoding of character n-grams ($n \leqslant 5$) and then the output is connected to a feed-forward network. Character n-grams have a lot of redundant information. Therefore, we apply dropout techniques to these sparse features to prevent overfitting. The same approach will apply to all the tokens to convert them to a floating-point vector. There are two feed-forward networks for converting a token to a floating vector which makes the model's training timely, so we apply 80% dropout to the feed-forward networks to make them lightweight, which is beneficial for the chatbot setting. According to work by [37], the DIET classifier adds a special classification token __CLS__ to the end of each sentence. The main idea of the __CLS__ token is to summarize the entire input. We establish the initial embedding for the __CLS__ token when using ConveRT because ConveRT is also trained as a sentence encoder. In addition to information from individual word embeddings, this adds additional contextual information for the entire sentence. We set the corresponding output embedding of the BERT [CLS] token for out-of-the-box pre-trained BERT.

Transformer: The model employs a two-layer transformer [138] with relative position attention to encode context throughout the entire sentence [124]. The input to the transformer architecture must have the same dimension as the layers of the transformer. To match the dimension of the transformer layers, which is 256, the concatenated features are then sent through another fully connected layer with shared weights across all sequence stages.

Named Entity Recognition: Conditional Random Field (CRF) [64] predicts a sequence of entity labels y_{entity}. The loss of the CRF calculates through the following equation in which α corresponds to the input sequence of tokens.

$$L_E = L_{CRF}(\alpha, y_{entity})$$

where $LCRF(.)$ indicates negative log-likelihood for a CRF [65].

Intent Classification: The transformer converts CLS token to a_{CLS}. The a_{CLS} and intent labels y_{intent} are embedded into a single semantic vector space $h_{CLS} = E(a_{CLS})$, $h_{intent} = E(y_{intent})$, where $h \in \mathbb{R}^{20}$. The goal is maximizing the similarity $S_I^+ = h_{CLS}^T h_{intent}^+$ for target label y_{intent}^+ while minimizing the similarities $S_I^- = h_{CLS}^T h_{intent}^-$ for non-target label y_{intent}^-. The intent loss calculates with the following equation where the sum is taken over the set of negative samples Ω_I^- and the average $< . >$ is taken over all the intent labels [50], [139], [152].

$$L_I = - < S_1^+ - \log(e^{S_I^+} + \sum_{\Omega_I^-} e^{S_I^-}) >$$

Masking: The DIET model aims not only to classify intents and entities quite well but also to understand general language. So some regularization techniques should apply. Hence, the notion of masking is added during training [156]. According to work by [37], [135], the architecture adds training objective to predict randomly masked input tokens. The DIET architecture opts for 15% of the input tokens randomly in a sequence, and in 70% of cases, the tokens replace with the vector corresponding to the special mask token $\_\_MASK\_\_$, in 10% cases, the tokens replace with a random token, and in remaining 20%, the original token will keep. The output of the $\_\_MASK\_\_$ token from the transformer, a_{MASK}, for each selected token y_{token} will use in the following equation for calculating the mask loss [50], [139], [152].

$$L_M = - < S_M^+ - \log(e^{S_M^+} + \sum_{\Omega_M^-} e^{S_M^-}) >$$

where the similarity with the target label y_{token}^+ is $S_M^+ = h_{MASK}^T h_{token}^+$ while $S_M^- = h_{MASK}^T h_{token}^-$ are the similarities with the negative sample y_{token}^-. The corresponding embedding vectors $h \in \mathbb{R}^{20}$ are $h_{MASK} = E(a_{MASK})$, $h_{token} = E(y_{token})$. The Ω is a sum over all the incorrect tokens, and the average $< . >$ is over all the tokens.

Total Loss: The total loss calculates with the following equation. The architecture trains to minimize the L_{total}.

$$L_{total} = L_I + L_M + L_E$$

Batching: In order to reduce class imbalance [57], the architecture employs a balanced batching method [139]. This is because some intents may be more common than others. The batch size will increase throughout the training as another form of regularization [128]. After training the DIET architecture, one can use it for predicting intent and entities of unseen examples. Consider "Know about Anxiety" as an example. By passing all the intents and entities to the DIET architecture (Figure 3.6) and choosing the ones that have the highest similarity, the intent and entities will predict.

3.3.3 Regular Expression

In order to make the entity extraction more effective, regular expressions are used alongside entity extraction from DIET architecture. A list of lookup tables has been defined within the MIRA Chatbot consisting of entities like cities and countries. The user's message is then reviewed via a regular expression, and if one of the words can be found in the lookup table, it is marked as an entity.

3.4 MIRA Dataset

As explained in the previous chapter, no specific dataset is available for training a mental health chatbot. Therefore, we aimed to create a dataset. The dataset consists of 91 intents with 2,292 examples when writing this manuscript. The dataset may expand more in the future. We applied the steps shown in Fig. 3.7 for creating the MIRA Dataset.

1. **Intent Definition**: The intents were defined according to the use cases of the MIRA chatbot. For example, one of the responsibilities of the chatbot is handling suicidal cases, so we created intent for it named "suicidal" intent. The dataset consists of 91 intents (A.1).

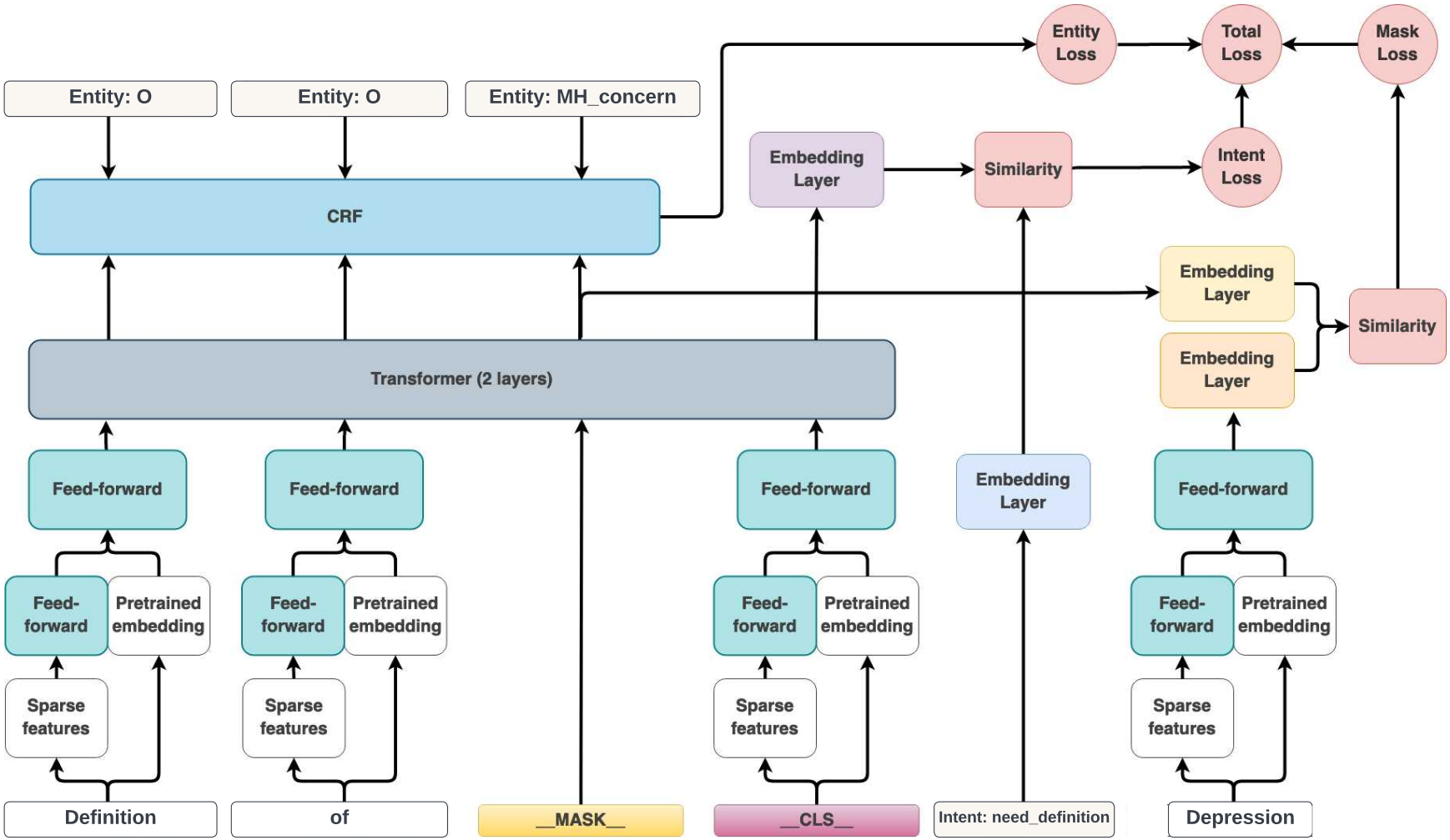

Figure 3.5: (Training phase) A representation of the DIET architecture in the training phase. The intent of "Definition of Depression" is "need_definition," and "Depression" is an entity named "MH_concern." The feed-forward layers' weights are distributed among tokens [18].

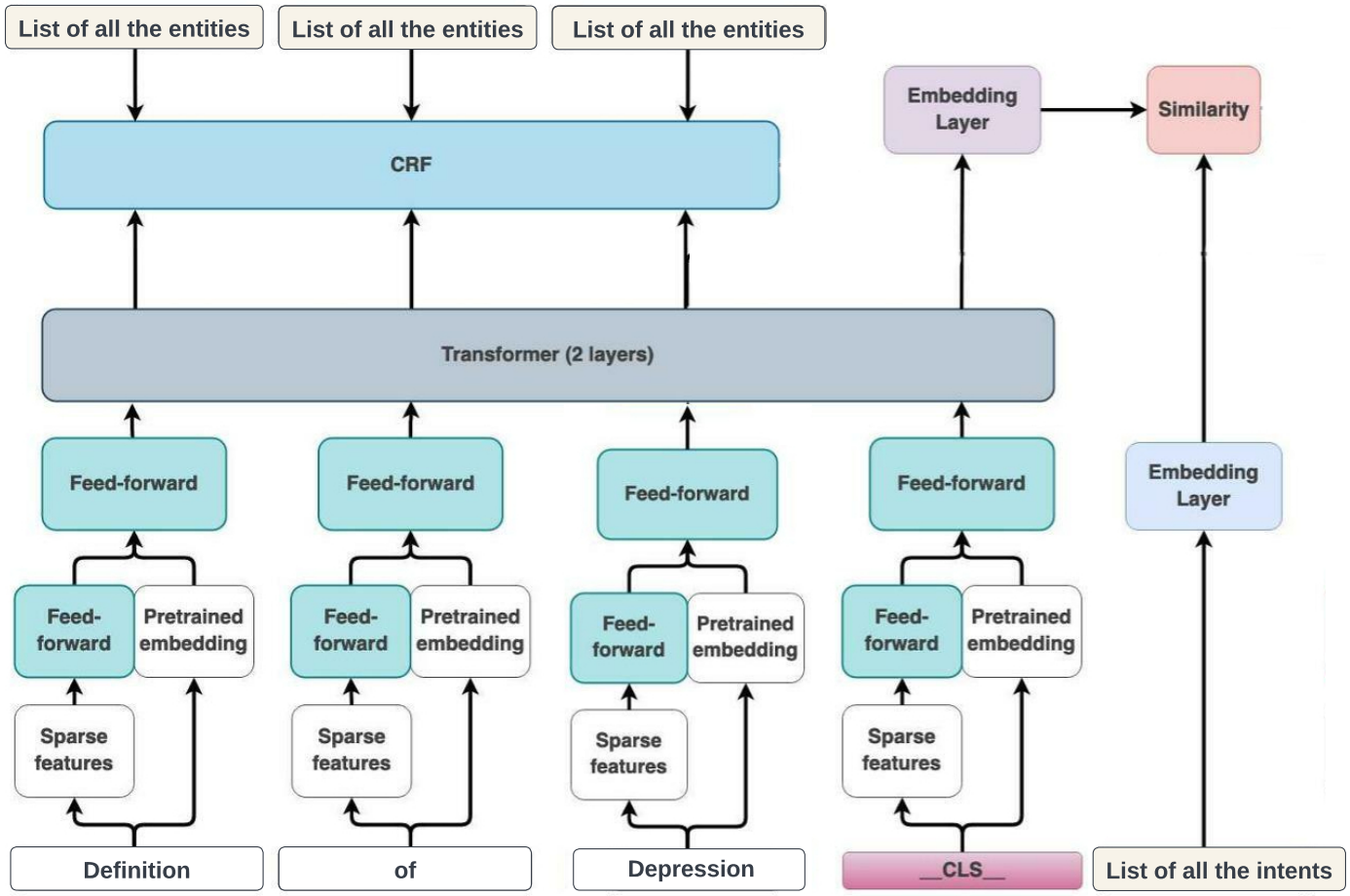

Figure 3.6: (Inference/test phase) A representation of the DIET architecture in the test phase. The intent of "Know about Anxiety" is "need_definition," and "Anxiety" is an entity named "MH_concern."

2. **Example Association**: After defining all the intents, we tried to define 3 to 5 examples for each intent in which 357 examples were defined. For example for the "suicidal" intent the following examples are defined:

    - I want to die.
    - I want to kill myself.
    - Life is not worth living anymore.

3. **Entity Anotation**: In this step, important keywords that should detect with the chatbot were annotated. For example, if a user types "I want to know about the symptoms of anxiety," the "anxiety" should mark as an entity.

4. **Data Augmentation**: After applying the intent definition, example association, and entity annotation, a small dataset was generated. To increase the accuracy of intent detection and entity extraction, we expanded the dataset by employing some data augmentation techniques. As shown in Fig. 3.8, the dataset is divided into the train set and test set, and the data augmentation techniques are only applied to the train set. It is worth mentioning that the utilized data augmentation techniques are categorized into four groups (for checking the complete list of used data augmentation techniques see A.2).

    - **EDA**: The EDA (Easy Data Augmentation) techniques are proposed in [147] for boosting performance on text classifications tasks. Synonym replacement, random insertion, random swap, and random deletion are four of the EDA's simple but powerful operations.
    - **Synonym Replacement**: This common type of data augmentation is the transformation of text into paraphrases by swapping out specific terms with synonyms. The work in [62] introduces one of the earliest uses of this replacement in the context of data augmentation. They used probable synonyms from WordNet to replace words [84]. According to the authors, the meaning of a phrase is

mainly preserved if one original token is replaced. Based on work by [62], we randomly select some words in a sentence and replace them with their synonyms using WordNet.

- **Embedding Replacement**: Comparable to synonym substitution, embedding replacement techniques look for words that best match the text's context while also maintaining the text's core ideas. To do this, words from the examples are translated into a latent representation space, where words from related contexts are placed closer together [46].
- **Replacement by Language Models**: By anticipating subsequent or missing words based on the prior or surrounding context, language models represent language (classical and respectively masked language modeling) [10]. Language models provide for a more localized replacement as opposed to embedding replacements by word embeddings that take into consideration a global context [80].

We utilized the approaches mentioned above for augmenting the MIRA Dataset.

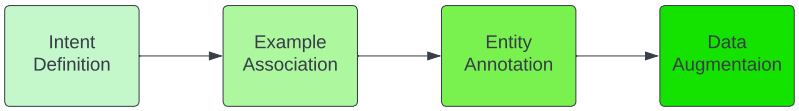

Figure 3.7: Steps of creating the MIRA Dataset.

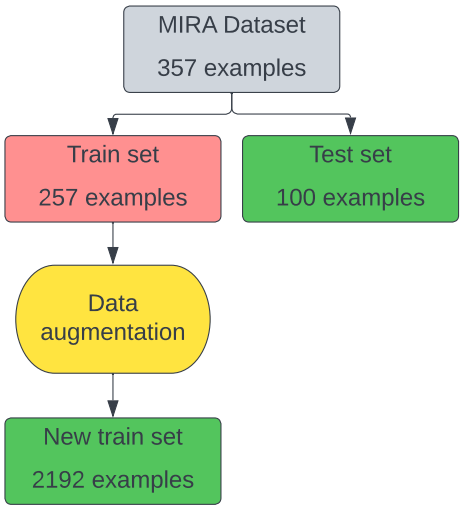

Figure 3.8: Dividing MIRA Dataset to the training set and test set and applying data augmentation techniques to the training set

Chapter 4

Evaluation of the MIRA Chatbot

Two main tasks are defined within MIRA Chatbot, intent detection, and entity extraction. Hence, improving the performance of intent detection and entity extraction has a direct impact on the performance of the chatbot and users' satisfaction. This chapter presents the results of intent detection and entity extraction using baseline machine learning models and (DIET architecture) with data augmentation and without data augmentation. It also examines the performance of the MIRA chatbot. Due to class imbalance in MIRA Database, reporting accuracy cannot evaluate the performance of intent detection and entity extraction. Therefore, a macro-averaged F1-score is utilized for comparing all baseline models and DIET architecture.

4.1 Performance of the Proposed Models

The results of the intent detection and entity extraction tasks on the MIRA Dataset (with and without data augmentation) are reported in this section. The first two subsections include the results of the basic machine learning models (i.e., Naive Bayes, logistic regression, and support vector machine), feed-forward neural Network, RNN, and LSTM. The feed-forward neural network is a two-layer fully connected neural network with 20, and 30 neurons in the first and second layers, respectively, and an output layer. The RNN is an embedding layer followed by an RNN with a dense layer with 10 neurons

and an output layer, while the LSTM is an embedding layer and an LSTM with a 10 neurons dense layer followed by an output layer. The feed-forward neural network, RNN, and LSTM are trained for 50 epochs. The following subsections report the results of the DIET architecture while using different pre-trained embedding models. The DIET architecture is trained on the train data for 100 epochs. It is worth mentioning that, for all the models, 28% of the data is selected randomly as test data (at least one example for each intent) and the remaining is used as training data. Also, because of class imbalance in both intent detection and entity extraction tasks, a macro-averaged F1-score is utilized for both tasks. An acceptable model should perform well on intent detection and entity extraction tasks. So the Average Intent and Entity (AIE) score, an average of macro-averaged F1-score of intent detection and entity extraction, is defined as shown in the following:

$$AIE = \frac{intent\ detection\ F1\text{-}score + entity\ extraction\ F1\text{-}score}{2}$$

4.1.1 Baseline Machine Learning Models (Without Data Augmentation)

The first step is converting the text data to a numeric version. Thus, a count vector has been employed, which Converts a collection of text documents to a matrix of token counts. Naive Bayes, logistic regression, support vector machine, feed-forward neural Network, RNN, and LSTM have been used as the baseline models. The Naive Bayes model achieves an AIE score of 63.2% with an average F1-score of 49.1% and 77.3% for intent detection and entity extraction tasks, respectively, while logistic regression has an AIE score of 62.9% with an average F1-score of 48.6% for intent detection and 77.3% for entity extraction. As shown in Table 4.1 and 4.10, SVM with a linear kernel achieves the highest AIE score of 66.1% with weighted F1-score of 54.2% and 78.1% for intent detection and entity extraction, respectively. Table 4.2 shows the performance of different models in which SVM with linear kernel archives

Table 4.1: Intent detection F1-score, entity extraction F1-score and AIE score for SVM with respect to different kernels (without data augmentation).

Kernel	Intent F1-score	Entity F1-score	AIE score
Without Kernel	45.1%	77.3%	61.2%
Linear Kernel	**54.2%**	**78.1%**	**66.1%**
RBF Kernel	45.1%	77.3%	61.2%

Table 4.2: Intent detection F1-score, entity extraction F1-score and AIE score for Naive Bayes, logistic regression, and SVM models (without data augmentation).

Model	Intent F1-score	Entity F1-score	AIE score
Naive Bayes	49.1%	77.3%	63.2%
Logistic regression	48.6%	77.3%	62.9%
SVM (Linear Kernel)	**54.2%**	**78.1%**	**66.1%**

the highest AIE score. Table 4.3 shows the results of the feed-forward neural network, LSTM, and RNN, in which the feed-forward neural network achieves the AIE score of 44% with a weighted F1-score of 10.8% and 77.3% for intent detection and entity extraction, respectively. The RNN and LSTM yield AIE score of 6.5 with a weighted F1-score of 6.5% for intent detection and 77.2% for entity extraction. Since the number of training data in the MIRA Dataset is not sufficient for training a neural network-based model, the low F1-score of the feed-forward neural network, RNN, and LSTM is predictable. However, since the basic machine learning models do not require a huge amount of training data, a higher F1-score in comparison to neural network-based models is achieved.

Table 4.3: Intent detection F1-score, entity extraction F1-score and AIE score for FeedForward network, RNN, and LSTM (without data augmentation).

Model	Intent F1-score	Entity F1-score	AIE score
FeedForwad	**10.8 %**	**77.3%**	**44%**
RNN	6.5%	77.3%	41.9 %
LSTM	6.5%	77.3%	41.9%

4.1.2 Baseline Machine Learning Models (With Data Augmentation)

Like the previous subsection, Naive Bayes, logistic regression, SVM, feed-forward neural network, RNN, and LSTM are used to examine the performance of the data augmentation techniques on the models. The Naive Bayes model yielded an AIE score of 72.1% and a weighted F1-score of 56.7% and 87.5% for intent detection and entity extraction. Also, logistic regression achieved an AIE score of 70.1% and an weighted F1-score of 52.8% for intent detection and 87.5% for entity extraction. Moreover, the results of SVM while using different kernels are report in Table 4.4, in which the linear kernel had the best AIE score (74.1%) with an average weighted F1-score of 59.8% for intent detection and 88.5% for entity extraction. According to Table 4.7, SVM with a linear kernel yields the best AIE score among basic machine learning models.

The results of the feed-forward neural network, RNN, and LSTM are reported in Table 4.5. The feed-forward neural network achieves the AIE score of 54.8% with a weighted F1-score of 22.1% for intent detection and 87.5% for entity extraction. The RNN and LSTM yield AIE score of 49.3% with a weighted F1-score of 11.2% and 87.5% for intent detection and entity extraction, respectively. By applying data augmentation techniques a small improvement was achieved for the feed-forward neural network, RNN, and LSTM. However, the results are not comparable to the results of basic machine learning models (i.e., Naive Bayes, logistic regression, and SVM) since the amount of training data after the augmentation is not sufficient enough for training a neural network based model. In this case, using pre-trained embedding models like ConveRT

Table 4.4: Intent detection F1-score, entity extraction F1-score and AIE score for SVM with respect to different kernels (with data augmentation).

Kernel	Intent F1-score	Entity F1-score	AIE score
Without Kernel	45.6%	87.5%	66.5%
Linear Kernel	**59.8%**	**88.5%**	**74.1%**
RBF Kernel	45.6%	87.5%	66.5%

Table 4.5: Intent detection F1-score, entity extraction F1-score and AIE score for FeedForward network, RNN, and LSTM (with data augmentation).

Model	Intent F1-score	Entity F1-score	AIE score
FeedForwad	**22.1%**	**87.5%**	**54.8%**
RNN	11.2%	87.5%	49.3 %
LSTM	11.2%	87.5%	49.3%

that have been trained on a large corpus is useful.

Table 4.6 compares the AIE score of SVM with respect to different kernels with and without data augmentation in which the linear kernel with data augmentation had the best AIE score of 74.1%.

4.1.3 DIET Architecture (Without Data Augmentation)

Table 4.8 depicts the results of DIET architecture (without augmented dataset) with different embedding (i.e., Count Vector featurizer, ConveRT featurizer,

Table 4.6: AIE for SVM concerning different kernels.

Kernel	AIE Score without data augmentatio	AIE Score with data augmentatio
without Kernel	61.2%	66.5%
Linear Kernel	66.1%	**74.1%**
RBF Kernel	61.2%	66.5%

Table 4.7: Intent detection F1-score, entity extraction F1-score and AIE for Naive Bayes, logistic regression, and SVM models (with data augmentation).

Model	Intent F1-score	Entity F1-score	AIE score
Naive Bayes	56.7%	87.5%	72.1%
Logistic regression	52.8%	87.5%	70.1%
SVM (Linear Kernel)	**59.8%**	**88.5%**	**74.1%**

Table 4.8: Intent detection F1-score, entity extraction F1-score and AIE for DIET architecture (without data augmentation).

Model	Intent F1-score	Entity F1-score	AIE score
CountVector featurizer	66.9%	57%	61.9%
ConveRT featurizer	**76%**	**65.6%**	**70.8%**
Bert embedding	72%	59%	65.5%

and Bert embedding) in which ConveRT featurizer has the highest average F1-score of 76% and 65.6% for intent detection and entity extraction tasks, respectively and it has AIE score of 70.8%.

4.1.4 DIET Architecture (With Data Augmentation)

Table 4.9 shows the results of the DIET architecture (with the augmented dataset) with different embedding (i.e., CountVector featurizer, ConveRT featurizer, and Bert embedding) in which conveRT featurizer yields the highest averaged F1-score for both intent detection (99.1%) and entity extraction (95.4%) tasks. The AIE score of the DIET architecture with ConveRT featurizer is 97.2% which has a 23.8% improvement compared to the best result of basic machine learning models (i.e., SVM with an AIE score of 73.4%). Table 4.10 compares the results of basic machine learning models and DIET architecture with and without data augmentation. For all the models, we see

Table 4.9: Intent detection F1-score, entity extraction F1-score and AIE for DIET architecture (with data augmentation).

Model	Intent F1-score	Entity F1-score	AIE score
CountVector featurizer	78.8%	68.4%	73.6%
Bert embedding	82.3%	72.8%	77.5%
ConveRT featurizer	**99.1%**	**95.4%**	**97.2%**

a positive impact of data augmentation on AIE score except Naive Bayes. For DIET architecture (ConveRT featurizer), an improvement of 26.4% is achieved.

Table 4.10 compares the results of basic machine learning models, the feed-forward neural network, RNN, LSTM, and DIET architecture with and without data augmentation. Data augmentation positively impacted all the models. The DIET architecture with ConveRT featurizer achieves the highest AIE score of 97.2% so it is used for intent detection and entity extraction within the MIRA Chatbot.

4.2 Performance of the MIRA Chatbot

In this section, variables like the number of conversational turns have been reported to evaluate the performance of the MIRA chatbot. It is worth mentioning that the variables are reported from May 2nd, 2022, the date the chatbot was released to the public.

Number of Conversational Turns: One method of user engagement evaluation is to measure the average number of daily conversational turns (i.e., the number of back-and-forth between a user and the bot), as reported in Fig. 4.1. The average number of daily conversational turns is 3.4. The minimum average daily conversational turns are 1, while the maximum average daily conversational turns are 15. As the initial pilot of this chatbot, there is no other data to compare these results to, so this particular information has

Table 4.10: AIE score with and without data augmentation.

model	AIE Score without data augmentation	AIE Score with data augmentation
Naive Bayes	63.2%	72.1%
Logistic regression	62.9%	70.1%
SVM (linear kernel)	66.1%	74.1%
FeedForward	44%	54.8%
RNN	41.9%	49.3%
LSTM	41.9%	49.3%
DIET architecture (Countvector feature)	61.9%	73.6%
DIET architecture (ConveRT featurizer)	**70.8%**	**97.2%**
DIET architecture (Bert embedding)	65.5%	77.5%

minimal impact on the analysis provided in this report. Also, since one can receive a resource by at least two turns, 3.4 average conversational turns do not provide any conclusion. The results demonstrated here, however, will support future research. It is determined that alternative methods of user engagement evaluation are necessary.

If a user is offered at least one resource, it implies that the user is motivated to continue chatting. Hence, user engagement can be considered the percentage of users offered with at least one resource. The average number of daily users offered with at least one resource is 30.7%, as reported in Fig. 4.2, therefore, the user engagement can be considered as 30.7%. The minimum number of daily users offered with at least one resource is 0.7%, while the maximum is 100%.

To interpret the 30.7% of user engagement, a complete evaluation of users' actions should be considered. Hence, Fig. 4.3 is provided in which 32% of the users have consented to the information, while 6.8% have not consented. Also, 61% of the users have not started a conversation, which needs a thorough evaluation of the users' behavior to understand why 61% of the users

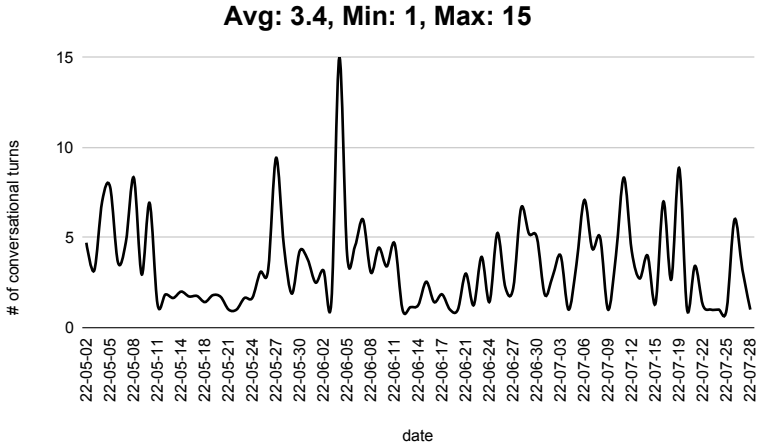

Figure 4.1: Number of conversational turns per day.

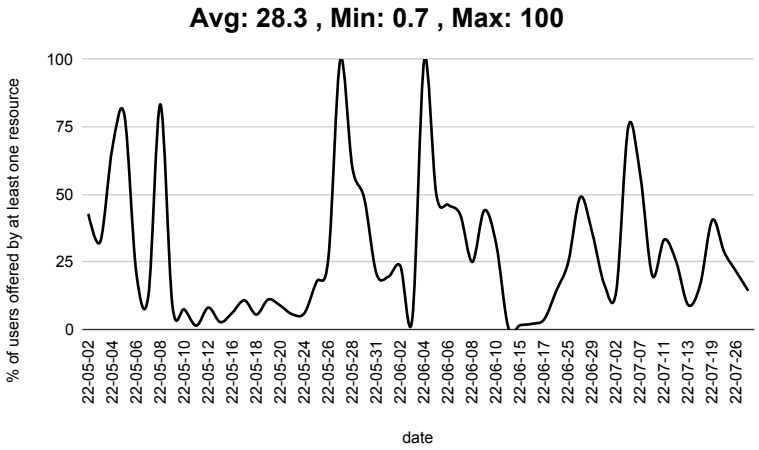

Figure 4.2: The percentage of users offered with at least one resource daily.

have not started a conversation. One possibility is that some web crawlers are using the chatbot for webpage indexing or other purposes. These web crawlers originate from all over the world. Since Canada is the target country in the MIRA project, any request from outside of Canada can be considered as a web crawler. So, extraction of location from IP addresses that sent a request to the chatbot can be helpful. Fig. 4.4 shows the location of the first 900 requests to the MIRA Chatbot, in which most of the request were from outside of Canada. Hence, we can hypothesize that most of the 61% users that did not start a conversation with the chatbot were web crawlers.

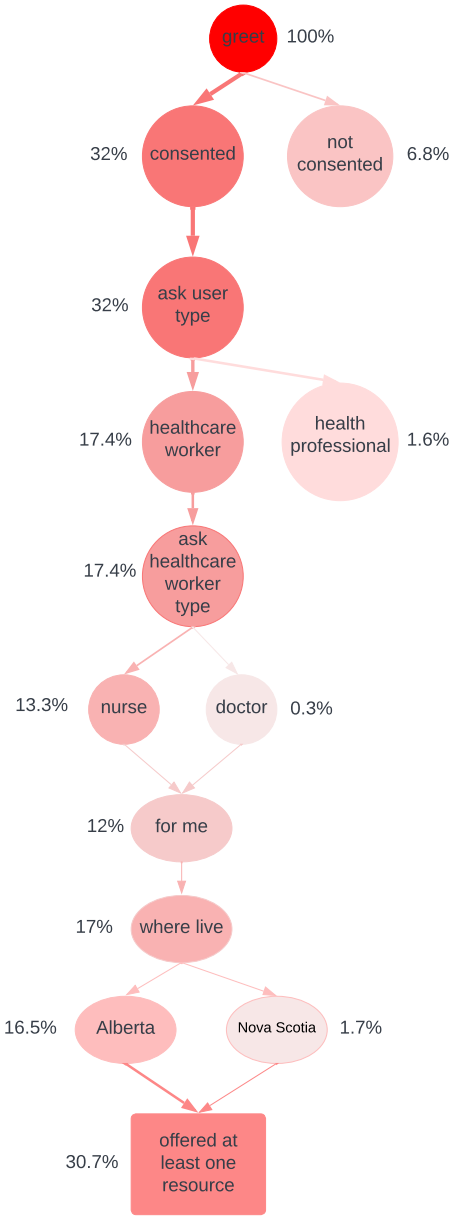

Figure 4.3: The steps users passed to reach at least one resource. The number near each node indicates the percentage of the users who reached that node.

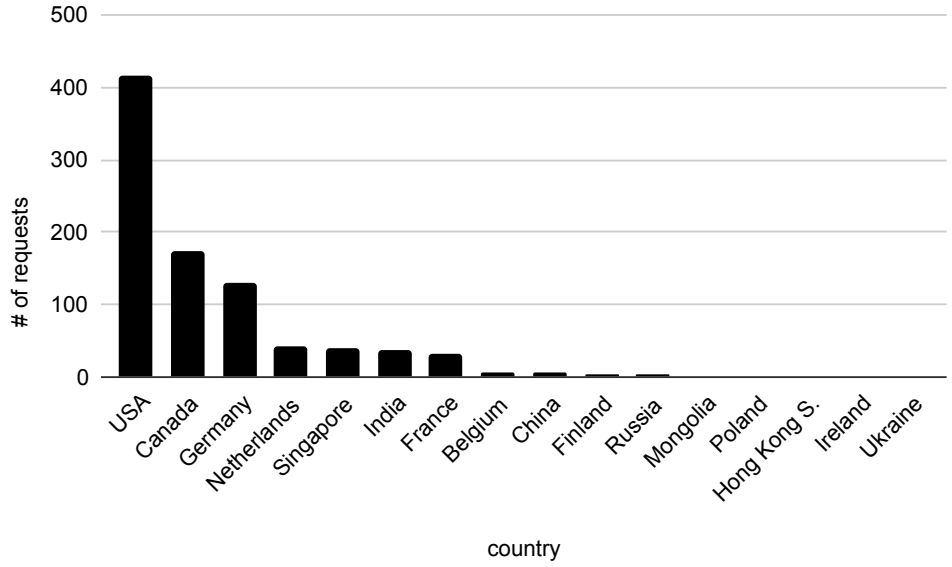

Figure 4.4: Location of first 900 requests to the MIRA Chatbot.

Intent Detection Accuracy: As discussed in the previous section, the intent detection F1-score of the DIET classifier with ConveRT featurizer is 99.1%. More evaluation of intent detection performance in detecting users' intent is required. Therefore, the accuracy of detected intents is reported in Fig. 4.5. The average intent detection accuracy of 91% is achieved. It is worth mentioning that the intents with an accuracy below 60% are not acceptable, and they are classified as the "NLU_fallback" class, and the chatbot asks the users to rephrase their sentences. Hence, the minimum intent detection accuracy is 60%. Also, the maximum intent detection accuracy is 99.9%.

Entity Extraction Accuracy: The DIET classifier yielded the average F1-score of 95.4% for entity extraction. To calculate the performance of the trained DIET classifier in extracting entities within MIRA Chatbot Fig. 4.6 is provided, which shows the accuracy of each entity (i.e., city, age, job, MH_concern, who, and resource_type) with an average accuracy of 92.9%. The City entity achieved the highest accuracy of 97.8%, while the MH_concern entity had the lowest accuracy of 87.7%. Hence, adding more training data for

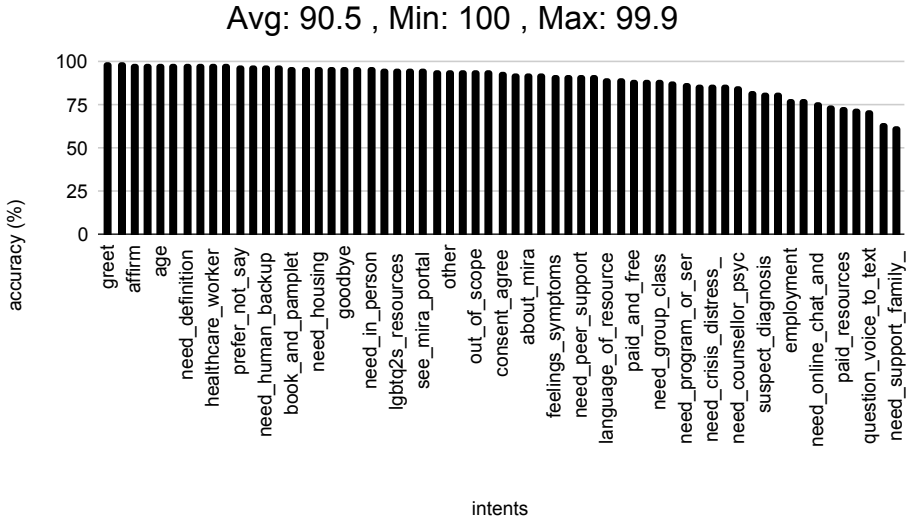

Figure 4.5: Intent detection accuracy.

improving the accuracy of the MH_concern entity is required, which can be considered for future work.

Number of Entities Extracted by DIET Classifier and Regular Expression: A regular expression component has been added to improve the overall performance of the entity extraction. Fig. 4.7 depicts the number of extracted entities by regular expression and DIET classifier. The greater number of extracted entities, 2530, is achieved with the regular expression component, while the DIET classifier has extracted 1665 entities. The combination of extracted entities by regular expression and DIET classifier is used for the MIRA Chatbot.

Conversation Length: Another measure of user engagement is average conversation length. As illustrated in Fig. 4.8, the daily conversation length is measured by the duration between a user's first and last message in seconds. The conversation length measured for all days is an average of 54.3 seconds. The date with the lowest average conversation was recorded at 0.1 seconds, and the maximum average time of 276.7 seconds. It would be beneficial to

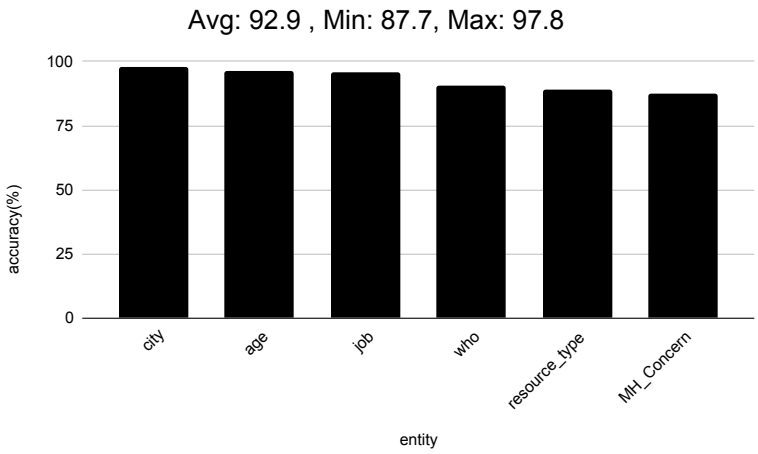

Figure 4.6: Entity extraction accuracy.

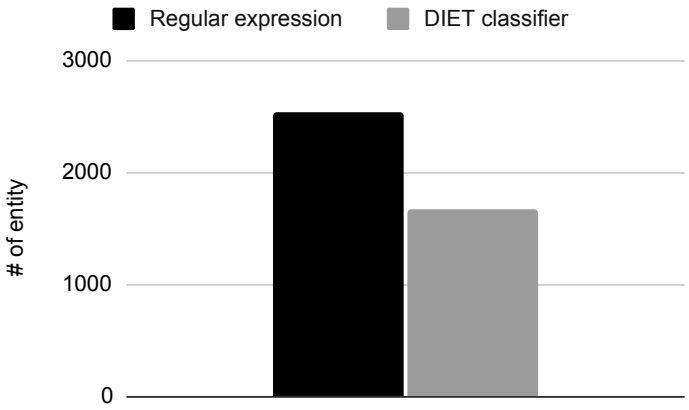

Figure 4.7: Number of extracted entities by regular expression and DIET classifier.

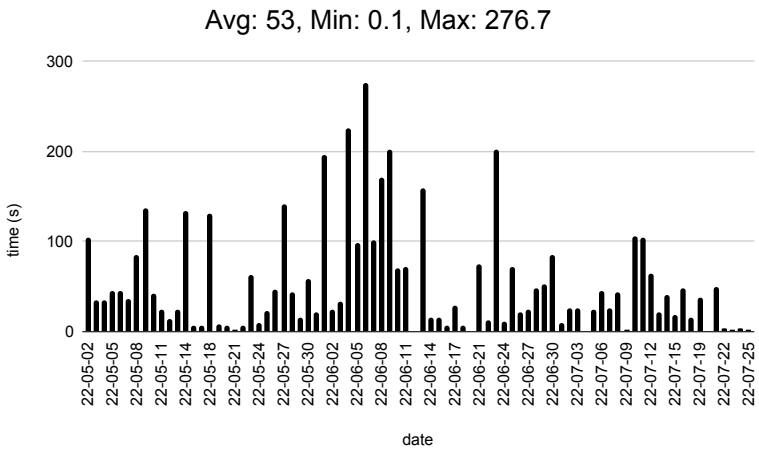

Figure 4.8: Conversation length per day.

analyze the average time needed for a user to chat with MIRA to receive a resource. Comparing the calculated average time to the average conversation length reveals whether techniques are needed to increase conversation length.

Number of Complete Versus Incomplete Conversations: If a user reaches a point where the chatbot replies with a goodbye message, the conversation will be marked as a complete conversation. Otherwise, it is marked as an incomplete conversation. Fig. 4.9 shows a much higher number of complete conversations, 2268, from a total of 2305 conversations. This phenomenon is because the users' motivation declines once the chatbot provides them with a proper resource. The complete conversations are beneficial for evaluating the chatbot since the chatbot asks the users to rate the chatbot at the end. Hence, Employing strategic techniques to motivate the users to have a complete conversation can be considered for future work.

Initial Load Time: Initial load time is defined as the time the chatbot requires to start the conversation. The average load time per day is depicted in Fig. 4.10. There was an acceptable average of 1.1 seconds for initial load time for all dates recorded. The lowest average load time in a day was reported

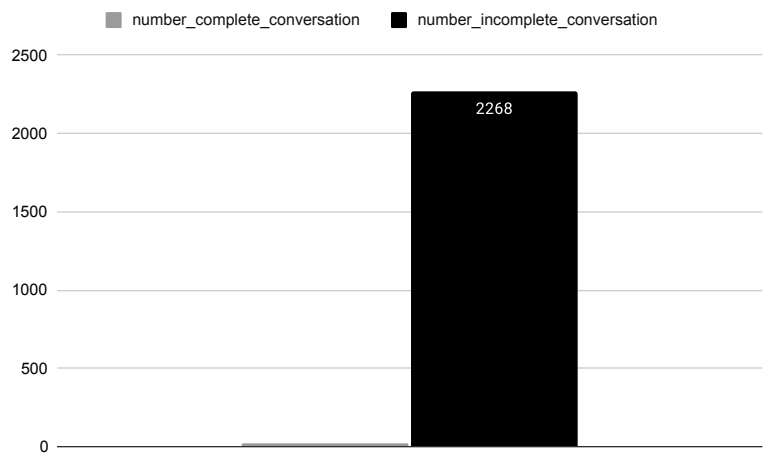

Figure 4.9: Number of complete versus incomplete conversations.

to be 0.1 seconds, whereas the maximum average load time was 12.1 seconds. The initial load time could be decreased by completing a thorough evaluation of the timing of different parts of the MIRA chatbot. Strategic methods, such as increasing the hardware resources (i.e., CPU, RAM), could decrease the initial load time.

Provinces/Cities that Used the Chatbot: Asking users' locations can be beneficial for customizing the offered resources. Moreover, evaluation of several users connected from different provinces/cities can reveal the effectiveness of the advertisements. Hence, Fig. 4.11 is provided, which shows Alberta and Nova Scotia had the highest number of users as they are the target provinces for the MIRA project. According to Fig. 4.11 more advertisements are needed for other provinces like Ontario and Quebec.

Number of Resource Types: Evaluation of a number of requested resources can be beneficial since more users can take advantage of high-quality resources by improving them. Fig. 4.12 reports a number of resource types in which doctor resource type has been requested more often.

Job: Another statistic that can help to customize resources is the user's job.

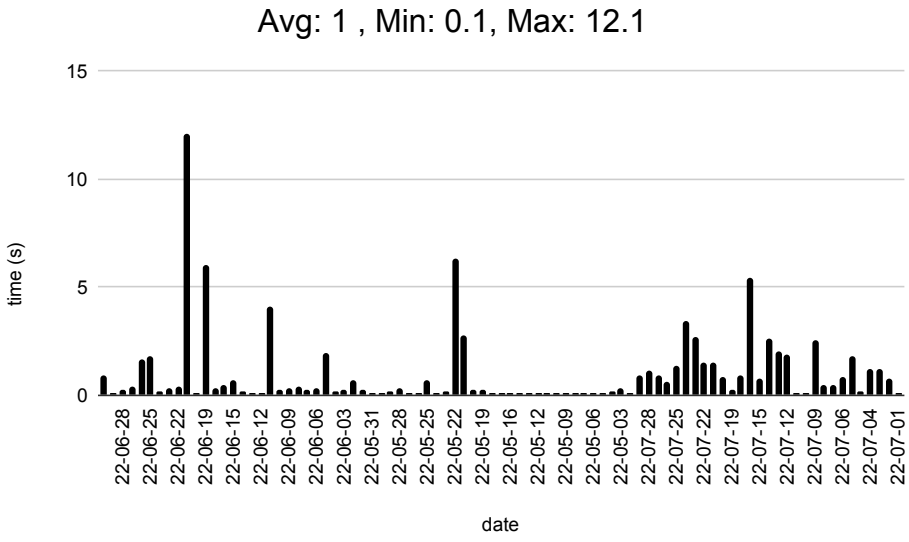

Figure 4.10: The average initial load time.

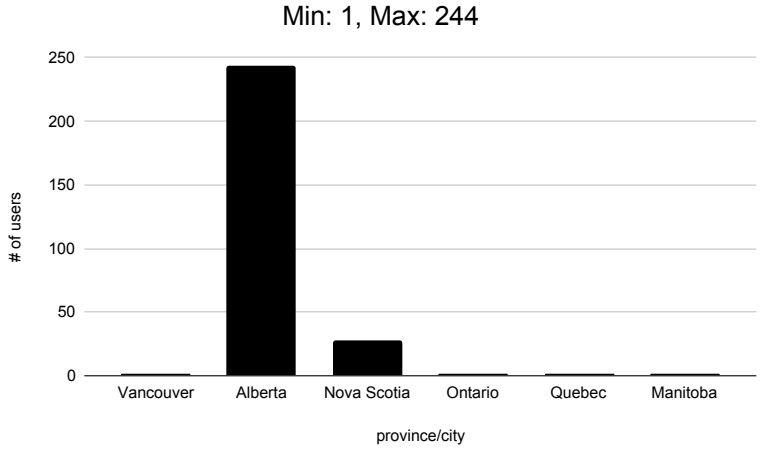

Figure 4.11: Number of users connected from different provinces/cities.

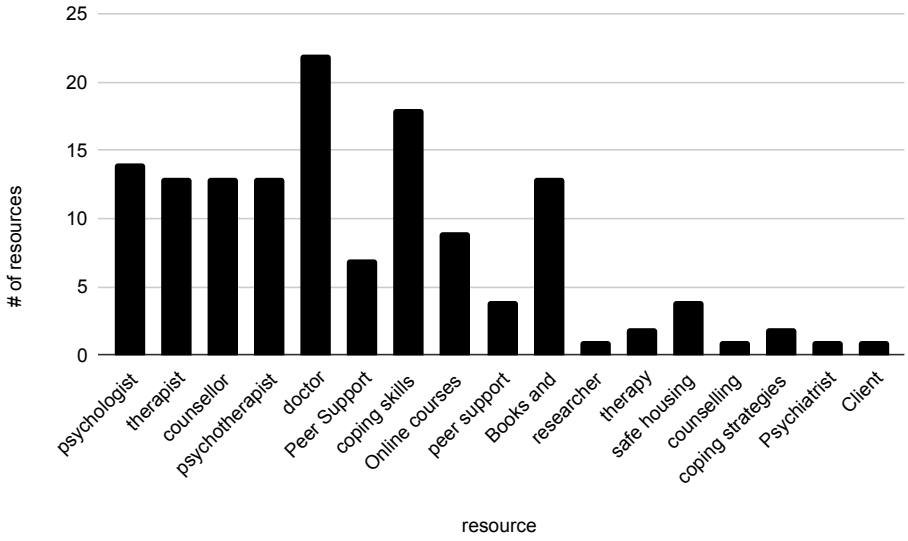

Figure 4.12: Number of resource types requested by the users.

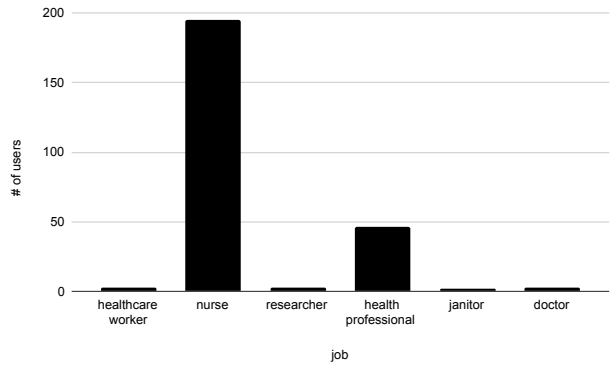

Figure 4.13: Number of jobs.

Hence, in Fig. 4.13 the number of jobs are reported in which most of the users are nurses. So, providing more resources for nurses can help most clients.

Who Entity: The chatbot asks the users whether they are looking for information for themselves or someone else. If users want information for someone else, the chatbot asks them to provide more information. In Fig. 4.14, a number of different values for each "who" entity is provided. The "family member" and "husband" had the highest number. Hence, providing more resources for them can improve the user's satisfaction of most of the users.

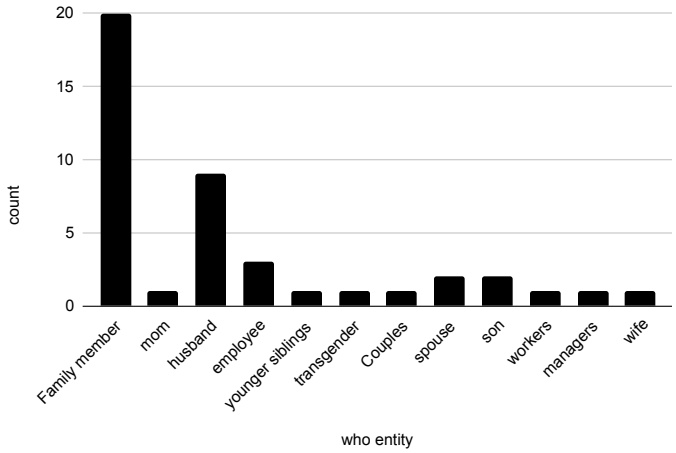

Figure 4.14: Total number of each "who" entity.

Age: It would be beneficial to consider making the chatbot's behavior responsive according to a user's age so as to provide more satisfactory services. As illustrated in Fig. 4.15, most users who engaged with the app each day were over 18.

Extracted Entities: Reporting the number of extracted entities can be useful since one can prioritize improving the resources related to the most extracted entities. Fig. 4.16 reports the number of extracted entities in which depression has the highest number. Moreover, Table 4.11 shows each entity label's count, average, minimum, and maximum, in which MH_Concern has the highest count. So, most users were looking for resources related to depression and MH_Concern. Hence, improvement of the resources related to depression and MH_Concern can enhance most users' user satisfaction.

Star Rating: For evaluating the whole system's performance, a star rating is added at the end of the conversation with the MIRA chatbot. According to Fig. 4.17 the average daily star rating of 4.6 is achieved.

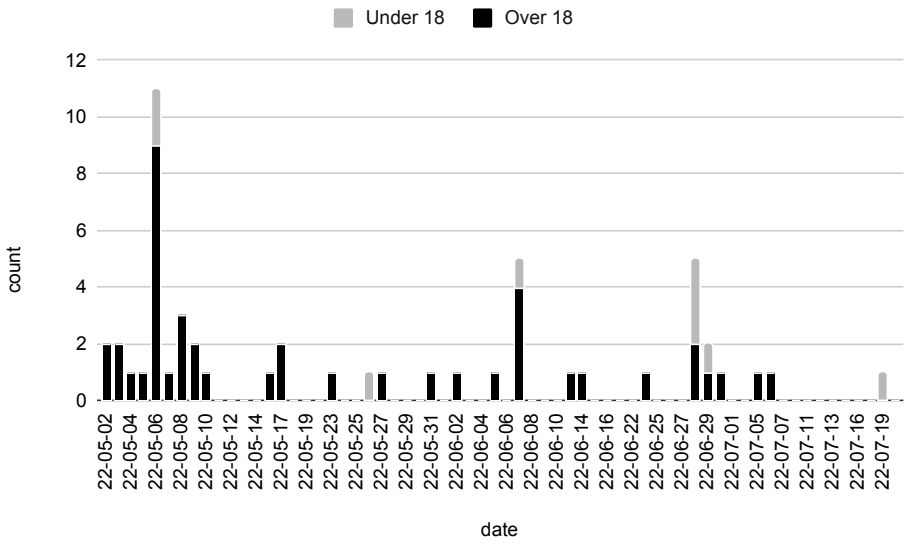

Figure 4.15: Age of the users.

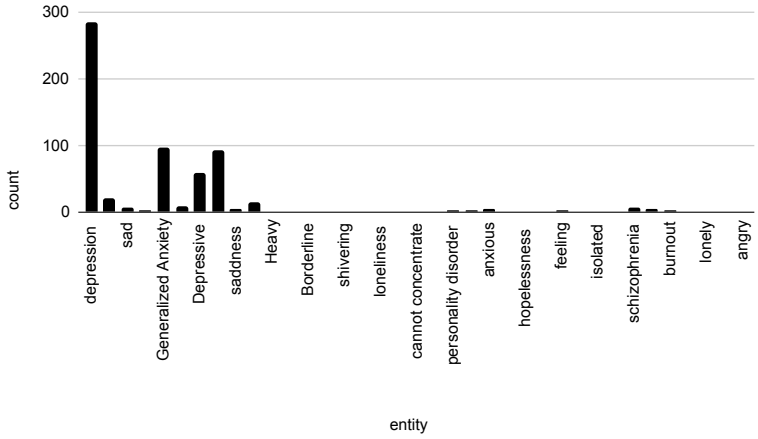

Figure 4.16: Number of extracted entities.

Table 4.11: Number of entity labels.

Entities	Count	Average	Minimum	Maximum
mental health concern	453	7.1	0	71
resource type	449	7	0	71
city	273	4.3	0	72
job	266	4.2	0	73
age	52	0.8	0	11
who	42	0.7	0	9

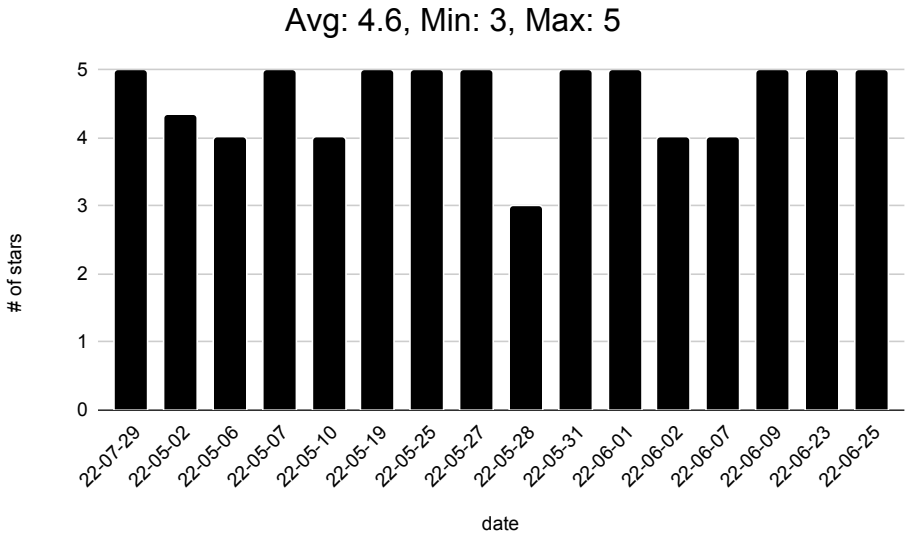

Figure 4.17: Average star rating per day.

Chapter 5

Conclusion and Future Work

The world is undergoing a period of significant growth in technological innovation. Starting with the Internet, technological networks and systems have emerged as so complex and disruptive that they have transformed not only our governing and economic structures, but also our perception of self, community, and day-to-day life. With 8 million global deaths attributed annually to mental illness [141], there is an urgency to identify effective and timely service options that reduce and eliminate barriers, including through health system navigation, as well as investigate innovations where the technology may present constructive novel solutions.

We have developed a mental health chatbot (MIRA chatbot) to help healthcare workers and their families with mental health issues by offering resources. The system uses a transformer-based architecture (DIET architecture) to detect intent and extract features simultaneously. The chatbot is intended to be a source of information, not a replacement for medical advice. Alongside MIRA chatbot, a resource portal was developed by one of the MIRA team members so that experts can add or alter the resource information. We contributed to the designing of the MIRA resource portal and the makeup of the interface for communication between the MIRA chatbot and MIRA resource portal. Moreover, we participated in developing the user interface of the MIRA chatbot, but being out of scope of this thesis, we did not explicitly report on them. The BERT and ConveRT featurizers are compared within the DIET architecture, in which the ConveRt featurizer achieved the highest averaged

intent and entity F1-score of 97.2%.

From the chatbot evaluation, we found out that 61% of the users did not start a conversation with the chatbot. Through evaluation of users' IP addresses, we figured out that most of the requests to the MIRA Chatbot came from countries that are not among our target countries for the MIRA project. Hence, we hypothesize that some web crawlers are using the chatbot. Therefore, 61% can be bots, not real users. The averaged accuracy of intent detection and entity extraction are 90.5%, and 92.9%, respectively, which shows the effectiveness of the DIET architecture in detecting the intent of users and extracting information from users' utterances.

If one wants to use MIRA Chatbot for another use cases, the following items should change:

- Add enough resources to the MIRA resource portal about the use case.
- Create or use the existing datasets for training the DIET architecture.
- Construct a tree of all the actions that the MIRA Chatbot should follow.
- Define a list of responses that the MIRA Chatbot should use for replying to the clients.

By applying the following items, one can use MIRA Chatbot for other use cases.

A direction for future consideration is the incorporation of emotional intelligence into dialogue generation to better imitate human conversational patterns and appropriately respond to emotional input. The existing chatbot can be criticized for being limited in response length or for producing generic or noncommittal versus empathetic or emotionally intelligent responses. Future studies should explore the integration of empathetic response generation that appropriately categorizes a client's current emotional state based on their input utterance, considers the desired target emotion to guide clients toward, and subsequently generates an emotionally intelligent response back to clients incorporating these considerations. Multilabel emotion mining may be considered to support this categorization [52], [116].

Although deep learning models currently can conduct language processing tasks such as tagging, text classification, machine translation, and question answering, existing, state-of-the-art models are criticized for lacking "explainability" - more specifically being able to describe how the algorithm came to a particular result or action, which is considered a key pillar in the discourse around ethical AI development [24], [52]. Future studies must seek to improve methods of explainable natural language processing.

There are some features in most chatbots like asking the chatbot to tell a joke or asking the chatbot about the weather. To have these kinds of features within MIRA chatbot, one can add a proper intent like "joke" intent to the list of intents and program the chatbot to have the features. Adding these features is considered for future work.

There are several anticipated limitations of note that we consider unavoidable. Digital interventions are not accessible to all Canadians, and there are barriers to their use, including technical issues with connectivity; lack of access to electrical or technological infrastructure because of cost, service provision, and natural disasters; and distrust of technology regarding the use of data or protection of anonymity [130].

References

[1] 2022. [Online]. Available: https://github.com/howdyai/botkit.

[2] "A national commitment to recovery from the disease of addiction in canada,," 2015. [Online]. Available: https://www.ccsa.ca/sites/default/files/2019-04/CCSA-Recovery-Oriented-System-of-Care-Resource-2017-en.pdf.

[3] A. A. Abd-Alrazaq, A. Rababeh, M. Alajlani, B. M. Bewick, and M. Househ, "Effectiveness and safety of using chatbots to improve mental health: Systematic review and meta-analysis," *Journal of medical Internet research*, vol. 22, no. 7, e16021, 2020.

[4] S. A. Abdul-Kader and J. Woods, "Question answer system for online feedable new born chatbot," in *2017 Intelligent Systems Conference (IntelliSys)*, IEEE, 2017, pp. 863–869.

[5] A. Agarwal, S. Maiya, and S. Aggarwal, "Evaluating empathetic chatbots in customer service settings," *arXiv preprint arXiv:2101.01334*, 2021.

[6] R. Agarwal and M. Wadhwa, "Review of state-of-the-art design techniques for chatbots," *SN Computer Science*, vol. 1, no. 5, pp. 1–12, 2020.

[7] E. H. Almansor and F. K. Hussain, "Survey on intelligent chatbots: State-of-the-art and future research directions," in *Conference on Complex, Intelligent, and Software Intensive Systems*, Springer, 2019, pp. 534–543.

[8] M. Bagchi, "Conceptualising a library chatbot using open source conversational artificial intelligence.," *DESIDOC Journal of Library & Information Technology*, vol. 40, no. 6, 2020.

[9] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[10] M. Bayer, M.-A. Kaufhold, and C. Reuter, "A survey on data augmentation for text classification," *ACM Comput. Surv.*, Jun. 2022, Just Accepted, ISSN: 0360-0300. DOI: 10.1145/3544558. [Online]. Available: https://doi.org/10.1145/3544558.

[11] Y. Bengio, R. Ducharme, and P. Vincent, "A neural probabilistic language model," *Advances in neural information processing systems*, vol. 13, 2000.

[12] T. W. Bickmore, D. Schulman, and C. Sidner, "Automated interventions for multiple health behaviors using conversational agents," *Patient education and counseling*, vol. 92, no. 2, pp. 142–148, 2013.

[13] T. W. Bickmore, R. A. Silliman, K. Nelson, *et al.*, "A randomized controlled trial of an automated exercise coach for older adults," *Journal of the American Geriatrics Society*, vol. 61, no. 10, pp. 1676–1683, 2013.

[14] D. Birrell and D. Heenan, "Implementing the transforming your care agenda in northern ireland within integrated structures," *Journal of Integrated Care*, vol. 20, Nov. 2012. DOI: 10.1108/14769011211285156.

[15] L. Bradeško and D. Mladenić, "A survey of chatbot systems through a loebner prize competition," in *Proceedings of Slovenian language technologies society eighth conference of language technologies*, Institut Jožef Stefan Ljubljana, Slovenia, 2012, pp. 34–37.

[16] S. Bucci, M. Schwannauer, and N. Berry, "The digital revolution and its impact on mental health care," *Psychology and Psychotherapy: Theory, Research and Practice*, vol. 92, no. 2, pp. 277–297, 2019.

[17] P. Budzianowski and I. Vulić, *Hello, it's gpt-2 – how can i help you? towards the use of pretrained language models for task-oriented dialogue systems*, 2019. DOI: 10.48550/ARXIV.1907.05774. [Online]. Available: https://arxiv.org/abs/1907.05774.

[18] T. Bunk, D. Varshneya, V. Vlasov, and A. Nichol, *Diet: Lightweight language understanding for dialogue systems*, 2020. DOI: 10.48550/ARXIV.2004.09936. [Online]. Available: https://arxiv.org/abs/2004.09936.

[19] ——, *Diet: Lightweight language understanding for dialogue systems*, 2020. DOI: 10.48550/ARXIV.2004.09936. [Online]. Available: https://arxiv.org/abs/2004.09936.

[20] B. Byrne, K. Krishnamoorthi, C. Sankar, *et al.*, "Taskmaster-1: Toward a realistic and diverse dialog dataset," *arXiv preprint arXiv:1909.05358*, 2019.

[21] S. Cabarkapa, S. E. Nadjidai, J. Murgier, and C. H. Ng, "The psychological impact of covid-19 and other viral epidemics on frontline healthcare workers and ways to address it: A rapid systematic review," *Brain, behavior, & immunity-health*, vol. 8, p. 100 144, 2020.

[22] J. Cahn, "Chatbot: Architecture, design, & development," *University of Pennsylvania School of Engineering and Applied Science Department of Computer and Information Science*, 2017.

[23] G. Cameron, D. Cameron, G. Megaw, *et al.*, "Best practices for designing chatbots in mental healthcare: A case study on ihelpr," in *Proceedings of the 32nd International BCS Human Computer Interaction Conference*, ser. HCI '18, Belfast, United Kingdom: BCS Learning amp; Development Ltd., 2018. DOI: 10.14236/ewic/HCI2018.129. [Online]. Available: https://doi.org/10.14236/ewic/HCI2018.129.

[24] "Canadian internet use survey," 2019. [Online]. Available: https://www150.statcan.gc.ca/n1/daily-%20quotidien/191029/dq191029a-eng.htm.

[25] "Centre for addiction and mental health, anxiety, feelings of depression and loneliness among canadians spikes to highest levels since spring 2020," 2022. [Online]. Available: https://www.camh.ca/en/camh-news-and-stories/anxiety-depression-loneliness-among-canadians-spikes-to-highest-levels.

[26] R. Chada, *Simultaneous paraphrasing and translation by fine-tuning transformer models*, 2020. DOI: 10.48550/ARXIV.2005.05570. [Online]. Available: https://arxiv.org/abs/2005.05570.

[27] Y.-N. Chen, T. Bedrax-Weiss, D. Hakkani-Tur, *et al.*, "Proceedings of the first workshop on nlp for conversational ai," in *Proceedings of the First Workshop on NLP for Conversational AI*, 2019.

[28] K. Chlasta, K. Wołk, and I. Krejtz, "Automated speech-based screening of depression using deep convolutional neural networks," *Procedia Computer Science*, vol. 164, pp. 618–628, Dec. 2019. DOI: 10.1016/j.procs.2019.12.228.

[29] K. Cho, B. van Merrienboer, C. Gulcehre, *et al.*, *Learning phrase representations using rnn encoder-decoder for statistical machine translation*, 2014. DOI: 10.48550/ARXIV.1406.1078. [Online]. Available: https://arxiv.org/abs/1406.1078.

[30] K. M. Colby, "Ten criticisms of parry," *ACM SIGART Bulletin*, no. 48, pp. 5–9, 1974.

[31] K. M. Colby, S. Weber, and F. D. Hilf, "Artificial paranoia," *Artificial Intelligence*, vol. 2, no. 1, pp. 1–25, 1971, ISSN: 0004-3702. DOI: https://doi.org/10.1016/0004-3702(71)90002-6. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0004370271900026.

[32] R. Csaky, *Deep learning based chatbot models*, Nov. 2017. DOI: 10.13140/RG.2.2.21857.40801.

[33] ——, *Deep learning based chatbot models*, 2019. DOI: 10.48550/ARXIV.1908.08835. [Online]. Available: https://arxiv.org/abs/1908.08835.

[34] R. Csaky and G. Recski, "The gutenberg dialogue dataset," *arXiv preprint arXiv:2004.12752*, 2020.

[35] K. Denecke, S. Vaaheesan, and A. Arulnathan, "A mental health chatbot for regulating emotions (sermo) - concept and usability test," *IEEE Transactions on Emerging Topics in Computing*, vol. 9, no. 3, pp. 1170–1182, 2021. DOI: 10.1109/TETC.2020.2974478.

[36] M. Denton, J. Ploeg, J. Tindale, *et al.*, "Where would you turn for help? older adults' awareness of community support services," *McMaster University, Social and Economic Dimensions of an Aging Population Research Papers*, vol. 27, Jan. 2010. DOI: 10.3138/cja.27.4.359.

[37] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, 2018. DOI: 10.48550/ARXIV.1810.04805. [Online]. Available: https://arxiv.org/abs/1810.04805.

[38] H. Dihingia, S. Ahmed, D. Borah, S. Gupta, K. Phukan, and M. K. Muchahari, "Chatbot implementation in customer service industry through deep neural networks," in *2021 International Conference on Computational Performance Evaluation (ComPE)*, IEEE, 2021, pp. 193–198.

[39] "Disconnected relationships between primary care and community-based health and social services and system navigation for older adults: A qualitative descriptive study," 2020.

[40] R. Edwards, T. Bickmore, L. Jenkins, M. Foley, and J. Manjourides, "Use of an interactive computer agent to support breastfeeding," *Maternal and child health journal*, vol. 17, Jan. 2013. DOI: 10.1007/s10995-013-1222-0.

[41] S. Fernandes, R. Gawas, P. Alvares, M. Femandes, D. Kale, and S. Aswale, "Survey on various conversational systems," in *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, IEEE, 2020, pp. 1–8.

[42] R. T. Fielding, *Architectural styles and the design of network-based software architectures*. University of California, Irvine, 2000.

[43] L. Fryer and R. Carpenter, "Bots as language learning tools," *Language Learning & Technology*, vol. 10, no. 3, pp. 8–14, 2006.

[44] C.-W. Goo, G. Gao, Y.-K. Hsu, *et al.*, "Slot-gated modeling for joint slot filling and intent prediction," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 753–757. DOI: 10.18653/v1/N18-2118. [Online]. Available: https://aclanthology.org/N18-2118.

[45] F. Guo, A. Metallinou, C. Khatri, A. Raju, A. Venkatesh, and A. Ram, "Topic-based evaluation for conversational bots," 2018. DOI: 10.48550/ARXIV.1801.03622. [Online]. Available: https://arxiv.org/abs/1801.03622.

[46] A. Harris and S. H. Jones, "Words," in *Writing for Performance*, Springer, 2016, pp. 19–35.

[47] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.

[48] M. Henderson, I. Casanueva, N. Mrkšić, P.-H. Su, T.-H. Wen, and I. Vulić, *Convert: Efficient and accurate conversational representations from transformers*, 2019. DOI: 10.48550/ARXIV.1911.03688. [Online]. Available: https://arxiv.org/abs/1911.03688.

[49] ——, *Convert: Efficient and accurate conversational representations from transformers*, 2019. DOI: 10.48550/ARXIV.1911.03688. [Online]. Available: https://arxiv.org/abs/1911.03688.

[50] M. Henderson, I. Vulić, D. Gerz, *et al.*, *Training neural response selection for task-oriented dialogue systems*, 2019. DOI: 10.48550/ARXIV.1906.01543. [Online]. Available: https://arxiv.org/abs/1906.01543.

[51] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013, vol. 398.

[52] C. Huang, A. Trabelsi, X. Qin, N. Farruque, L. Mou, and O. Zaıane, "Seq2Emo: A sequence to multi-label emotion classification model," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online: Association for Computational Linguistics, Jun. 2021, pp. 4717–4724. DOI: 10.18653/v1/2021.naacl-main.375. [Online]. Available: https://aclanthology.org/2021.naacl-main.375.

[53] K. Humphreys and E. Klaw, "Can targeting nondependent problem drinkers and providing internet-based services expand access to assistance for alcohol problems? a study of the moderation management self-help/mutual aid organization," *Journal of studies on alcohol*, vol. 62, pp. 528–32, Aug. 2001. DOI: 10.15288/jsa.2001.62.528.

[54] S. Hussain, O. Sianaki, and N. Ababneh, "A survey on conversational agents/chatbots classification and design techniques," in Mar. 2019, pp. 946–956, ISBN: 978-3-319-98284-7. DOI: 10.1007/978-3-030-15035-8_93.

[55] B. Hutchison, J.-F. LEVESQUE, E. Strumpf, and N. Coyle, "Primary health care in canada: Systems in motion," *The Milbank Quarterly*, vol. 89, no. 2, pp. 256–288, 2011.

[56] K. Hwerbi, "An ontology-based chatbot for crises management: Use case coronavirus," *arXiv preprint arXiv:2011.02340*, 2020.

[57] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intell. Data Anal.*, vol. 6, no. 5, pp. 429–449, Oct. 2002, ISSN: 1088-467X.

[58] Q. Jiao and S. Zhang, "A brief survey of word embedding and its recent development," in *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, vol. 5, 2021, pp. 1697–1701. DOI: 10.1109/IAEAC50856.2021.9390956.

[59] "Kids can't wait: 2020 report on wait lists and wait times for child and youth mental health care in ontario," 2020. [Online]. Available: https://cmho.org/wp-content/uploads/CMHO-Report-WaitTimes-%202020.pdf.

[60] S.-W. Kim and J.-M. Gil, "Research paper classification systems based on tf-idf and lda schemes," *Hum.-Centric Comput. Inf. Sci.*, vol. 9, no. 1, pp. 1–21, Dec. 2019, ISSN: 2192-1962. DOI: 10.1186/s13673-019-0192-7. [Online]. Available: https://doi.org/10.1186/s13673-019-0192-7.

[61] I. Kissos and N. Dershowitz, "Ocr error correction using character correction and feature-based word classification," in *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, IEEE, 2016, pp. 198–203.

[62] O. Kolomiyets, S. Bethard, and M.-F. Moens, "Model-portability experiments for textual temporal analysis," in *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, ACL; East Stroudsburg, PA, vol. 2, 2011, pp. 271–276.

[63] K. Kypri and H. M. McAnally, "Randomized controlled trial of a web-based primary care intervention for multiple health risk behaviors," *Preventive medicine*, vol. 41, no. 3-4, pp. 761–766, 2005.

[64] J. Lafferty, A. Mccallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," Jan. 2001, pp. 282–289.

[65] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, *Neural architectures for named entity recognition*, 2016. DOI: 10.48550/ARXIV.1603.01360. [Online]. Available: https://arxiv.org/abs/1603.01360.

[66] L. Laranjo, A. G. Dunn, H. L. Tong, *et al.*, "Conversational agents in healthcare: A systematic review," *Journal of the American Medical Informatics Association*, vol. 25, no. 9, pp. 1248–1258, 2018.

[67] A. Leibing, "Aging in contemporary canada," *The Canadian Review of Sociology*, vol. 40, no. 4, p. 489, 2003.

[68] C. LHIN, *ntegrated health service plan 2013–2016*. 2012.

[69] Y. Li and T. Yang, "Word embedding for understanding natural language: A survey," in *Guide to big data applications*, Springer, 2018, pp. 83–104.

[70] D. Lieberman and S. Massey, "A technological approach to reaching a hidden population of problem drinkers," *Psychiatric services (Washington, D.C.)*, vol. 59, pp. 297–303, Apr. 2008. DOI: 10.1176/appi.ps.59.3.297.

[71] S. Linke, E. Murray, C. Butler, and P. Wallace, "Internet-based interactive health intervention for the promotion of sensible drinking: Patterns of use and potential impact on members of the general public," *Journal of medical Internet research*, vol. 9, e10, Feb. 2007. DOI: 10.2196/jmir.9.2.e10.

[72] P. Lison and J. Tiedemann, "Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles," 2016.

[73] Y. Liu, M. Ott, N. Goyal, *et al.*, *Roberta: A robustly optimized bert pretraining approach*, 2019. DOI: 10.48550/ARXIV.1907.11692. [Online]. Available: https://arxiv.org/abs/1907.11692.

[74] N. Love, "The linguistic thought of jr firth," *Stellenbosch Papers in Linguistics*, vol. 15, no. 1, pp. 31–60, 1986.

[75] R. Lowe, N. Pow, I. Serban, and J. Pineau, "The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems," *arXiv preprint arXiv:1506.08909*, 2015.

[76] G. Lucas, J. Gratch, A. King, and L.-P. Morency, "It's only a computer: Virtual humans increase willingness to disclose," *Computers in Human Behavior*, vol. 37, pp. 94–100, Aug. 2014. DOI: 10.1016/j.chb.2014.04.043.

[77] D. Lukovnikov, A. Fischer, and J. Lehmann, *Pretrained transformers for simple question answering over knowledge graphs*, 2020. DOI: 10.48550/ARXIV.2001.11985. [Online]. Available: https://arxiv.org/abs/2001.11985.

[78] D. M, *How Canada is encouraging self-isolation to prevent the spread of COVID-19*. World Health Organization, 2020. [Online]. Available: https://globalnews.ca/news/6634584/canada-coronavirus-covid-19-self-isolation.

[79] M. F. L. Macadam, "Moving toward health service integration: Provincial progress in system change for seniors," 2009.

[80] V. Marivate and T. Sefara, "Improving short text classification through global augmentation methods," in *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, Springer, 2020, pp. 385–399.

[81] J. McAuley and A. Yang, "Addressing complex and subjective product-related queries with customer reviews," in *Proceedings of the 25th International Conference on World Wide Web*, 2016, pp. 625–635.

[82] "Mental illness will be 'next wave' of covid-19 pandemic, epidemiologist says," 2020. [Online]. Available: `https://www.cbc.ca/news/canada/british-columbia/months-%20isolation-mental-health-covid-1.5521649`.

[83] T. Menzies, A. Dekhtyar, J. Distefano, and J. Greenwald, "Problems with precision: A response to "comments on 'data mining static code attributes to learn defect predictors'"," *IEEE Transactions on Software Engineering*, vol. 33, no. 9, pp. 637–640, 2007. DOI: `10.1109/TSE.2007.70721`.

[84] G. F. MILLER, R. Beckwith, and C. Fellbaum, "C., gross, d., and miller, k. 1990. introduction to wordnet: An on-line lexical database," *J. Lexicog*, vol. 3, no. 4, pp. 234–244, 1993.

[85] "Months of abuse, exhaustion have burnt-out nurses leaving their jobs," [Online]. Available: `https://www.cbc.ca/player/play/1945590851847`..

[86] "Mood disorders society of canada, "quick facts: Mental illness addiction," mdsc, 2019. [Online]. Available: `https://mdsc.ca/docs/MDSC_Quick_Facts_4th_Edition_EN.pdf`.

[87] C. Moreno, T. Wykes, S. Galderisi, *et al.*, "How mental health care should change as a consequence of the covid-19 pandemic," *The Lancet Psychiatry*, vol. 7, no. 9, pp. 813–824, 2020.

[88] C. M. Morin, M. LeBlanc, M. Daley, J. Gregoire, and C. Merette, "Epidemiology of insomnia: Prevalence, self-help treatments, consultations, and determinants of help-seeking behaviors," *Sleep medicine*, vol. 7, no. 2, pp. 123–130, 2006.

[89] R. Mossabir, R. Morris, A. Kennedy, C. Blickem, and A. Rogers, "A scoping review to understand the effectiveness of linking schemes from healthcare providers to community resources to improve the health and well-being of people with long-term conditions," *Health Social Care in the Community*, vol. 23, Dec. 2014. DOI: `10.1111/hsc.12176`.

[90] W. Narrow, D. Regier, D. Rae, R. Manderscheid, and B. Locke, "Use of services by persons with mental and addictive disorders: Findings from the national institute of mental health epidemiologic catchment area program," *Archives of general psychiatry*, vol. 50, pp. 95–107, Mar. 1993.

[91] U. Naseem, I. Razzak, K. Musial, and M. Imran, "Transformer based deep intelligent contextual embedding for twitter sentiment analysis," *Future Generation Computer Systems*, vol. 113, pp. 58–69, 2020.

[92] "National institute on alcohol abuse and alcoholism, understanding the impact of alcohol on human health and well-being, bethesda: National institute on alcohol abuse and alcoholis," 2014.

[93] W. S. Noble, "What is a support vector machine?" *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.

[94] A. Nuez Ezquerra, "Implementing chatbots using neural machine translation techniques," B.S. thesis, Universitat Politècnica de Catalunya, 2018.

[95] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[96] M. Paul McGinnis, M. M. Davis, M. DeSordi, and M. Thomas, "Integrating primary care practices and community-based resources to manage obesity," 2014.

[97] H. V. Peachey D and A. O, "An imperative for change: Access to psychological services for canada. report to the canadian psychological association," *Canadian Psychological Association*, 2013.

[98] J. T. Pearson C and A. J, "Mental and substance use disorders in canada," statistics canada," 2015. [Online]. Available: `https://www150.statcan.gc.ca/n1/pub/82-624-x/2013001/article/11855-eng.htm`.

[99] A. Perevalov, D. Kurushin, R. Faizrakhmanov, and F. Khabibrakhmanova, "Question embeddings based on shannon entropy: Solving intent classification task in goal-oriented dialogue system," *arXiv preprint arXiv:1904.00785*, 2019.

[100] F. Peters *et al.*, "Master thesis: Design and implementation of a chatbot in the context of customer support," 2018.

[101] J. Ploeg, M. Denton, J. Tindale, *et al.*, "Older adults' awareness of community health and support services for dementia care," *Canadian Journal on Aging/La Revue canadienne du vieillissement*, vol. 28, no. 4, pp. 359–370, 2009.

[102] "Position paper on the occasion of the 10th anniversary of the 2004 10-year plan to strengthen health care in canada," 2014. [Online]. Available: `https://www.waittimealliance.ca/wp-content/uploads/2014/09/WTA-Fall-Event-2014-Position-Paper-English-FINAL.pdf`.

[103] S. Provoost, H. M. Lau, J. Ruwaard, H. Riper, *et al.*, "Embodied conversational agents in clinical psychology: A scoping review," *Journal of medical Internet research*, vol. 19, no. 5, e6553, 2017.

[104] "Psychology works fact sheet: Depression," 2017. [Online]. Available: https://cpa.ca/docs/File/Publications/FactSheets/PsychologyWorksFactSheet_Depression.pdf.

[105] W. Qader, M. M. Ameen, and B. Ahmed, "An overview of bag of words;importance, implementation, applications, and challenges," Jun. 2019, pp. 200–204. DOI: 10.1109/IEC47844.2019.8950616.

[106] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[107] F. Radlinski, K. Balog, B. Byrne, and K. Krishnamoorthi, "Coached conversational preference elicitation: A case study in understanding movie preferences," 2019.

[108] R. Raine, "Making a clever intelligent agent: The theory behind the implementation," in *2009 IEEE International Conference on Intelligent Computing and Intelligent Systems*, IEEE, vol. 3, 2009, pp. 398–402.

[109] K. Ramesh, S. Ravishankaran, A. Joshi, and K. Chandrasekaran, "A survey of design techniques for conversational agents," in *International conference on information, communication and computing technology*, Springer, 2017, pp. 336–350.

[110] C. on the Reform of Ontario's Public Services, D. Drummond, and Ontario, *Public Services for Ontarians: A Path to Sustainability and Excellence : Executive Summary*. Queen's Printer for Ontario, 2012, ISBN: 9781443589055. [Online]. Available: https://books.google.ca/books?id=dxb3wQEACAAJ.

[111] A. Reid, *Worry, Gratitude Boredom: As COVID-19 affects mental, financial health, who fares better? Who is worse?* World Health Organization, 2020. [Online]. Available: https://angusreid.org/covid19-mental-health/.

[112] J. D. Rennie, L. Shih, J. Teevan, and D. R. Karger, "Tackling the poor assumptions of naive bayes text classifiers," in *Proceedings of the 20th international conference on machine learning (ICML-03)*, 2003, pp. 616–623.

[113] R. Rifkin and A. Klautau, "In defense of one-vs-all classification," *J. Mach. Learn. Res.*, vol. 5, pp. 101–141, Dec. 2004, ISSN: 1532-4435.

[114] A. Sahebi, B. Nejati-Zarnaqi, S. Moayedi, K. Yousefi, M. Torres, and M. Golitaleb, "The prevalence of anxiety and depression among healthcare workers during the covid-19 pandemic: An umbrella review of meta-analyses," *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, vol. 107, p. 110 247, 2021.

[115] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," 2014.

[116] A. E. Samy, S. R. El-Beltagy, and E. Hassanien, "A context integrated model for multi-label emotion detection," *Procedia computer science*, vol. 142, pp. 61–71, 2018.

[117] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, *Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter*, 2019. DOI: 10.48550/ARXIV.1910.01108. [Online]. Available: https://arxiv.org/abs/1910.01108.

[118] C. Sanmartin, C. Houle, S. Tremblay, and J.-M. Berthelot, "Changes in unmet health care needs," *Health Rep*, vol. 13, no. 3, pp. 15–21, 2002.

[119] R. C. Schank, "Conceptual dependency: A theory of natural language understanding," *Cognitive Psychology*, vol. 3, no. 4, pp. 552–631, 1972, ISSN: 0010-0285. DOI: https://doi.org/10.1016/0010-0285(72)90022-9. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0010028572900229.

[120] J. Schmidhuber, "Learning complex, extended sequences using the principle of history compression," *Neural Computation*, vol. 4, no. 2, pp. 234–242, 1992.

[121] F. Sebastiani, "Machine learning in automated text categorization," *ACM computing surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, 2002.

[122] T. Shangipour ataei, S. Javdan, and B. Minaei-Bidgoli, "Applying transformers and aspect-based sentiment analysis approaches on sarcasm detection," in *Proceedings of the Second Workshop on Figurative Language Processing*, Online: Association for Computational Linguistics, Jul. 2020, pp. 67–71. DOI: 10.18653/v1/2020.figlang-1.9. [Online]. Available: https://aclanthology.org/2020.figlang-1.9.

[123] T. Shao, Y. Guo, H. Chen, and Z. Hao, "Transformer-based neural network for answer selection in question answering," *IEEE Access*, vol. 7, pp. 26 146–26 156, 2019.

[124] P. Shaw, J. Uszkoreit, and A. Vaswani, *Self-attention with relative position representations*, 2018. DOI: 10.48550/ARXIV.1803.02155. [Online]. Available: https://arxiv.org/abs/1803.02155.

[125] B. A. Shawar and E. Atwell, "Chatbots: Are they really useful?" *LDV Forum*, vol. 22, pp. 29–49, 2007.

[126] C. A. Simpson and J. A. Tucker, "Temporal sequencing of alcohol-related problems, problem recognition, and help-seeking episodes," *Addictive Behaviors*, vol. 27, no. 5, pp. 659–674, 2002.

[127] S. Singh and H. Beniwal, "A survey on near-human conversational agents," *Journal of King Saud University - Computer and Information Sciences*, 2021, ISSN: 1319-1578. DOI: https://doi.org/10.1016/j.jksuci.2021.10.013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1319157821003001.

[128] S. L. Smith, P.-J. Kindermans, C. Ying, and Q. V. Le, *Don't decay the learning rate, increase the batch size*, 2017. DOI: 10.48550/ARXIV.1711.00489. [Online]. Available: https://arxiv.org/abs/1711.00489.

[129] "Statistics canada, health fact sheets: Mental health care needs," 2018, Stats Can, 2019. [Online]. Available: https://www150.statcan.gc.ca/n1/pub/82-625-x/2019001/article/00011-eng.htm.

[130] G. Strudwick, S. Sockalingam, I. Kassam, *et al.*, "Digital interventions to support population mental health in canada during the covid-19 pandemic: Rapid review," *JMIR mental health*, vol. 8, no. 3, e26550, 2021.

[131] A. Sunderland and L. C. Findlay, "Perceived need for mental health care in canada: Results from the 2012 canadian community health survey-mental health.," *Health reports*, vol. 24 9, pp. 3–9, 2013.

[132] P. Suta, X. Lan, B. Wu, P. Mongkolnam, and J. Chan, "An overview of machine learning in chatbots," *International Journal of Mechanical Engineering and Robotics Research*, vol. 9, no. 4, pp. 502–510, 2020.

[133] S. Suthaharan, "Support vector machine," in *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*. Boston, MA: Springer US, 2016, pp. 207–235, ISBN: 978-1-4899-7641-3. DOI: 10.1007/978-1-4899-7641-3_9. [Online]. Available: https://doi.org/10.1007/978-1-4899-7641-3_9.

[134] D. A. Sutton, H. Moldofsky, and E. M. Badley, "Insomnia and health problems in canadians," *Sleep*, vol. 24, no. 6, pp. 665–670, 2001.

[135] W. L. Taylor, ""cloze procedure": A new tool for measuring readability," *Journalism Quarterly*, vol. 30, no. 4, pp. 415–433, 1953. DOI: 10.1177/107769905303000401. eprint: https://doi.org/10.1177/107769905303000401. [Online]. Available: https://doi.org/10.1177/107769905303000401.

[136] I. Tenney, D. Das, and E. Pavlick, "Bert rediscovers the classical nlp pipeline," 2019. DOI: 10.48550/ARXIV.1905.05950. [Online]. Available: https://arxiv.org/abs/1905.05950.

[137] A. N. Vaidyam, H. Wisniewski, J. D. Halamka, M. S. Kashavan, and J. B. Torous, "Chatbots and conversational agents in mental health: A review of the psychiatric landscape," *The Canadian Journal of Psychiatry*, vol. 64, no. 7, pp. 456–464, 2019.

[138] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[139] V. Vlasov, J. E. M. Mosig, and A. Nichol, *Dialogue transformers*, 2019. DOI: 10.48550/ARXIV.1910.00486. [Online]. Available: https://arxiv.org/abs/1910.00486.

[140] M. Wahde and M. Virgolin, *Conversational Agents: Theory and Applications*. Feb. 2022.

[141] E. R. Walker, R. E. McGee, and B. G. Druss, "Mortality in mental disorders and global disease burden implications: A systematic review and meta-analysis," *JAMA psychiatry*, vol. 72, no. 4, pp. 334–341, 2015.

[142] R. S. Wallace, "The anatomy of a.l.i.c.e.," in *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*, R. Epstein, G. Roberts, and G. Beber, Eds. Dordrecht: Springer Netherlands, 2009, pp. 181–210, ISBN: 978-1-4020-6710-5. DOI: 10.1007/978-1-4020-6710-5_13. [Online]. Available: https://doi.org/10.1007/978-1-4020-6710-5_13.

[143] M. Wan and J. McAuley, "Modeling ambiguity, subjectivity, and diverging viewpoints in opinion question answering systems," in *2016 IEEE 16th international conference on data mining (ICDM)*, IEEE, 2016, pp. 489–498.

[144] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, *Glue: A multi-task benchmark and analysis platform for natural language understanding*, 2018. DOI: 10.48550/ARXIV.1804.07461. [Online]. Available: https://arxiv.org/abs/1804.07461.

[145] Q. Wang, B. Li, T. Xiao, *et al.*, *Learning deep transformer models for machine translation*, 2019. DOI: 10.48550/ARXIV.1906.01787. [Online]. Available: https://arxiv.org/abs/1906.01787.

[146] A. Watson, T. Bickmore, A. Cange, A. Kulshreshtha, J. Kvedar, *et al.*, "An internet-based virtual coach to promote physical activity adherence in overweight adults: Randomized controlled trial," *Journal of medical Internet research*, vol. 14, no. 1, e1629, 2012.

[147] J. Wei and K. Zou, *Eda: Easy data augmentation techniques for boosting performance on text classification tasks*, 2019. DOI: 10.48550/ARXIV.1901.11196. [Online]. Available: https://arxiv.org/abs/1901.11196.

[148] J. Weizenbaum, "Eliza—a computer program for the study of natural language communication between man and machine," *Commun. ACM*, vol. 9, no. 1, pp. 36–45, Jan. 1966, ISSN: 0001-0782. DOI: 10.1145/365153.365168. [Online]. Available: https://doi.org/10.1145/365153.365168.

[149] ——, "Eliza—a computer program for the study of natural language communication between man and machine," *Commun. ACM*, vol. 9, no. 1, pp. 36–45, Jan. 1966, ISSN: 0001-0782. DOI: 10.1145/365153.365168. [Online]. Available: https://doi.org/10.1145/365153.365168.

[150] *WHO Coronavirus (COVID-19) Dashboard*. World Health Organization. [Online]. Available: https://covid19.who.int.

[151] T. C. Wild, A. B. Roberts, and E. L. Cooper, "Compulsory substance abuse treatment: An overview of recent findings and issues," *European Addiction Research*, vol. 8, no. 2, pp. 84–93, 2002.

[152] L. Wu, A. Fisch, S. Chopra, K. Adams, A. Bordes, and J. Weston, *Starspace: Embed all the things!* 2017. DOI: 10.48550/ARXIV.1709.03856. [Online]. Available: https://arxiv.org/abs/1709.03856.

[153] Z. Xu, C. Sun, Y. Long, *et al.*, "Dynamic working memory for context-aware response generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 9, pp. 1419–1431, 2019.

[154] R. Yan, "" chitty-chitty-chat bot": Deep learning for conversational ai.," in *IJCAI*, vol. 18, 2018, pp. 5520–5526.

[155] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, *Xlnet: Generalized autoregressive pretraining for language understanding*, 2019. DOI: 10.48550/ARXIV.1906.08237. [Online]. Available: https://arxiv.org/abs/1906.08237.

[156] R. Yoshihashi, W. Shao, R. Kawakami, S. You, M. Iida, and T. Naemura, *Classification-reconstruction learning for open-set recognition*, 2018. DOI: 10.48550/ARXIV.1812.04246. [Online]. Available: https://arxiv.org/abs/1812.04246.

[157] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *ieee Computational intelligenCe magazine*, vol. 13, no. 3, pp. 55–75, 2018.

[158] J. Zhang, H. Luan, M. Sun, *et al.*, *Improving the transformer translation model with document-level context*, 2018. DOI: 10.48550/ARXIV.1810.03581. [Online]. Available: https://arxiv.org/abs/1810.03581.

Appendix A

A.1 List of Defined Intents in the Mira Chatbot

1. need_definition
2. need_screening
3. share_diagnosis
4. need_domestic_abuse_support
5. need_medication
6. need_program_or_service_in_person
7. suicidal_self
8. suicidal_other
9. need_support_family_or_friend
10. need_program_or_services_COVID_specific
11. need_psychiatrist
12. need_counsellor_psychotherapist
13. need_psychologist
14. need_healer
15. follow_up_survey

16. copy_of_transcript
17. survey_and_transcript
18. need_housing
19. need_abuse_online_chat
20. need_abuse_financial_support
21. need_abuse_rights
22. need_addiction_substance_use_programs
23. need_human_backup
24. feelings_symptoms
25. paid_and_free
26. see_mira_portal
27. need_specialist
28. need_online_chat_and_text
29. need_crisis_distress_support
30. need_group_class
31. need_symptom_list
32. need_program_services
33. greet
34. goodbye
35. affirm
36. deny
37. for_someone_else

38. help_from_another_person
39. online_courses_and_webinar
40. access_consent_info
41. consent_agree
42. see_consent_form_again
43. have_question_for_study_team
44. continue
45. where_live
46. live_somewhere
47. prefer_not_say
48. for_me
49. book_and_pamplet
50. paid_resources
51. need_information
52. suspect_diagnosis
53. need_prevalence
54. need_causes
55. need_coping_skills
56. need_peer_support
57. need_doctor
58. need_types_of_disorders
59. need_comparison

60. need_info_comorbid
61. need_treatment_info
62. need_text_to_voice
63. question_voice_to_text
64. withdraw_consent
65. end_conversation
66. employment
67. military/veteran
68. healthcare_worker
69. need_in_person
70. eliza
71. something_else
72. about_mira
73. need_virtual
74. confused
75. satisfied
76. need_info_other_province
77. need_program_services
78. mood_great
79. family_member
80. other
81. employer_resources

82. female_resources
83. male_resources
84. lgbtq2s_resources
85. new_canadian
86. age
87. free_resources
88. language_of_resource
89. show_resource
90. out_of_scope

A.2 List of Utilized Data Augmentation Techniques

- Substitute character by pre-defined a Optical Character Recognition (OCR) [61] error
- Substitute character by keyboard distance
- Insert character randomly
- Substitute character randomly
- Swap characters randomly
- Delete characters randomly Substitute words by spelling mistakes words dictionary
- Substitute words by WordNet's synonym
- Substitute word by contextual word embeddings (DistilBERT [117])
- Substitute word by contextual word embeddings (RoBERTA [73]) Delete words randomly

- Swap words randomly Split word to two tokens randomly
- Insert word by contextual word embeddings (RoBERTA)
- Insert word by contextual word embeddings (Distilbert)
- Insert word by contextual word embeddings (BERT)
- Insert word randomly by word embeddings similarity
- Insert sentence by contextual word embeddings (XLNet [155])
- Insert sentence by contextual word embeddings (GPT-2 [17])
- Back translation augmenter
- Substitute word by contextual word embeddings (BERT)

A.3 Mira Tree

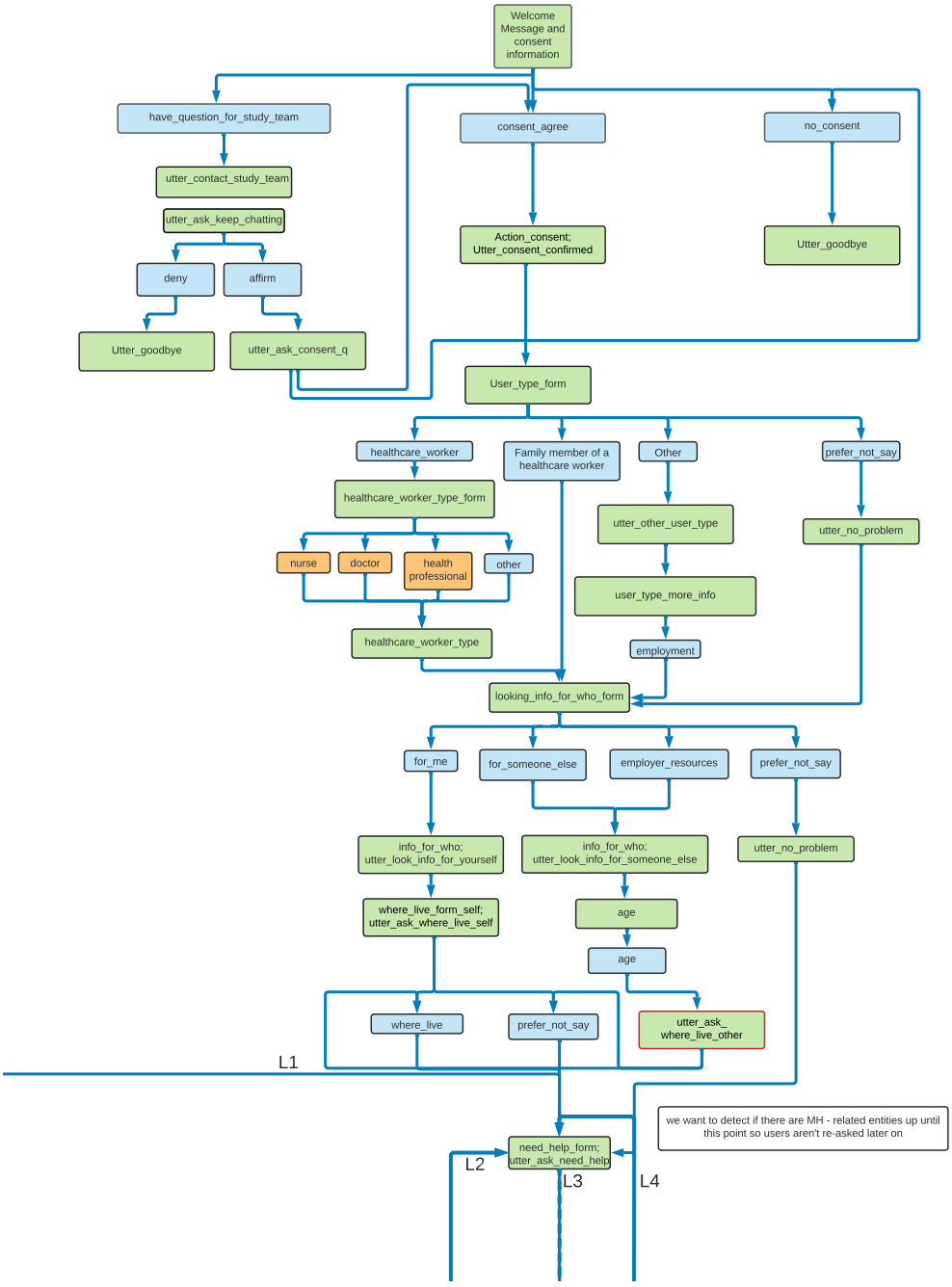

Figure A.1: MIRA decision tree (part 1).

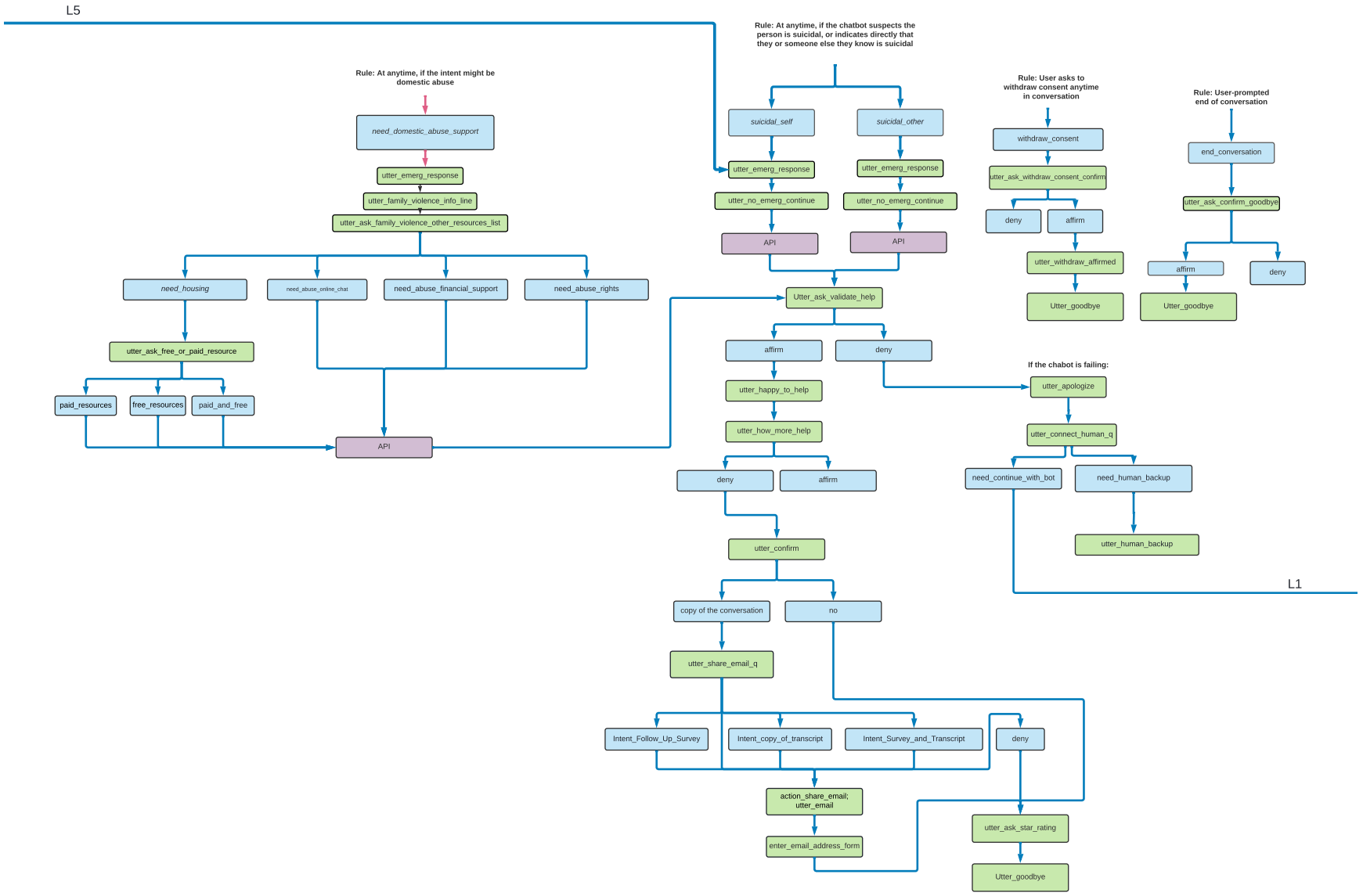

Figure A.2: MIRA decision tree (part 2).

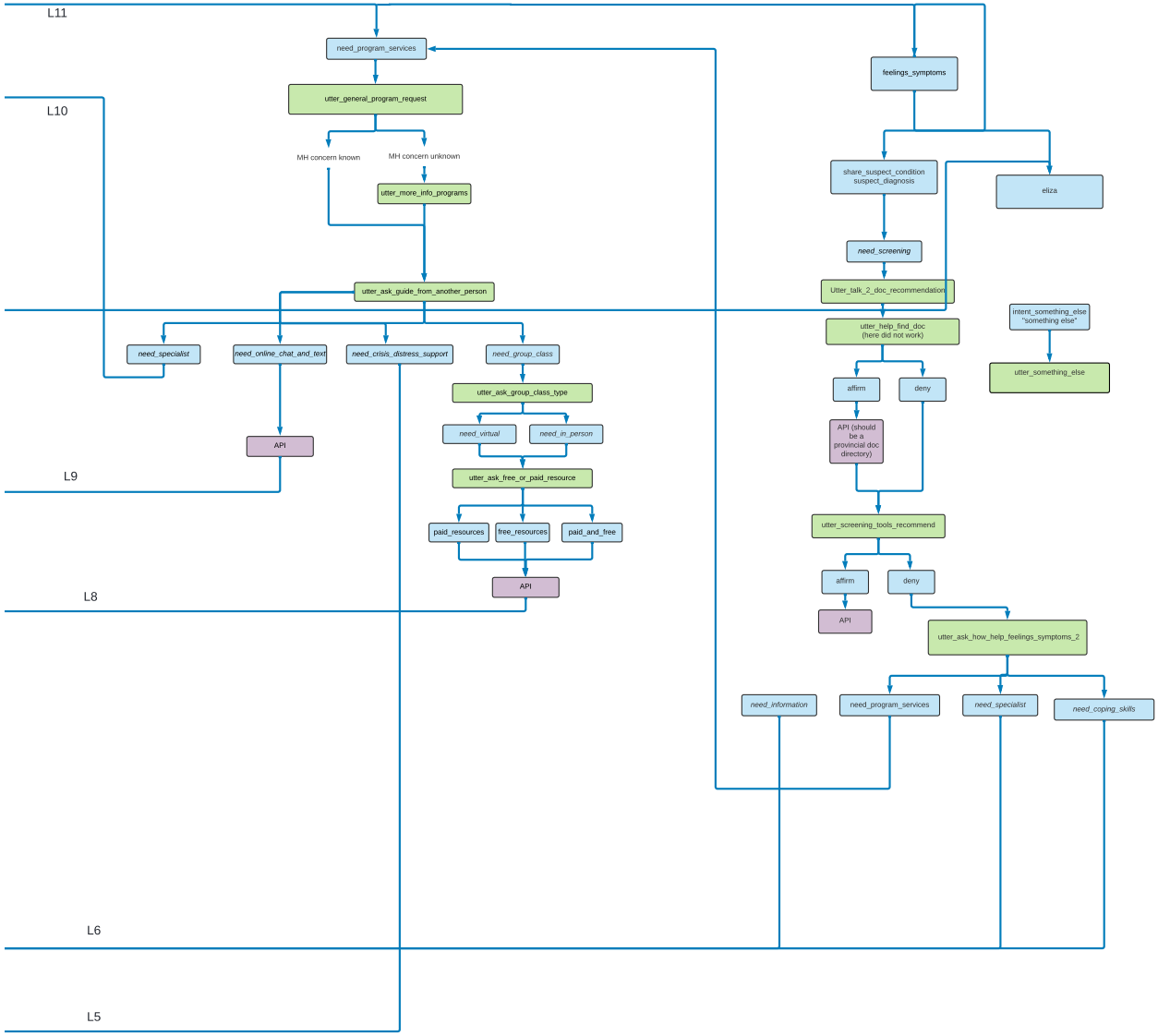

Figure A.3: MIRA decision tree (part 3).

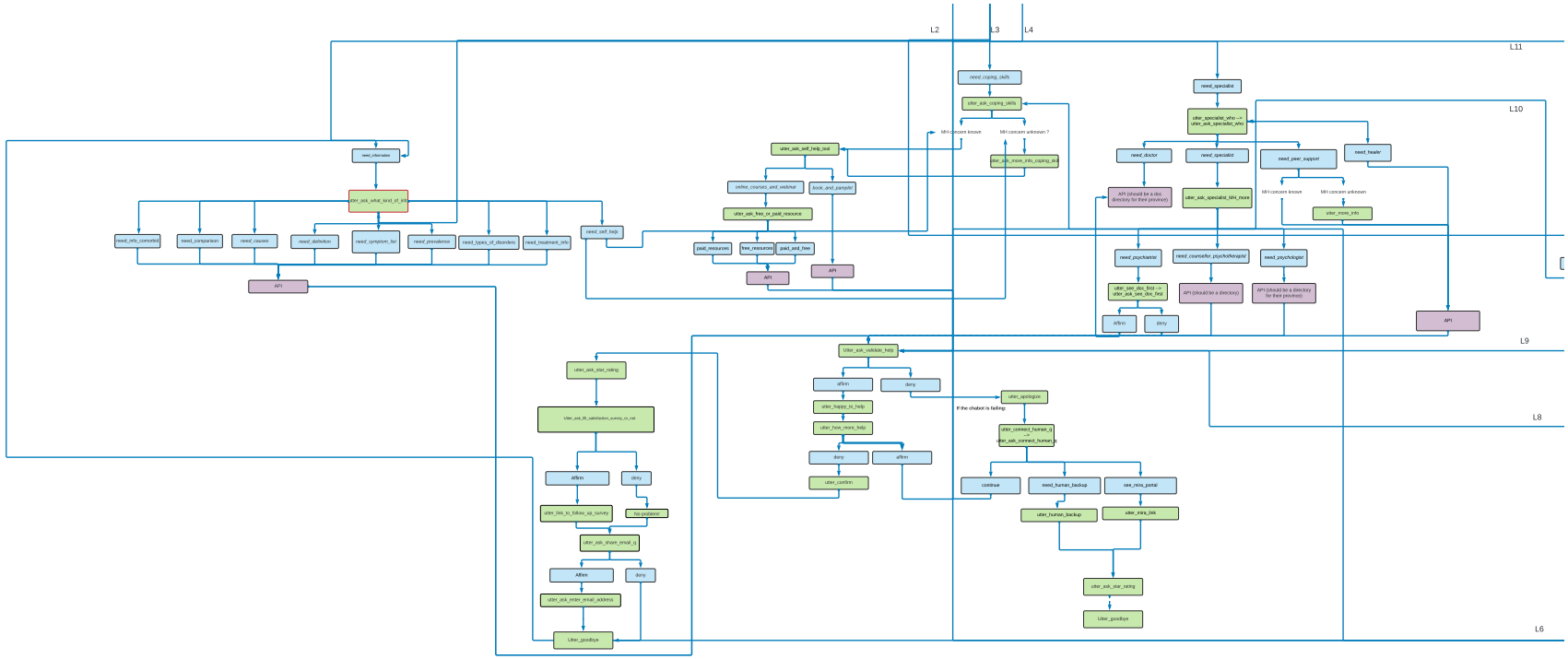

Figure A.4: MIRA decision tree (part 4).