Czech Technical University in Prague
Faculty of Electrical Engineering

Department of Computer Science
Study Programme: Open Informatics

Processing of dialogue data from the field of addiction counselling practice

MASTER'S THESIS

Author: Bc. Patrik Jankuv
Supervisor: doc. Ing. Daniel Novák, Ph.D.
Year: 2023

MASTER‘S THESIS ASSIGNMENT

I. Personal and study details

Student's name: **Jankuv Patrik** Personal ID number: **483838**

Faculty / Institute: **Faculty of Electrical Engineering**

Department / Institute: **Department of Computer Science**

Study program: **Open Informatics**

Specialisation: **Software Engineering**

II. Master’s thesis details

Master’s thesis title in English:

Processing of dialogue data from the field of addiction counselling practice

Master’s thesis title in Czech:

Zpracování dialogových dat z oblasti adiktologické poradenské praxe

Guidelines:

1) Convert audio recordings of psychologist-client consultations into text form. Perform review of the available systems, compare their quality and select a suitable candidate.
2) Anonymise the data obtained, taking into account GDPR regulations.
3) Pre-process the data for the language conversation system.
The aim is to design a support system for social service workers to make their work more efficient when chatting online with clients.

Bibliography / sources:

1. Miovský,M., Čablová,L., & Jurystová,L. (2015).Časná diagnostik aa krátké intervencev adiktologii.In K. Kalina(Ed.),Klinickáadiktologie(286–293).Praha:GradaPublishing.
2. H. Brendryen, P. Kraft, and H. Schaalma, “Looking Inside the Black Box: Using Intervention Mapping to Describe the Development of the Automated Smoking Cessation Intervention ‘Happy Ending’,” The Journal of Smoking Cessation, vol. 5, no. 1, pp. 29–56, Jun. 2010.
3. Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Name and workplace of master’s thesis supervisor:

doc. Ing. Daniel Novák, Ph.D. Analysis and Interpretation of Biomedical Data FEE

Name and workplace of second master’s thesis supervisor or consultant:

Date of master’s thesis assignment: **22.09.2023** Deadline for master's thesis submission: **09.01.2024**

Assignment valid until: **16.02.2025**

doc. Ing. Daniel Novák, Ph.D. Supervisor’s signature	Head of department’s signature	prof. Mgr. Petr Páta, Ph.D. Dean’s signature

III. Assignment receipt

The student acknowledges that the master’s thesis is an individual work. The student must produce his thesis without the assistance of others, with the exception of provided consultations. Within the master’s thesis, the author must state the names of consultants and include a list of references.

. | Date of assignment receipt | Student’s signature

Author declaration

I declare that the presented work was developed independently and that I have listed all sources of information used within it in accordance with the methodical instructions for observing the ethical principles in the preparation of university theses.

Prague the .....

.....
Bc. Patrik Jankuv

Acknowledgment

I would like to thank doc. Ing. Daniel Novák, Ph.D., for the guidance of my thesis and for the suggestive suggestions that enriched it. I would also like to thank my colleagues from the project of which this thesis is a part, the following: Klára Losenická, Cheng Kang, MSc., Fabián Bodnár, Ing. Jindřich Prokop and Bc. Štěpán Bořek. Of course, I want to thank my family and friends for their support throughout my studies.

Bc. Patrik Jankuv

Title:
Processing of dialogue data from the field of addiction counselling practice

Author: Bc. Patrik Jankuv

Supervisor: doc. Ing. Daniel Novák, Ph.D.
Analysis and Interpretation of Biomedical Data

Abstract: The master's thesis focuses on the processing of dialogue data from the field of addiction counselling practice. The project's primary objectives are to convert audio recordings of addiction specialist-client consultations into text form, anonymize the data in compliance with GDPR regulations, and preprocess the converted data for integration into a language conversation system. The ultimate goal is to develop a robust and adaptable support system tailored for workers engaged in online consultations with clients, aiming to streamline retrieving and utilizing insights from past consultations and facilitate real-time assistance. The thesis delves into the exploration of available systems for speech-to-text conversion, assessment of their quality, and meticulous selection of an optimal solution. It also addresses the intricacies of anonymizing sensitive data and ensuring its readiness for advanced language processing algorithms. The thesis discusses the theoretical underpinnings of data anonymization and legal and ethical considerations and proposes a methodological framework for the anonymization process.
Key words: Natural language processing, Speech to text, Anonymization

Název práce:
Zpracování dialogových dat z oblasti adiktologické poradenské praxe

Abstrakt: Magisterská práce se zaměřuje na zpracování dat z dialogů z oblasti poradenské praxe v oblasti závislostí. Hlavním cílem projektu je převést zvukové záznamy konzultací specialistů na závislosti s klienty do textové podoby, anonymizovat data v souladu s předpisy GDPR a předzpracovat převedená data pro integraci do systému jazykové konverzace. Konečným cílem je vyvinout robustní a přizpůsobivý podpůrný systém přizpůsobený pracovníkům zapojeným do online konzultací s klienty, jehož cílem je zefektivnit získávání a využívání poznatků z minulých konzultací a usnadnit pomoc v reálném čase. Práce se zabývá průzkumem dostupných systémů pro převod řeči na text, posouzením jejich kvality a pečlivým výběrem optimálního řešení. Zabývá se také složitostmi anonymizace citlivých dat a zajištěním jejich připravenosti pro pokročilé algoritmy zpracování jazyka. Práce se zabývá teoretickými základy anonymizace dat a právními a etickými aspekty a navrhuje metodický rámec procesu anonymizace.
Klíčová slova: Zpracování přirozeného jazyka, Rozpoznávání řeči, Anonymizace

Contents

List of abbreviations

ANN	Artificial Neural Network
API	Application Programming Interface
BERT	Bidirectional Encoder Representations from Transformers
CNN	Convolutional Neural Network
CER	Character Error Rate
DL	Deep Learning
DNN	Deep Neural Network
EU	European Union
GDPR	General Data Protection Regulation
GPU	Graphics Processing Unit
GPT	Generative Pre-trained Transformer
HMM	Hidden Markov Model
LLM	Large Language Model
LSTM	Long Short-Term Memory
ML	Machine Learning
NER	Named Entity Recognition
NLPO	Narodni Linka Pro Odvikani
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
RAG	Retrieval-Augmented Generation
RNN	Recurrent Neural Network
SER	Sentence Error Rate
STT	Speech to Text
WAV	Waveform Audio File Format
WER	Word Error Rate

List of Figures

Introduction

The field of addiction services and support has seen significant advancements in methodology and treatment approaches. The National Quitline, a pivotal resource in the Czech Republic, stands as a testament to this progress. Over the past years, its collaboration with professionally trained psychologists and addiction specialists has facilitated the rehabilitation and prevention of various addictive behaviors, assisting over 6,000 clients and yielding commendable outcomes [1].

Amidst this success, the transformation of audio records of psychologist-client consultations into a more accessible format presents a crucial opportunity. The collaboration between the Czech Technical University and The National Quitline has unveiled an invaluable reservoir of information within the 8000 recorded calls between 2020 and 2022. These conversations, encapsulating rich insights and diverse experiences related to addiction, hold the potential to not only enhance the existing methodology but also pave the way for more efficient and effective support systems.

The goal of this master's thesis is to address this challenge by employing cutting-edge techniques in natural language processing and conversational data processing. The primary objectives outlined for this project are threefold: first, to convert the audio recordings into text form, enabling comprehensive analysis and utilization of the contained information; second, to anonymize the data in compliance with GDPR regulations, safeguarding the privacy and confidentiality of the individuals involved; and third, to preprocess the converted data, preparing it for integration into a language conversation system.

The ultimate aim is to develop a robust and adaptable support system tailored for social service workers engaged in online consultations with clients. This system will not only streamline the retrieval and utilization of insights from past consultations but also facilitate real-time assistance, thereby bolstering the efficiency and efficacy of addiction counseling services.

This thesis embarks on a journey to explore the available systems for audio-to-text conversion, assess their quality, and meticulously select an optimal solution. Additionally, it will delve into the intricacies of anonymizing sensitive data while ensuring its readiness for advanced language processing algorithms.

Chapter 1

Speech to Text

1.1 Speech to Text

Speech to text is a technology that converts spoken words into text[2]. It is a complex technology that has been around for many years, but it has only recently become practical for everyday use. STT[1] is used in various applications, including dictation software, voice-to-text messaging, transcription services, virtual assistants, smart speakers, etc.

1.2 Techniques for STT

1.2.1 Hidden Markov models

Hidden Markov models have emerged as a preferred approach in modeling stochastic processes and sequences across various applications, including speech and handwriting recognition [3][4], computational molecular biology [5], and natural language modeling [6][7]. An important work in the utilization of HMM within the domain of speech recognition was achieved through the work of Lawrence in 1989 [8]. Another significant contribution in the realm of speech recognition using HMM was presented by Gales, Mark, Young, Steve, and other researchers in their 2008 work titled "The Application of Hidden Markov Models in Speech Recognition."[9].

An HMM[2] represents a stochastic process where the concealed underlying stochastic process becomes indirectly ascertainable through the analysis of observed symbol sequences from a distinct set of stochastic processes. These models encapsulate hidden states delineating unobservable or latent attributes within the modeled process. Widely employed in analyzing usage patterns, activity profiles, and transitions between process states, HMM-based methodologies facilitate the prediction of the most probable sequence of states. Functioning as a stochastic model for discrete events,

[1] Speech to Text
[2] Hidden Markov Model

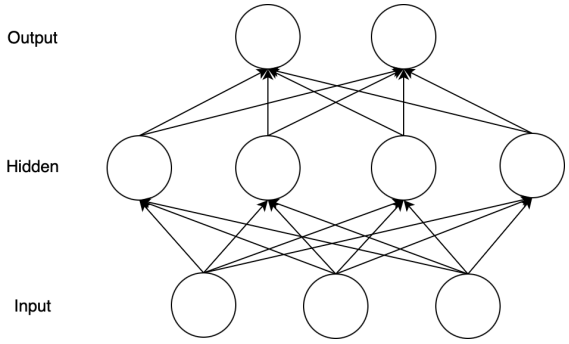


Figure 1.1: An artificial neural network architecture. The arrows represent the connections between the layers, and the text labels indicate the type of layer [21].

HMM operates as a variant of the Markov chain, governing state transitions solely based on the current state of the system. Notably, the states within an HMM remain hidden or can only be inferred from the observed symbols [10].

1.2.2 Artificial Neural Networks

In the late 1980s, neural networks emerged as an appealing approach for acoustic modeling in speech recognition [11]. Since their introduction, neural networks have found extensive application across various facets of speech recognition, encompassing phoneme classification using multi-objective evolutionary algorithms, isolated word recognition [12], audiovisual speech recognition, audiovisual speaker recognition, and speaker adaptation [13].

Unlike HMMs, neural networks operate with fewer explicit assumptions regarding the statistical properties of features, rendering them an attractive model for speech recognition. Their capacity to estimate probabilities for speech feature segments enables discriminative training efficiently and naturally. However, while early neural networks demonstrated efficacy in classifying short-time units like phonemes and isolated words, their limited capability in modeling temporal dependencies hindered their success in continuous recognition tasks [14].

Addressing this limitation, one approach involved utilizing neural networks as a preprocessing step, facilitating feature transformation or dimensionality reduction before HMM-based recognition [15]. Nevertheless, recent advancements in architectures like Long Short-Term Memory and related Recurrent Neural Networks [16][17][18], Time Delay Neural Networks [19], and transformers [20] have showcased notable improvements in tackling this challenge. These innovations have demonstrated enhanced performance in handling temporal dependencies, marking significant progress in the domain of continuous recognition tasks.

Neural networks represent a subset of machine learning models inspired by the organizational principles found in biological neural networks present in animal brains. ANN[3] consist of interconnected artificial neurons, resembling the neurons in the brain, where signals pass through connections known as edges [22]. These artificial

[3]Artificial Neural Network

neurons process received signals using non-linear functions and transmit outputs to connected neurons. The connections between neurons, akin to synapses, possess adjustable weights that modify signal strength. Neurons within layers perform various transformations on inputs, propagating signals from the input layer to the output layer, possibly traversing multiple layers. When an ANN comprises a minimum of 2 hidden layers, it is termed a deep neural network [21].

1.2.3 Deep Neural Network

A deep neural network represents an artificial neural network structure featuring multiple hidden layers positioned between the input and output layers [23]. Comparable to shallow neural networks, DNN[4]s possess the capability to model intricate non-linear relationships. These architectures of DNNs foster the creation of compositional models, wherein additional layers facilitate the synthesis of features from lower layers. This attribute grants DNNs an extensive learning capacity, enabling the effective modeling of intricate patterns inherent in speech data [24].

Notably, a pivotal advancement in the realm of large vocabulary speech recognition materialized in 2010 through the collaboration of industrial and academic researchers, marking the success of DNNs. This achievement was realized by employing large output layers within the DNN framework, constructed based on context-dependent HMM states formulated through decision trees [25][26]. A comprehensive elucidation of this development, along with the contemporary state-of-the-art insights as of October 2014, is available in a recent Springer publication by Microsoft Research [27]. Additionally, comprehensive background discussions on automatic speech recognition and the transformative influence of various machine learning paradigms, notably encompassing deep learning, are expounded upon in recent overview articles [28].

1.3 STT tools evaluation metrics and aspects

Evaluation of speech-to-text tools involves assessing the performance and accuracy of systems designed to convert spoken language into written text. This evaluation process plays a crucial role in determining the efficacy and reliability of these tools. The evaluation typically encompasses multiple aspects, including transcription accuracy, language model quality, punctuation and capitalization correctness, and overall system performance.

Evaluators employ diverse methodologies, such as comparing the output of speech-to-text tools against manually transcribed reference texts or using objective metrics like Word Error Rate to quantify accuracy. The following sections contain key parameters to compare STT tools usable for the problem.

[4]Deep Neural Network

1.3.1 Word Error Rate

The Word Error Rate is the predominant evaluation metric for STT systems. It quantifies the percentage of incorrect words within a transcription relative to the total number of input words. These incorrect words are the result of erroneous insertions, replacements, or deletions made by the system during the transcription process. The WER[5] is formally defined by equation 1.1:

$$WER = \frac{I + R + D}{H} \tag{1.1}$$

where I is the number of inserted words, R is the number of replaced words, D is the number of deleted words, and H is the number of hits. Despite its popularity, WER is limited to accuracy at the word level [29].

1.3.2 Character Error Rate

The Character Error Rate is a metric for evaluating the accuracy of STT systems. It is calculated by counting the number of character errors that occur in a transcription compared to a reference transcript. CER[6] is a strict metric that provides a detailed view of the accuracy of individual characters in the transcription. The CER is formally defined by equation [30]:

$$CER = \frac{S + D + I}{N} \tag{1.2}$$

where S is number of substitutions, D is number of deletions, I is number of insertions, N is total number of characters in the reference transcript.

A lower CER value indicates a more accurate transcription. For instance, a CER of 5% means that the system correctly transcribed 95% of the characters in the reference transcript. While CER provides a granular view of character accuracy, it doesn't consider the overall structure of the transcribed text, such as grammaticality and sentence fluency.

1.3.3 Sentence Error Rate

Sentence Error Rate is a metric for evaluating the overall fluency of an STT system's transcription. It measures the percentage of sentences in the transcription that contain at least one error. A sentence error can involve any combination of word errors, insertions, deletions, or modifications. SER[7] provides a broader assessment of the

[5] Word Error Rate
[6] Character Error Rate
[7] Sentence Error Rate

STT system's ability to produce grammatically correct and coherent sentences. The SER is formally defined by equation [31]:

$$SER = \frac{E}{S} * 100 \tag{1.3}$$

where E is number of errors (combination of substitutions, deletions, insertions, and modifications), S is total number of sentences in the reference transcript. A lower SER value indicates a more fluent transcription. For instance, a SER of 2% means that the system accurately transcribed 98% of the sentences in the reference transcript. SER provides a more holistic view of the STT system's ability to produce grammatically correct and coherent sentences.

1.3.4 Language support

Language support is a crucial aspect when it comes to speech-to-text (STT) tools. The ability of these tools to accurately transcribe spoken language relies heavily on their proficiency in recognizing and processing different languages. Language support entails recognizing and transcribing words and understanding the grammatical rules, punctuation conventions, and nuances specific to each language.

1.3.5 Speaker recognition

Speaker recognition involves identifying individuals based on voice traits. This technology aims to determine the identity of a speaker, answering the question of 'Who is speaking?' Voice recognition, a term often used interchangeably, may encompass speaker recognition or speech recognition. It's important to note the distinctions between speaker verification, which focuses on confirming an individual's identity, and identification, which involves recognizing a speaker's identity [32].

1.3.6 Price

Based on price, we can divide STT tools into free and paid. Free tools mostly use open-source licences. This availability has an effect on the absence of customer support. The quality of documentation might be different based on the product. The paid products, on the other side, are very easy-to-use applications. Customer support can help with problems. However, the paid amount usually depends on the length of transcripted records. If it is a necessary transcript of a massive amount of records might price skyrise on high values.

Based on our experience, we can say that the difference between paid and free software is not only the comfort of using but mainly the quality of transcriptions. Most of the free tools support English or another main language, Czech usually absent.

Chapter 2

Anonymization

2.1 Introduction

In the digital era, where data is abundantly generated and collected, the ethical handling of sensitive information becomes paramount, especially in fields dealing with personal and confidential data. This chapter delves into the crucial aspect of anonymization in the context of the transcribed calls from the National Quitline. Our objective is to ensure that the privacy and confidentiality of the individuals involved in these calls are rigorously protected while allowing for the valuable insights derived from these conversations to be utilized for academic and clinical purposes.

Anonymization, in the context of this study, refers to the process of removing or altering personally identifiable information from the transcriptions of calls, thereby ensuring that the identities of the individuals cannot be traced or inferred. This process is vital not only to comply with data protection regulations, such as the General Data Protection Regulation in the European Union, but also to uphold ethical standards in research.

The challenge lies in effectively anonymizing the data without compromising the integrity and the utility of the information for research purposes. This involves identifying and categorizing the types of sensitive information present in the transcriptions, developing methodologies to remove or obscure this information, and ensuring that the anonymized data remains coherent and meaningful for subsequent analysis.

In this chapter, we will explore the theoretical underpinnings of data anonymization, review the legal and ethical considerations specific to handling sensitive health-related data, and propose a methodological framework for the anonymization process. We will also discuss the implications of this process for data quality and the potential challenges in balancing privacy concerns with research needs. This will include examining various techniques and tools available for data anonymization and assessing their applicability and effectiveness in the context of our project with the National Quitline.

2.2 Legal and Ethical Considerations in Handling Sensitive Health-Related Data

In the context of the National Quitline data anonymization project, these considerations are particularly relevant as the project involves transcriptions of confidential calls between individuals seeking support for addiction and trained professionals. The project's objective is to extract valuable insights from these conversations while protecting the privacy and confidentiality of the individuals involved.

Therefore, adhering to legal and ethical guidelines and regulations is crucial while handling this sensitive health-related data. This involves developing methodologies for anonymization that ensure that the data is still helpful for research purposes while protecting the identities of those involved in the calls.

2.2.1 General Data Protection Regulation

The General Data Protection Regulation is an EU[1] regulation that sets out rules for the processing of personal data by organizations operating within the EU. Personal data includes health data, which is considered to be a special category of personal data due to its sensitive nature. The GDPR[2] imposes stricter requirements for the processing of health data compared to other types of personal data [33].

The GDPR establishes seven key principles that must be adhered to when processing personal data. These principles are foundational to GDPR compliance and serve as guidelines for organizations handling personal data [34]:

1. **Lawfulness, fairness, and transparency:** Personal data must be processed lawfully, fairly, and in a transparent manner. This means that organizations must have a legal basis for processing personal data, and they must be transparent about how they are collecting and using it [35].

2. **Purpose limitation:** Personal data must be collected for specified, explicit, and legitimate purposes and not further processed in a way that is incompatible with those purposes. This means that organizations must only collect personal data for the purposes they have told the individual about, and they cannot use it for other purposes without the individual's consent[36][37].

3. **Data minimization:** Personal data must be adequate, relevant, and limited to what is necessary in relation to the purposes for which they are processed. This means that organizations should only collect the personal data that they actually need for the purposes they have stated [33].

4. **Accuracy:** Personal data must be accurate and, where necessary, kept up to date. This means that organizations must take steps to ensure that the personal data they hold is accurate and up to date [33].

[1] European Union
[2] General Data Protection Regulation

5. **Storage limitation:** Personal data shall be kept in a form that permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed. This means that organizations should only store personal data for as long as they need it for the purposes they have stated [35].

6. **Integrity and confidentiality:** Personal data must be processed in a manner that ensures appropriate security of the personal data, including protection against unauthorized or unlawful processing and against accidental loss, destruction, or damage, using appropriate technical or organizational measures. This means that organizations must take steps to protect personal data from unauthorized access, use, disclosure, alteration, or destruction [35].

7. **Accountability:** The data controller is responsible for complying with the principles outlined above and must be able to demonstrate compliance. This means that organizations must have procedures in place to ensure that they are complying with the GDPR, and they must be able to demonstrate this to regulators [38].

2.2.2 Ethical Dilemmas and Challenges

Balancing Privacy and Innovation

In the healthcare sector, the pursuit of innovation often intersects with the preservation of patient privacy, creating a delicate balancing act. While technological advancements in healthcare hold immense potential for improving patient care, enabling personalized medicine, and accelerating medical research, they also raise concerns about the protection of sensitive personal health information.

The tension between privacy and innovation is particularly evident in the realm of big data and artificial intelligence [39]. The vast troves of healthcare data generated through electronic health records, wearable devices, and genomic sequencing offer valuable insights for understanding disease patterns, developing new treatments, and tailoring care to individual patients. However, this data collection, storage, and analysis raise concerns about potential misuse, unauthorized access, and the erosion of patient privacy.

To balance privacy and innovation, healthcare organizations and technology developers must prioritize data security measures, implement robust consent procedures, and establish clear data governance frameworks. Transparency and accountability are crucial in fostering public trust and ensuring patient data is used ethically and responsibly [40].

2.2.3 Data Sharing and Use

The sharing of healthcare data for research or commercial purposes can generate valuable insights and contribute to the development of new treatments and diag-

nostic tools. However, it also raises ethical concerns about patient privacy, informed consent, and potential commercial exploitation.

When sharing healthcare data, it is essential to obtain explicit and informed consent from patients, clearly explaining the purpose of data use, potential risks and benefits, and the safeguards in place to protect privacy [41]. Data minimization principles should be followed, ensuring that only the data necessary for the specific purpose is shared. Additionally, data-sharing agreements should establish clear guidelines for data storage, access controls, and reporting requirements [42].

The commercialization of healthcare data also raises ethical concerns, as it can lead to the monetization of patient information and potentially exacerbate healthcare disparities. It is crucial to ensure that commercial interests do not overshadow patient privacy and that data-sharing agreements are structured to prioritize patient benefits and not solely commercial gains [43].

2.2.4 Handling Sensitive Information

Certain types of healthcare data, such as genetic or mental health information, are particularly sensitive due to their potential to reveal personal traits, predispositions, or past experiences. Handling this information requires heightened ethical considerations and additional safeguards to protect patients from discrimination, stigma, or potential misuse.

Mental health information is sensitive, as it can reveal personal struggles, diagnoses, or treatment histories. Handling this information requires a high degree of confidentiality and respect for patient privacy. Mental health professionals should have clear guidelines for data management, ensure patient consent for data sharing, and protect against potential stigma or discrimination [44].

In conclusion, navigating the ethical dilemmas and challenges surrounding healthcare data handling requires a nuanced approach that balances innovation with privacy, promotes responsible data sharing, and ensures the protection of sensitive information. By prioritizing patient trust, transparency, and ethical principles, healthcare organizations can harness the power of data to improve patient care while safeguarding the fundamental rights of individuals.

2.3 Data anonymization

Data anonymization, a vital aspect of information sanitization, primarily aims to safeguard privacy. This process involves meticulously removing personally identifiable information from datasets. Anonymization stands at the fore of efforts to balance the utility of big data with the imperative of protecting individual privacy.

In the era of digital information, data mining has become an integral part of various industries, offering unprecedented insights and opportunities for business and

research. However, this surge in data utilization brings forth significant challenges, particularly concerning the privacy and security of individuals. The increasing concerns about privacy in data mining necessitate stringent measures to restrict the use of data containing personal identifiers, a crucial step in protecting individual privacy [45].

Anonymization emerges as a pivotal process in ensuring that the valuable insights drawn from data do not come at the cost of compromising sensitive information. The balance between maintaining the utility of data and minimizing the risk of disclosing sensitive information is a delicate and crucial aspect of data analytics [46]. It is evident in scenarios such as energy companies using smart meter data, where operational necessities must be weighed against the imperative of data privacy. The trend towards external sharing or release of consumption data further underscores the importance of anonymization in such contexts [47].

The healthcare sector, in particular, exemplifies the critical need for anonymization. The potential risks associated with patient privacy in big data analytics highlight the urgency of implementing effective anonymization strategies before data is analyzed or shared [48]. The sharing and publication of data raise concerns about privacy breaches. Combining multiple datasets that have been published or shared with third parties can potentially reveal a comprehensive profile of an individual. For instance, medical information can be inferred by linking patient data to public voter registration lists or by combining survey responses from the public with published data [49].

Moreover, research findings indicate that simple demographic attributes like birth date, zip code, and gender could lead to identifying almost 60% of individuals, demonstrating the ease of conducting linkage attacks to infer sensitive information [50]. These specific attributes hold the potential to uniquely identify a large portion of the population through a process known as a linkage attack. In such an attack, the aim is to infer sensitive attributes of the targeted individuals [51].

As privacy requirements become more stringent, the number of techniques required to safeguard privacy also increases, sometimes exponentially. Privacy-preserving techniques such as classification, anonymization, association rule mining, and clustering have been proposed to address these concerns. In large datasets, it is crucial to strike a balance between maintaining data utility and ensuring that the privacy-preserving techniques employed do not introduce excessive computational complexity [49].

2.3.1 Anonymization Techniques

In the realm of data privacy, especially concerning the anonymization of sensitive information, it's imperative to understand the challenges and nuances involved in the process. Anonymization is a critical step in ensuring that individuals' private information remains protected when datasets are released. However, the effectiveness of anonymization is contingent upon the methodologies employed to prevent various forms of data disclosure.

Types of Information Disclosure:

(A) **Identity Disclosure:** This form of disclosure arises from inadequate anonymization techniques. For instance, simplistic algorithms that replace 'A' with '1', 'B' with '2', and so on, may inadvertently lead to re-identification. This happens when specific records in the anonymized dataset can be linked back to their original identities, revealing sensitive information [52].

(B) **Attribute Disclosure:** This occurs when anonymized data inadvertently reveals new information about an individual. A classic example is an anonymized dataset of employee records indicating that all employees over a certain age received a bonus. If it is known that a specific employee falls into this age group, their receipt of the bonus becomes discernible, despite their individual record being indistinguishable within the dataset [49].

(C) **Inference Disclosure:** Inference disclosure is a risk when confidential information is deduced by correlating anonymized data with other datasets. This form of disclosure involves indirect leakage of sensitive data through inferences drawn from the released anonymized data [53].

2.3.2 Data structure

Structured data

Structured data refers to information that is organized in a defined manner, often in databases or spreadsheets. This data is typically easier to analyze due to its consistent format. Common techniques for anonymizing structured data include data masking, pseudonymization, and aggregation. These methods systematically transform or summarize data to prevent the identification of individuals [54].

Unstructured data

Unstructured data includes text, audio, video, and images. It lacks a predefined data model, making it more complex to process and anonymize. Techniques for unstructured data involve natural language processing, image processing, and other advanced methods to identify and obscure personal identifiers. These may include redaction, blurring, or altering elements within the data [55].

2.4 Anonymization of unstructured textual data

This work primarily focuses on the anonymization of unstructured textual data. A multitude of methods have been proposed to address this challenge, primarily classified into two categories: dictionary-based approaches and machine-learning methods. Considering the merits and advantages gleaned from existing research in this

domain, our approach to addressing this issue leans toward utilizing a machine-learning methodology [56].

The process of anonymizing unstructured data involves a sequential dual-step procedure. Firstly, it entails the identification of entities embedded within the dataset. To achieve this, we will employ a machine learning-based approach, specifically the Named Entity Recognition task, as an initial phase. Once these entities are successfully identified, the subsequent step involves the application of structured anonymization methods [56] [57].

2.5 Named Entity Recognition

Named Entity Recognition or NER[3], is a technique focused on identifying specific mentions, known as rigid designators, within text. These mentions typically fall into predetermined semantic categories like person, location, organization, etc [58]. NER serves not only as a standalone tool for extracting information from text but also holds a crucial role in numerous natural language processing applications. It contributes to text comprehension, information retrieval, automatic text summarization, question answering, machine translation, and the construction of knowledge bases [59]. The evolution of NER traces back to its early recognition as "Named Entity" during the sixth Message Understanding Conference, which involved identifying various entities like organizations, individuals, geographic locations, as well as expressions related to currency, time, and percentages [60].

2.5.1 Definition

Named Entity Recognition, refers to the process of identifying and categorizing specific words or phrases within text that distinctly represent individual items from a group sharing similar attributes [61]. These named entities encompass various categories such as organization, person, location names in general domains, and specialized terms like gene, protein, drug, and disease names in biomedical contexts. The primary objective of NER is to locate these entities within text and classify them into predefined categories [59].

In more formal terms, when provided with a sequence of tokens represented as $s =< w_1, w_2, ..., w_N >$, the task of NER involves producing a list of tuples, denoted as I_s, I_e, t_i. Each tuple within this list represents a named entity present in the sequence s. Here, $I_s \in [1, N]$ and $I_e \in [1, N]$ refer to the start and end indexes respectively, indicating the span of the identified named entity within the sequence. The t variable signifies the entity type belonging to a predefined category set, characterizing the recognized named entity [59]. The figure 2.1 shows an example where a NER system recognizes three named entities from the given sentence.

[3]Named Entity Recognition

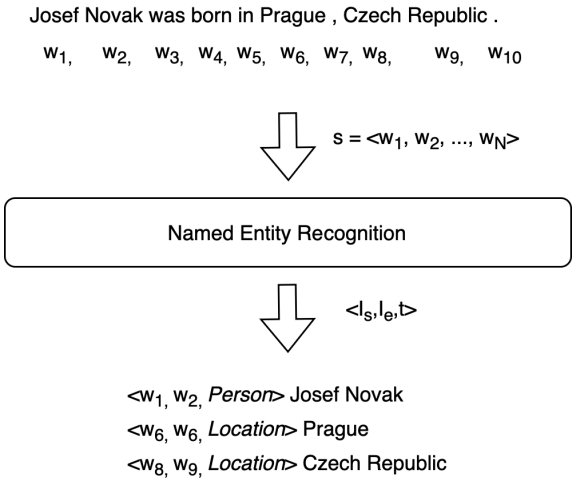


Figure 2.1: An illustration of the NER task [62]

2.5.2 Traditional approaches to NER

Traditional approaches to NER can be broadly categorized into rule-based, unsupervised learning, and feature-based supervised learning methods [58]. Rule-based methods rely on manually crafted rules, unsupervised methods utilize statistical techniques to learn patterns from unlabeled data, and supervised methods train models on labeled data to predict entity labels. These methods are being replaced by DL[4]-based approaches as deep learning evolves.

2.5.3 Deep Learning techniques for NER

Over the past few years, NER models based on Deep Learning have risen to prominence, establishing themselves as the frontrunners in achieving state-of-the-art results [59]. Deep Learning, unlike feature-based approaches, excels in autonomously uncovering latent features, thereby proving advantageous in NER tasks. To delve deeper, we provide a succinct introduction to the concept of Deep Learning and elucidate its relevance in NER. Subsequently, an overview of DL-based NER approaches is presented.

Deep Learning and NER

Deep learning is a subfield of machine learning that employs a hierarchical arrangement of processing layers to extract increasingly abstract representations from data. These layers, typically composed of artificial neural networks, involve two distinct phases: the forward pass and the backward pass. The forward pass calculates a weighted sum of inputs from the preceding layer and applies a non-linear transformation to the result. In contrast, the backward pass computes the gradient of an

[4]Deep Learning

objective function with respect to the weights of a multi-layered network stack using the chain rule [63].

A fundamental advantage of deep learning lies in its ability to learn representations and perform semantic composition, facilitated by both vector representation and neural processing. This enables machines to process raw data and autonomously uncover latent representations and processing pipelines necessary for classification or detection tasks [63].

2.5.4 Recurrent Neural Networks

A recurrent neural network represents one of the two primary types of artificial neural networks, distinguished by the information flow within its layers. Unlike the unidirectional feedforward neural network, it operates bidirectionally, allowing output from certain nodes to influence subsequent input to those same nodes [64] [65] [66]. Their capability to utilize internal memory to handle diverse input sequences makes them applicable in tasks such as continuous handwriting recognition or speech analysis. While the term "recurrent neural network" specifically denotes networks with an infinite impulse response [67] [68].

A finite impulse recurrent network can be depicted as a directed acyclic graph that can be unwound and substituted with a strictly feedforward neural network. Conversely, an infinite impulse recurrent network is a directed cyclic graph that cannot be unwound [69].

Both infinite-impulse and finite-impulse networks can incorporate additional stored states controlled directly by the network. These states can be replaced by another network or graph introducing time delays or feedback loops [70]. These controlled states are termed gated states or gated memory, integral to long short-term memory networks and gated recurrent units. These networks are also referred to as Feedforward Neural Networks. RNN[5] possess theoretical Turing completeness, enabling them to execute arbitrary programs for processing diverse input sequences [71].

Word embeddings are provided for a bidirectional LSTM[6] like on 2.2 figure. In this setup, l_i denotes the word i along with its left context, while r_i represents the word i with its right context. Combining these two vectors results in a representation of the word i within its contextual surroundings, termed as c_i [62].

2.5.5 Convolutional Neural Network

A convolutional neural network is a structured form of a feed-forward neural network that autonomously learns feature engineering through filter optimization. Unlike earlier neural networks, where issues like vanishing gradients and exploding gradients

[5]Recurrent Neural Network
[6]Long Short-Term Memory

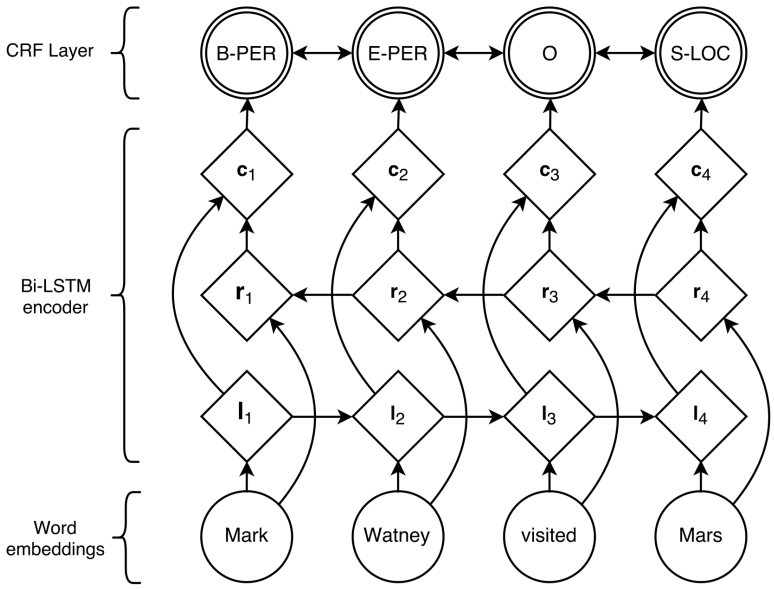


Figure 2.2: The architecture of the RNN network dedicated to NER [62].

occurred during backpropagation, CNN[7]s prevent these problems by employing regulated weights across fewer connections [72][73].

RNNs operate sequentially, computing along the input sequence's positions. This sequential nature causes heavy dependency on preceding time steps, limiting their exploitation of GPU parallelism during training and testing. Consequently, RNNs incur higher time costs for these tasks. In contrast, CNNs process all sentence words simultaneously in a feed-forward manner, optimizing GPU[8] parallelism effectively.

However, in NER tasks, CNNs have received less attention due to their focus on capturing local context over long-term context, unlike the more adept Long Short-Term Memory networks inherent in RNNs. CNNs, while capable of expanding their receptive field through techniques like stacking convolutional layers or using dilated convolutional layers, still struggle to capture comprehensive global context, particularly in sentences of varying lengths. This limitation impedes CNNs from matching the performance of LSTMs in NER applications [74].

2.5.6 Deep Transformers

Commonly, neural sequence labeling models rely on intricate convolutional or recurrent networks comprising both encoders and decoders. However, the Transformer model, introduced by Vaswani et al. [75], breaks away from this convention by eschewing recurrent and convolutional structures entirely. Instead, it employs stacked self-attention mechanisms and point-wise, fully connected layers to construct fundamental components for both the encoder and decoder. Empirical evaluations across different tasks demonstrate the superiority of Transformers in terms of quality,

[7]Convolutional Neural Network
[8]Graphics Processing Unit

achieved with notably reduced training time [75][76][77].

Language model embeddings pretrained using Transformer architectures are ushering in a new era for NER. These embeddings, first and foremost, offer contextualization and serve as potential replacements for conventional embeddings like Google Word2vec and Stanford GloVe. Several studies, have shown promising results by combining traditional embeddings with these language model embeddings [78][79][80]. Furthermore, these specialized language model embeddings allow for additional fine-tuning by adding just one output layer, facilitating their application across various tasks, including NER and chunking [59].

We'll discuss two transformer models: Bidirectional Encoder Representations from Transformers and Generative Pre-trained Transformer. The figure 3.8 illustrates variances in their pre-training model architectures. Google's BERT employs a bidirectional Transformer (abbreviated as 'Trm'), while OpenAI's GPT utilizes a left-to-right Transformer.

Bidirectional Encoder Representations from Transformers

BERT[9] was unveiled in October 2018 by a team of researchers from Google [81]. At its core, BERT comprises a series of Transformer encoder layers [75], each containing numerous self-attention 'heads'. These heads independently compute key, value, and query vectors for every input token within a sequence, forming a weighted representation. The outputs from all heads within a given layer are amalgamated and processed through a fully connected layer. Every layer is enveloped by a skip connection and then subjected to layer normalization [82].

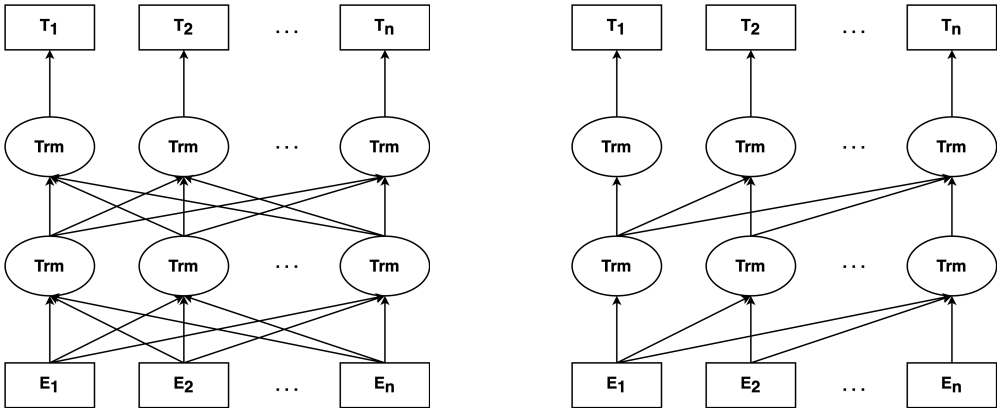
The traditional process for BERT involves a sequence of steps: pre-training and fine-tuning. During pre-training, two self-supervised tasks are utilized: masked language modeling, where randomly masked tokens are predicted, and next sentence prediction, which predicts the adjacency of two input sentences. When fine-tuning for specific applications, additional fully connected layers are usually added above the final encoder layer [82].

To generate input representations, each input word undergoes tokenization into wordpieces following [83]. Subsequently, a combination of three embedding layers—token, position, and segment—is employed to create a fixed-length vector. The special tokens and are used for classification predictions and to separate input segments, respectively [82].

Generative Pre-trained Transformer

Generative pretraining had been a well-established concept in machine learning applications [23][84]. However, the transformer architecture, crucial for this approach, only came into existence in 2017 through Google's team [81]. This development

[9]Bidirectional Encoder Representations from Transformers



(a) BERT **(b)** GPT

Figure 2.3: Differences in BERT and GPT models architectures

paved the way for the emergence of large language models like BERT, a pre-trained transformer lacking generative capabilities, being an "encoder-only" model [85]. In 2018, OpenAI introduced the first generative pre-trained transformer (GPT[10]-1) in their article "Improving Language Understanding by Generative Pre-Training" [86].

Before the advent of transformer-based architectures, the most successful neural NLP models relied heavily on supervised learning from extensively labeled data. This reliance on supervision restricted their applicability to well-annotated datasets, making training large language models excessively costly and time-consuming [86].

OpenAI took a semi-supervised approach, pioneering a large-scale generative system with a transformer model. This approach involved two stages: an unsupervised generative "pretraining" phase to establish initial parameters using a language modeling objective, followed by a supervised discriminative "fine-tuning" stage to adapt these parameters to a specific task [86].

2.6 NER metrics

In this work, the evaluation and benchmarking of NER models constituted a pivotal aspect of the research. Consequently, the ensuing section delineates fundamental NER metrics intended for use as objective benchmarks. These metrics serve as foundational criteria to assess and gauge the performance of NER models under scrutiny, facilitating a comprehensive and standardized evaluation process.

Precision

Precision refers to the accuracy of a model in its predictions. It quantifies the ratio of correctly identified positive instances (true positives) to all instances classified as positive. This metric illuminates the proportion of predicted entities that are accurately labeled, calculated as follows [87]:

[10]Generative Pre-trained Transformer

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives+False Positives}}$$

Recall

Recall measures a model's capability to predict actual positive classes. It represents the ratio of predicted true positives to the total instances that were indeed tagged as positive. The recall metric highlights the accuracy of predicted entities, calculated by [87]:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives+False Negatives}}$$

F1 score

The F1 score, a function combining Precision and Recall, becomes essential when seeking a balance between these metrics. It serves as a harmonizing factor between Precision and Recall, calculated using the formula [87]:

$$\text{F1 Score} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision+Recall}}$$

Chapter 3

Implementation part

3.1 Introduction

The primary objective of this endeavor involves the systematic processing of telephone conversations obtained from the national Quitline to generate structured data suitable for utilization as a foundational resource for Natural Language Processing applications. The schematic diagram delineates the successive stages constituting the procedural framework adopted for this purpose. Each stage necessitates a comprehensive assessment of prerequisites and selection of pertinent tools.

The initial phase encompasses the conversion of audio calls in WAV[1] format to textual transcripts. This step entails an exhaustive exploration of available tools conducive to this purpose, followed by a comparative evaluation of their efficacy using our dataset. Termed as the speech-to-text or speech recognition task, this phase also involves the crucial aspect of speaker identification within the recorded conversations.

Subsequently, the subsequent phase mandates the precise identification of speakers within these recordings, specifically distinguishing between addiction specialists and clients of the Quitline. This demarcation assumes critical significance in preserving the contextual integrity of the dataset, thereby potentially enhancing the performance of NLP[2] models in responding to specific inquiries.

The overarching aim of this project extends beyond the confines of the Czech language to encompass an international ambit. Consequently, the availability of data in the English language becomes pivotal. However, considering that the calls originating from the National Quitline are in Czech, a requisite part of the workflow involves the translation of these transcriptions into English.

[1] Waveform Audio File Format
[2] Natural Language Processing

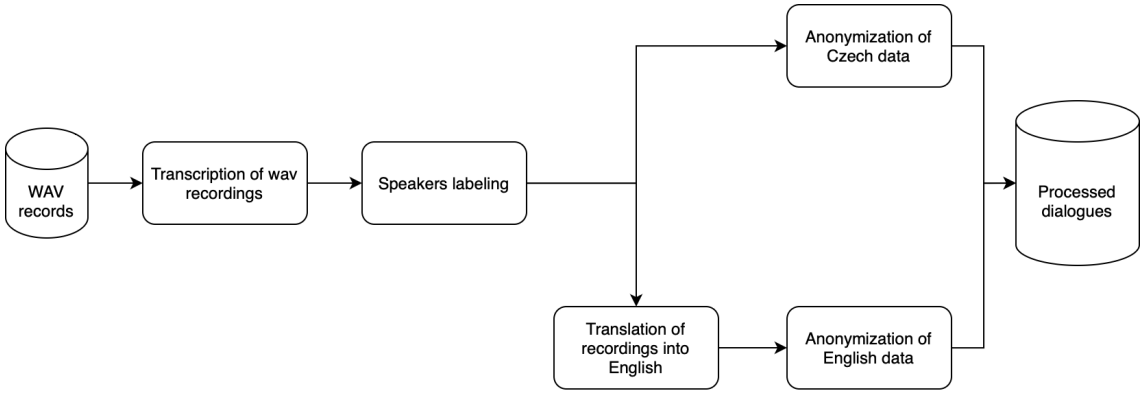


Figure 3.1: Workflow of processing dialogue date presented in the thesis

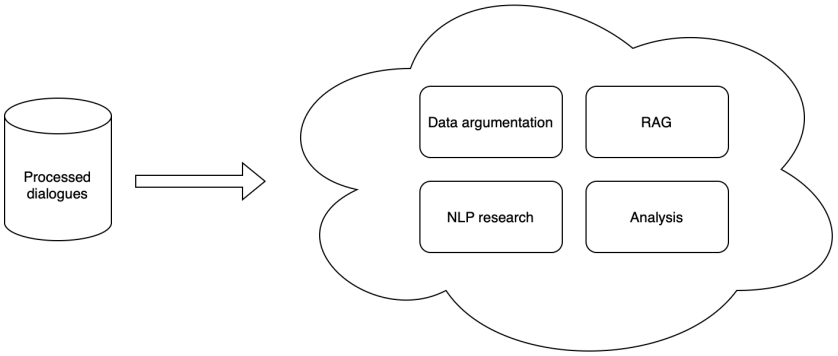


Figure 3.2: Possible usage of processed dialogues

The conclusive stage of the data processing involves the anonymization or removal of sensitive information embedded within the dataset. Given the existence of two datasets - the original Czech transcripts and their translated English counterparts - the strategy employed to address this concern will be tailored separately for the Czech and English datasets to ensure the preservation of data integrity and privacy.

Post-dialogue processing yields several potential applications, as depicted in Figure 3.2. One conceivable avenue for discussion involves employing the data for fine-tuning LLM[3] models. This dataset holds promise as a valuable resource in the development of models endowed with the expertise of trained addiction specialists. Another prospect pertains to Retrieval-Augmented Generation (RAG[4]). Furthermore, the utilization of this data for comprehensive data analysis may yield insightful perspectives into the operations of the Quitline. Given the recent advancements in NLP, additional unexplored possibilities may also emerge.

[3]Large Language Model
[4]Retrieval-Augmented Generation

3.2 Simple analysis of records

3.2.1 Metadata in filenames

The filename formats of the WAV files provide valuable insights into the nature of the records. The most pertinent information extracted from the filenames is the date, time, name of operator and phone number of the calling client. To maintain consistent client representation across multiple calls, the phone number was deemed the primary identifier. This decision was based on the premise that client information, rather than the specific addiction specialist involved, would be more relevant for the intended analyses. Therefore, tracking client interactions based solely on their phone numbers was deemed sufficient for the project's objectives. The filenames were presented in two distinct formats:

DT_{date}_TM_{time}O_{addiction_specialist_name}_TA_F_{phone number}_.{unique_id_of_record}.WAV

Figure 3.3: Format A of filenames

DT_{date}_TM_{time}F_{phone_number}___TA_O_202_{addiction_ specialist_name}_QU_{diagnose}_1627301701.{unique_id_of_record}.WAV

Figure 3.4: Format B of filenames

3.2.2 Analysis

For a more comprehensive understanding of the data, several initial observations were conducted. The ensuing figures potentially contain pertinent insights that could elucidate the underlying nature and characteristics of the dataset.

Distribution of calls on the Quitline over time

The figure 3.5 assists in delineating the temporal scope of the records, encompassing the time span from September 2020 to April 2022. This duration notably encapsulates the period encompassing the COVID-19 pandemic.

However, deriving conclusive inferences regarding the pandemic's impact on the volume of calls to the Quitline is challenging based solely on the depicted figure. Nevertheless, it is discernible that on a monthly basis, the Quitline consistently managed a call volume ranging between approximately 350 to nearly 500 calls. This steady range suggests a consistent demand for Quitline services throughout the recorded period, irrespective of the pandemic's occurrence. We can see a drop-down in calls in April 2022, which is caused by the fact that we don't have records from the whole month.

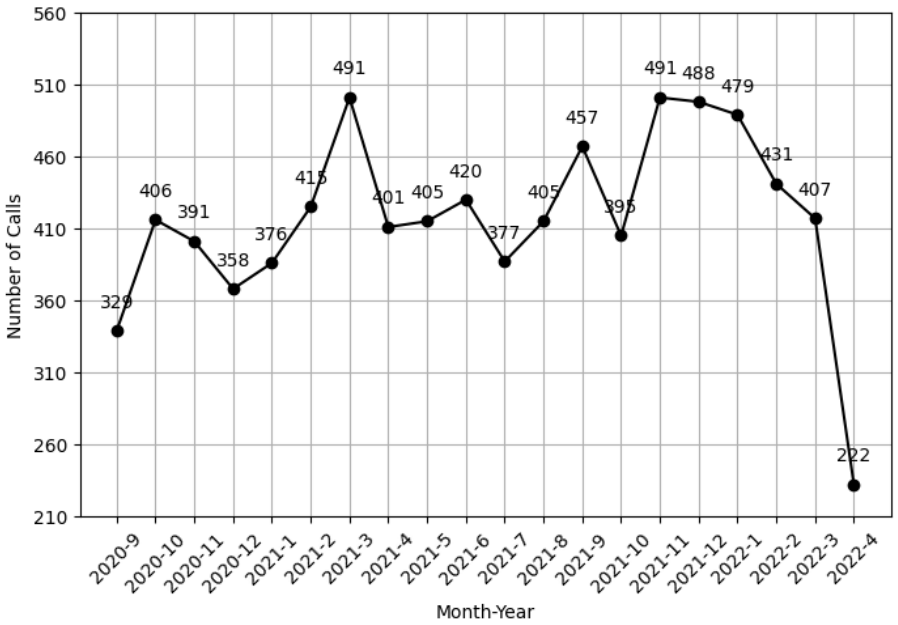


Figure 3.5: Number of calls on the Quitline by months

Frequency of calls on the Quitline by phone number

The figure 3.6 illustrates the distribution of phone numbers based on the frequency of their calls to the Quitline. A total of 2978 distinct phone numbers were included in the analysis. The predominant trend observed indicates that the vast majority of calls made to the Quitline were singular occurrences. Moreover, the data reveals where two callers contacted the Quitline 37 times, representing the maximum frequency within the dataset.

Cumulative length of time on the Quitline by phone numbers

The figure 3.7 displays the distribution of the aggregate call durations made to the Quitline per phone number. On average, clients expended approximately 31.29 minutes during their calls. The median call duration, representing the midpoint value, was recorded at 12.65 minutes. Remarkably, the maximum cumulative call duration observed within the dataset reached 505.4 minutes, 488.1 minutes and 449.5 minutes by clients. On the other hand, records contain 0.0 minutes, which we will exclude from further investigation. We decided to set a threshold of 3 minutes as the minimum duration for processing in the following steps (total 416 records with duration < 3 minutes).

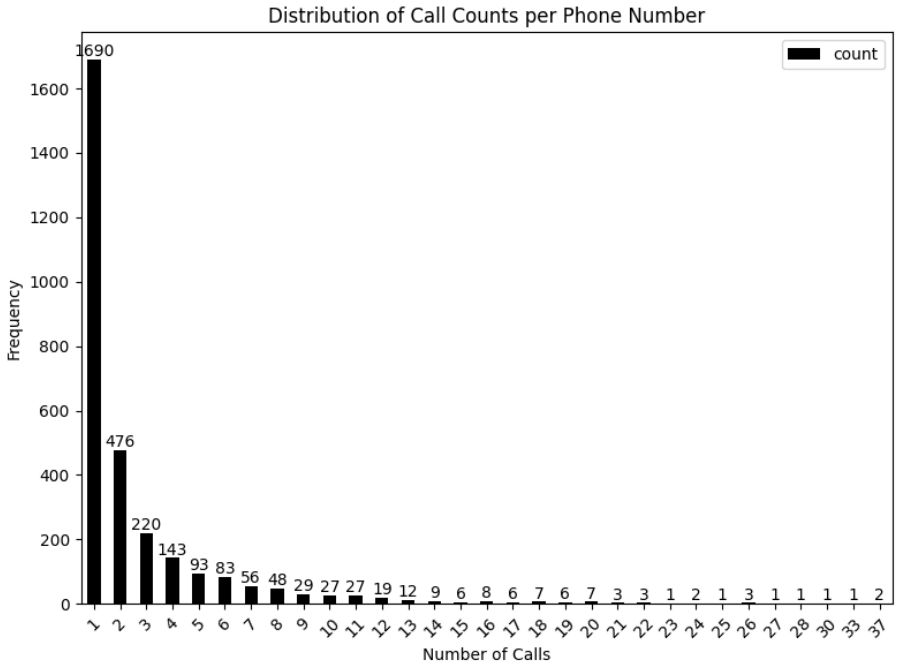


Figure 3.6: Distribution of Call Counts per Phone Number

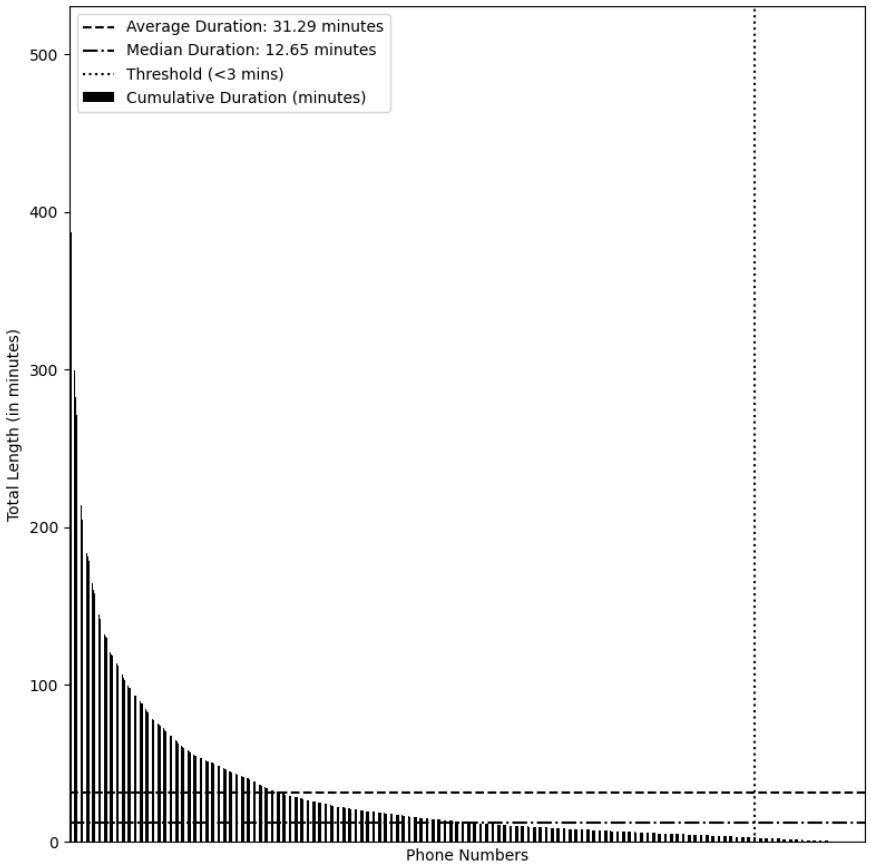


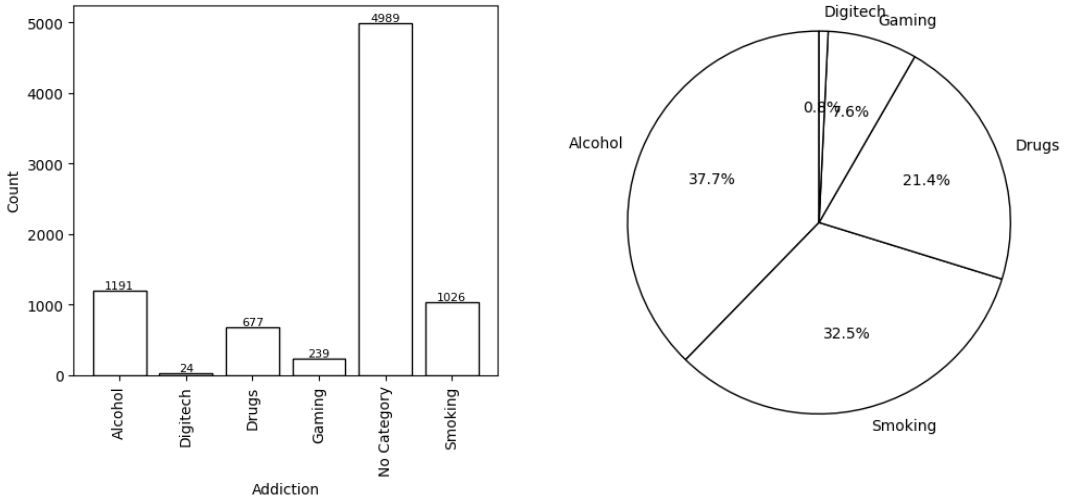
Figure 3.7: Distribution of Cumulative Duration of Calls by Phone Number on the Quitline

Distribution of addictions

The records, designated in the Format B within the file names, potentially contain information pertaining to addiction. Figure 3.8a illustrates the distribution of these records, highlighting the presence or absence of an assigned addiction category. It is noteworthy that a substantial majority of these records lack an assigned category. Nevertheless, a relatively sizeable sample still includes records with designated categories.

Figure 3.8b provides an overview of the distribution of categories within the records where an addiction category has been assigned. Notably, a predominant portion of these categorized records pertains to alcohol addiction. This observation might correlate with a noteworthy consumption of alcohol per capita, as indicated in the WHO report of [88][89]. Following closely behind is the category of smoking addiction, demonstrating a significant presence. Additionally, drug addiction emerges as the third most prevalent category.

It's important to note that gaming and digitech addictions collectively constitute less than 8% of the recorded cases, suggesting a lower prevalence compared to the aforementioned categories.



(a) Categorized and uncategorized records by addiction

(b) Distribution of categories by addiction

Figure 3.8: Categorization of calls to the Quitline depending on addiction

3.3 Transcription of WAV recordings

3.3.1 Tools for speech recognition

The following section presents a comprehensive list of STT tools that were considered in our research. The selection process for these tools was primarily guided by the parameters and criteria outlined in the theoretical part. We focused on factors deemed crucial for accurate and efficient transcription, aiming to identify tools that align with our research objectives.

Beey

Beey is an online tool developed by the Czech company Newton Technologies, offering voice recognition capabilities. It is primarily a web application featuring a transcription editor. The Beey editor is a robust tool that caters to a wide range of users, including journalists, video producers, lecturers, scribes, and individuals spanning from YouTubers to multinational media agencies. The tool supports 20 languages, including Czech, making it suitable for a diverse user base. Beey also facilitates team collaboration within the editor, enabling efficient workflow management. It is a paid tool available in two versions: the standard version with basic functions and the enterprise version, which offers additional features, including an API[5] [90].

Google Cloud Speech-to-Text

Google Cloud Speech-to-Text is an automatic speech recognition system developed by Google. It offers a comprehensive solution for converting spoken language into written text, and it has gained significant popularity and recognition in both academic and industrial domains. The system utilizes advanced deep learning models and neural networks to achieve high accuracy and performance in speech recognition tasks.

One key advantage of Google Cloud Speech-to-Text is its broad language support, as it is capable of transcribing audio in a multitude of languages and dialects. The core functionality of Google Cloud Speech-to-Text is delivered through a user-friendly API, allowing developers and researchers to integrate the system seamlessly into their applications or research projects. The API provides access to essential features such as real-time transcription, speaker diarisation, and word-level time alignment [91].

[5] Application Programming Interface

Microsoft Azure

This cloud-based service, provided by Microsoft, utilizes cutting-edge technology and machine learning algorithms to ensure accurate and reliable speech recognition. One of the distinguishing features of Microsoft Azure Speech to Text is its support for multiple languages, including Czech. However, it is important to note that Microsoft Azure Speech to Text operates on a paid model, requiring users to subscribe to a pricing plan to access its services. Additionally, it is worth mentioning that while the tool offers speaker recognition capabilities, this particular functionality is still in the beta stage, indicating that it is currently undergoing further development and refinement.

Sonix AI

Sonix AI is an automated STT software that can convert audio and video files into text in over 40 languages, including Czech. Sonix AI uses a deep learning model to transcribe audio accurately, even in noisy environments. The software also includes a speaker detection feature that can identify and label different speakers in a recording. Sonix AI also has an online editor that allows users to review and edit transcripts[92]. The software is available for a monthly subscription fee based on the amount of transcripted audio. Sonix also provides API with documentation available online [93].

Vosk

Vosk is an open-source speech recognition toolkit that can be used to transcribe audio into text. It is available for various platforms, including Linux, Windows, and macOS. Vosk is trained on a large dataset of audio recordings, and it can recognize speech in over 20 languages, including Czech. Vosk is a powerful tool that can be used for various purposes. Vosk is also a valuable tool for researchers working on speech recognition algorithms. Vosk is open-source, so it can be used to develop new speech recognition algorithms and improve existing algorithms' accuracy. Offline speech recognition: Vosk can be used to transcribe audio files that are stored on a computer or mobile device [94].

Whisper

Whisper AI is a speech recognition model developed by OpenAI. It is trained on a large dataset of diverse audio and is also a multitasking model that can perform multilingual speech recognition, speech translation, and language identification. Whisper AI is still under development, but it has already shown promising results. Whisper AI can be imported into Python and run on a local device with sufficient sources. Whisper supports multiple languages, including Czech, but not speaker detection [95].

Conclusion

In this stage, our primary objective was identifying and choosing tools aligned with our specific needs and requirements. After careful evaluation, we found that only two of the presented tools (Beey, Sonix) met our criteria, as highlighted in Table 3.1. Notably, both of these companies primarily specialize in the STT problem, which suggests their expertise in this domain.

One significant challenge we encountered with the other tools was their inadequacy in accurately handling speaker recognition. The tools that did not meet our requirements exhibited notable shortcomings in this aspect.

While we did not conduct a comprehensive quality assessment of the transcription output, we did attempt to transcribe a subset of the audio records. Based on this limited sample, we observed that Sonix and Beey outperformed the other tools in terms of transcription accuracy. However, it is crucial to acknowledge that this observation is based on a small number of records, and thus, it should be interpreted with caution.

Tool	Support Czech	Speaker diarisation	Price
Beey	YES	YES	175 Kcz/hour
Google Cloud Speech-to-Text	YES	Beta	$0.016/minute
Microsoft Azure	YES	NO	$1.4/hour
Sonix	YES	YES	$5/hour
Vosk	YES	NO	free
Whisper	YES	NO	$0.006/minute

Table 3.1: Overview of STT tools and their features

3.3.2 Comparison of Beey and Sonix

During this phase, our primary focus was on assessing the quality of transcription provided by the selected tools, namely Beey and Sonix, as determined in the previous phase. The main metric employed for evaluating the transcription quality was WER. To facilitate a fair and comprehensive comparison, we carefully selected a test sample of recordings that would serve as the basis for comparing transcriptions generated by the different tools.

Testing records

In order to conduct the quality testing, we opted to utilize the original call records themselves. Due to the substantial number of recordings available, we employed a random sampling approach to select recordings for evaluation. To ensure a comprehensive assessment, the recordings were divided into three distinct quality categories based on perceived audio quality:

- Good - clean sound without any significant defects
- Ok - a record contains some noise or other defects, but the content of dialogue is clear
- Bad - a record contains passages which it is hard to understand

The categorization of the recordings into their respective quality categories was determined subjectively by listeners who carefully reviewed each recording. Additionally, detailed notes were made regarding the content of the recordings to provide a comprehensive understanding of the overall content.

In total, we listened to 43 recordings using this methodology. Following the evaluation process, we subsequently selected three recordings from each quality category, resulting in a set of nine test recordings. The selection of these test recordings aimed to represent the range of quality observed in the larger dataset.

Testing

We utilized a set of selected recordings during the testing and conducted transcription attempts using both Beey and Sonix tools. Due to the paid nature of these tools, we utilized the unpaid versions, which had certain limitations. One such limitation was the maximum duration of the audio files for transcription, set at 30 minutes. Therefore, for recordings that exceeded this duration, we had to divide them into smaller sections to accommodate the tool's limitations.

Following the initial transcription process, we manually corrected the transcriptions to address any inaccuracies or errors. The corrected transcriptions were then compared with the original, uncorrected versions of the text. The quality of the resulting transcriptions was evaluated using the WER. The table provided includes the measured WER values for each transcription attempt, along with an indication of their respective quality.

Record	Audio Quality	Beey	Sonix
record 1	GOOD	7,3%	9,3%
record 2	OK	7,9%	4,6%
record 3	GOOD	6,2%	6,4%
record 4	OK	5,2%	11,3%
record 5	OK	20,3%	29,7%
record 6	BAD	15%	14,4%
record 7	OK	3,3%	3,6%
record 8	BAD	8%	4%
record 9	BAD	4%	3,4%

Table 3.2: WER of transcriptions

The table 3.2 presented illustrates the accuracy of each transcription attempt, with the second column indicating the quality of the corresponding audio recording. From the results, both Beey and Sonix were comparable in terms of transcription quality.

Beey outperformed Sonix in five cases, while Sonix exhibited higher accuracy in four cases.

Interestingly, we noted that the perceived quality of the audio recordings, as determined by human judgment, did not always align with the accuracy of the transcriptions. This discrepancy suggests that factors other than the overall quality of the recording, such as speech characteristics, background noise, or speaker variations, may have influenced the transcription accuracy.

One noteworthy observation is the transcription of record 5. Despite being classified as a "good" quality recording, it presented challenges for both tools. The conversation in this particular recording contained frequent pauses and fluctuations in the client's volume. These factors posed difficulties for accurate transcription, resulting in lower accuracy for both Beey and Sonix.

3.3.3 Tool selection for transcription

After careful analysis and consideration of the previous findings, we have made the decision to select Beey as the preferred tool for our problem of automated transcription of addiction quitline audio files. This choice was based on several key factors aligning with our requirements. One of the main reasons for selecting Beey is its support for the Czech language, which is essential for accurately transcribing the recorded conversations in our context. Additionally, Beey offers robust speaker recognition capabilities, allowing for effective identification and separation of individual speakers within the audio files. This feature is crucial for maintaining clarity and coherence in the transcriptions.

While both Beey and Sonix demonstrated similar levels of transcription quality, the company's location also influenced the decision to opt for Beey. Being a Czech company with a team in the Czech Republic provides us with more accessible communication and potential support from a local perspective. Furthermore, we thank our supervisor for his assistance in this process. Thanks to his effort, we were able to negotiate a discount with Newton Technology, the company behind Beey. As a result, we secured 1500 hours of transcription time at a highly favourable price point, which significantly contributes to the cost-effectiveness of our project.

Utilizing the Beey REST API, a total of 5160 records were transcribed. The resultant output comprised XML files, subsequently transformed into .txt file format to enhance usability and accessibility.

3.4 Translation of recordings into English

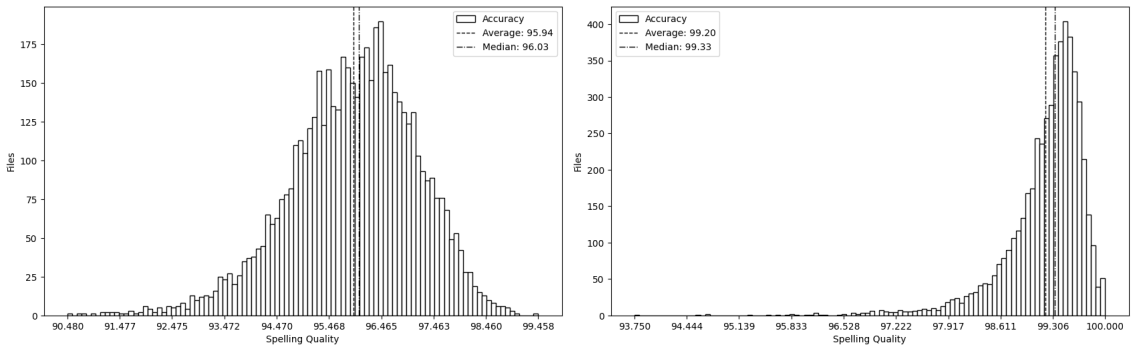
In the process of translating the recorded transcriptions, we leveraged a specific capability within Beey that facilitates the translation of text from one language to another. It's noteworthy that Beey relies on the Google Translate API for this purpose, establishing a consistent reliance on Google's translation tool. Notably, no

additional exploration or assessment of alternative translation models or tools was conducted. The determination was made that the solution offered by Google met the requisite standards of adequacy in terms of quality and ease of use, hence obviating the necessity for further investigation.

3.5 Transcriptions spelling quality

Following the transcription process, our focus shifted towards assessing the grammatical quality of the text. To accomplish this, we employed the open-source grammar spell tool, Hunspell. The evaluation of spelling quality was conducted through a straightforward formula: $spelling\ quality = \frac{words\ without\ error}{total\ words} * 100$. Each word underwent scrutiny by Hunspell, and any word unrecognized by the tool was deemed an error. In instances where a word contained multiple errors, it was considered as a singular error, given that the errors were confined to a single word.

However, it is imperative to acknowledge the limitations inherent in this approach. The simplicity of this test restricts its scope to the assessment of individual words, disregarding the intricate structures of sentences and the appropriate fit of words within a sentence. Linguistic phenomena such as grammatical cases, particularly prevalent in Slavic languages, and declension are regrettably overlooked. Addressing these nuanced issues would necessitate a more intricate approach. Nevertheless, for the specific case presented, the simplicity of our approach sufficed for the intended purpose.



(a) Czech dialogues **(b)** Translated English dialogues from Czech

Figure 3.9: Spelling quality of transcribed dialogues

Figure 3.9 presents an analysis of spelling quality observed in Czech and English dialogues, revealing discernible disparities between the two linguistic sets. The quality of Czech spelling spans within the range of 90% to 98%, whereas the majority of English dialogues exhibit a higher spelling quality, typically ranging between 98% and 100%.

3.6 Speakers labeling

In this phase, the objective was to enhance the transcribed dialogues by incorporating information regarding the speakers involved. Figure 3.10a illustrates the preliminary state of the dialogues subsequent to transcription via Beey and conversion into a text format. Each speaker is denoted by an index, commencing from 0 and extending sequentially.

However, discerning the roles of addiction specialists and clients from this format poses a challenge. Figure 3.10b portrays the desired outcome subsequent to identifying speakers within the dialogues. Notably, the presented dialogue is synthetic and generated by Google's BART model. Its purpose is to illustrate the structural arrangement of dialogues rather than representing authentic conversational content. This transformation enables a clearer delineation between different speakers, facilitating the identification of roles within the dialogue.

0: Hello, the national Quitline. How can I help you today?
1: Hi, I'm calling because I'm trying to quit smoking.
0: That's great that you're taking this important step. Quitting smoking is one of the best things you can do for your health.
1: I know, but it's so hard. I've tried to quit before, but I always end up going back to it...

(a) Unlabeled dialogue

Addiction Specialist: Hello, the national Quitline. How can I help you today?
Client2302: Hi, I'm calling because I'm trying to quit smoking.
Addiction Specialist: That's great that you're taking this important step. Quitting smoking is one of the best things you can do for your health.
Client2302: I know, but it's so hard. I've tried to quit before, but I always end up going back to it...

(b) Labeled dialogue

Figure 3.10: Comparison of unlabeled and labeled dialogues

3.6.1 A heuristic approach to the identification of addiction specialist

In order to correctly identify addiction specialists and calling clients, leveraging the ability of Beey to distinguish speakers in the records, a heuristic approach was employed. Observation revealed that the majority of addiction specialists introduce themselves with the phrase "Dobrý den, národní linka pro odvikaní...", which can be translate to "Hello, The National Quitline..." in English.

To address this, a script was developed to utilize the first two lines of the dialog to determine the speaker's identity. Specifically, the script aims to locate the phrase "Narodni linka pro odvikani", abbreviated as *NLPO*[6], within the dialog. Due to potential inaccuracies in transcription, Levenshtein distance was employed to measure

[6]Narodni Linka Pro Odvikani

the similarity between the extracted phrase and the intended phrase. If the Levenshtein distance between the extracted sequence of four words in the first two lines of the dialog and the *NLPO* phrase is less than 13, the extracted sequence is considered to be the *NLPO* phrase. Otherwise, the script skips speaker identification. This approach achieved a speaker identification rate of 81.3% across 4288 transcriptions. To verify the effectiveness of this approach, a sample of 84 transcriptions was randomly selected, and no error was detected on beginning of dialogues.

It is imperative to highlight that while the initial approach to identifying addiction specialists in the opening lines demonstrates a certain level of effectiveness, it became apparent that Beey, encountered challenges in consistently tracking speakers. In specific dialogues, instances arose where the addiction specialist was initially indexed at position 0, and the client at position 1; however, this ordering was reversed towards the conclusion of the text, with the addiction specialist indexed at 1 and the client at 0. The extent of this issue remains challenging to quantify without a comprehensive examination of the entire dataset, involving manual identification and rectification of records displaying such discrepancies.

During the analysis of transcriptions, this issue was observed on two occasions. Notably, both instances coincided with a commonality in the low quality of the audio file. In one instance, the complexity was compounded by the involvement of more than two speakers. The resolution of this matter necessitates careful scrutiny of the dataset and a discerning assessment of records presenting such inconsistencies.

3.6.2 A LLM approach to the identification of addiction specialist

For the remaining 872 unlabelled records, a GPT-4 model, specifically GPT-4 Basic with an input context of 8,192 tokens, was employed. The start of each conversation was presented to the model, and it was tasked with identifying the index of the corresponding addiction specialist. The effectiveness of this approach was evaluated using a sample of 33 records, yielding a lower accuracy rate compared to the heuristic approach, with 5 instances of misidentification. Nevertheless, this error rate was deemed acceptable and sufficient for the purpose of speaker classification. Alternatively, manual review of all 872 records could have been conducted; however, this option was deemed impractical due to the substantial number of records.

3.6.3 Clients identification

The identification process for clients within the transcription did not necessitate a specialized methodology. A straightforward premise guided this classification: the assumption that an addiction specialist exclusively engages in conversation with the client. Consequently, if a speaker was not identified in the preceding step as an addiction specialist, that individual was classified as the client by default. This uncomplicated rule allowed for the differentiation between the roles of addiction specialists and clients within the dialogue content.

It is customary for clients to make multiple calls to the Quitline. Leveraging metadata contained within file names, specifically pertaining to clients' phone numbers, enabled the maintenance of continuity in client-related information across these successive calls. Through the extraction of phone numbers embedded within file names, a systematic procedure was employed to assign each user a randomized client identifier, following a straightforward naming convention of "Client" accompanied by a random numerical designation. Consistently, all records associated with a particular phone number were linked to the same assigned client name, thereby ensuring coherence and traceability across multiple interactions stemming from the same client.

3.6.4 Limitations of Beey speaker recognition

Figure 3.11 delineates the distribution of identified speakers within the dialogues, revealing that, as per the transcription analysis, 53.5% of dialogues featured two speakers, while 37.6% involved three speakers, and a minority, 5.8%, comprised monologues with just one speaker. However, these findings diverge from our firsthand experience with phone calls on the Quitline.

Observations from actual calls on the Quitline contradict the representation in the data. Monologues consisting of only one person were infrequent. Typically, the calls constituted dialogues between two individuals, a pattern notably different from the indicated 5.8% of single-speaker dialogues. Moreover, while instances of three-speaker dialogues were identified in the transcriptions, they were disproportionately represented at 37%, conflicting with our subjective experiences. Upon scrutinizing these alleged three-speaker dialogues in the audio records, we consistently recognized only two speakers. Furthermore, records featuring four or more speakers were also noted.

In light of these discrepancies, decisions were made to rectify the identified issues. Firstly, dialogues featuring one or more than four speakers were excluded from the speaker identification results. Secondly, addressing the challenge posed by three-speaker dialogues, the approach involved merging non-specialist speakers following the identification of the addiction specialist. While acknowledging the limitations of this strategy, we presumed that errors were more likely to occur in recognizing the client's role, given that specialists operate within a controlled calling center environment. Common issues encountered in phone calls typically stem from noise or low-quality audio from the client's end.

3.7 Anonymization

3.7.1 Introduction

The practical part of this master thesis contains the anonymization of unstructured text data. Anonymization of text is a two-step process that involves recognizing attributes in the text and then using anonymization techniques on the recognized

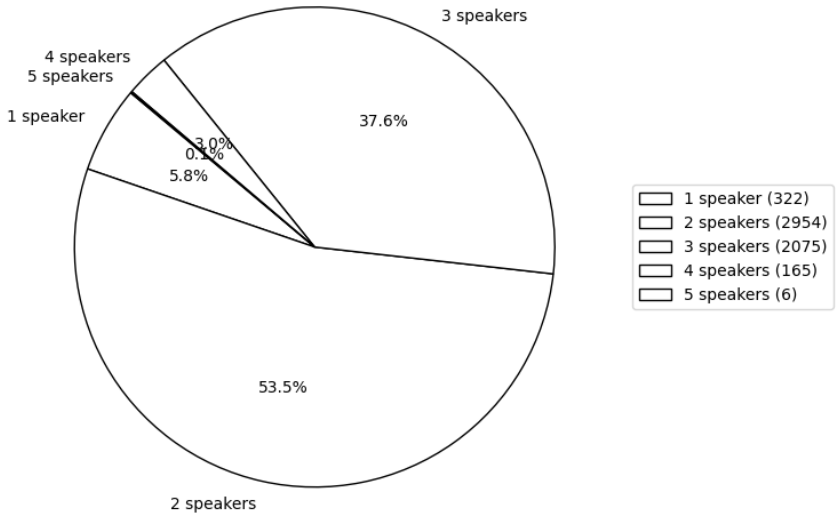


Figure 3.11: The distribution of number of speakers in the dialogues by Beey

entities, similar to structured data. We can convert the first step of attribute recognition to a Named Entity Recognition problem. As the primary goal of this thesis is not to implement our solution from scratch, we have decided to utilize existing solutions for NER. The NER is a common problem in ML[7], so many models are trained on different datasets. This section will discuss our tool selection process and how we applied anonymization techniques to existing NER models to achieve our goal.

3.7.2 NER models selection

In this study, an investigation into existing NER models for both English and Czech languages were conducted. The primary repository surveyed for these models was the Hugging Face platform, renowned for its extensive collection of models and libraries catering to diverse tasks such as text classification, token classification, question answering, translation, summarization, and conversational text generation. Notably, Hugging Face's Transformers library stands out for its specialized design aimed at Natural Language Processing applications. Additionally, complementary research encompassed exploration beyond the Hugging Face platform, encompassing open-source libraries like SpaCy. It is imperative to note that comparing these models is not straightforward due to the diverse nature of information provided in their respective documentations. This variance in available documentation necessitates a nuanced approach to comparative analysis.

[7]Machine Learning

3.7.3 NER models for Czech

DeepPavlov

The DeepPavlov model constitutes a BERT-based architecture that underwent training on the Slavic dataset BSNLP-2019 [96]. This model exhibits proficiency not only in NER for the Czech language but also extends its capabilities to encompass recognition functionalities for Russian, Polish, and Bulgarian languages. Evaluation metrics furnished by the developers of DeepPavlov reveal commendable performance metrics, showcasing an accuracy rate of 94.3%, a recall rate of 93.4%, and an Recognition Precision Match rate of 93.9% [96]. These metrics underscore the model's robustness in accurately identifying named entities across multiple Slavic languages, establishing its viability for diverse NER applications within this linguistic domain.

GPT models - ChatGPT-3.5/ChatGPT-4

Despite its primary design for conversational and generative tasks, ChatGPT-3.5/4 showcases promising adaptability for NER in the Czech language domain. Leveraging its comprehensive understanding of contextual nuances, semantic relations, and linguistic patterns, the model demonstrates a capacity to identify and classify named entities, encompassing entities like individuals, locations, organisations, and more, within the Czech text.

NameTag 2

NameTag 2 represents an open-source tool designed for NER, specializing in the identification of proper names within textual data and their categorization into predefined classes, encompassing a spectrum of entities, including individuals, geographical locations, organizations, among others. A pivotal characteristic distinguishing NameTag 2 pertains to its adeptness in recognizing nested entities of diverse complexities within textual contexts, showcasing a flexibility surpassing mere flat entity recognition [97].

Recent evaluations have underscored NameTag 2's exceptional performance across diverse linguistic corpora. Notably, it has exhibited remarkable proficiency in Czech (CNEC 2.0 corpus [98]), English (CoNLL corpus [99]), Dutch (CoNLL corpus), and Spanish (CoNLL corpus), demonstrating state-of-the-art capabilities [97].

NameTag 2 leverages the potent capabilities of BERT contextualized word embeddings, thereby augmenting its prowess in NER tasks. This tool demonstrates the proficiency to identify both flat and nested named entities, accommodating variable depths contingent upon the intricacies of the trained model. For user accessibility, NameTag 2 offers accessibility through the NameTag Web Application and NameTag REST Web Service, both hosted by LINDAT/CLARIN, facilitating ease of access and utilization [100].

small-e-czech-finetuned-ner-wikiann

In this study, the "small-e-czech-finetuned-ner-wikiann" model is explored, presenting a fine-tuned adaptation of the Seznam/small-e-czech model specifically tailored for NER on the wikiann dataset [101]. This finely tuned variant showcases its performance through the evaluation metrics obtained on the respective evaluation set [102].

The reported metrics for this fine-tuned model on the evaluation set include a loss of 0.2547, precision at 0.8713, recall at 0.8970, F1 score of 0.8840, and an accuracy of 0.9557. These metrics indicate the model's proficiency in accurately identifying named entities within the context of the wikiann dataset [102].

The Seznam/small-e-czech model is an Electra-small model, initially pretrained on a Czech web corpus developed at the Czech company Seznam.cz [103]. Primarily intended for general language understanding, this model requires fine-tuning on specific downstream tasks, including NER, to realize its full potential. At Seznam.cz, its deployment has notably enhanced web search ranking, rectified query typos, and detected clickbait titles, underscoring its utility in diverse applications. The release of this model is facilitated under the CC BY 4.0 license, enabling commercial utilization, and encouraging community engagement through issue reporting via their GitHub repository [104].

3.7.4 Benchmarking of NER models for Czech

The preceding compilation of models delineates those under consideration within the scope of this research for NER tasks specific to the Czech language. Notably, almost all these models share a common architectural foundation based on the BERT transformers. However, a key challenge arises from the fact that each model is trained on disparate datasets and possesses documentation with varying levels of coherence.

This divergence in training data and documentation impedes achieving an objective comparison, particularly when relying solely on publicly available information. In response to this challenge, we have undertaken a comprehensive comparative analysis, employing these diverse models on text samples selected from transcripted records. The primary objective is to assess their performance based on accuracy metrics, thereby contributing to a more informed understanding of their efficacy in Czech NER tasks. This approach seeks to mitigate the discrepancies arising from inconsistent documentation and disparate training data, fostering a more rigorous and objective evaluation of the models under consideration.

Evaluation data

Nine transcriptions were meticulously chosen, and entities were manually identified and selected within these transcripts. This selection process aimed to encompass diverse entities and capture typical characteristics found within records. A methodical

approach was adopted in the selection of these transcriptions to ensure representation across various entity types and encapsulate typical record attributes.

Evaluation of results

The test samples were employed for NER using the tools delineated in Section 3.7.3. An assessment was conducted to quantify the instances where the identified entities precisely matched those manually evaluated. The resulting table, denoted as Table 3.3, illustrates the proportion of entities accurately recognised by the NER tools, delineating cases where the words were correctly classified into their respective categories. This quantitative analysis provides a comprehensive overview of the precision of entity recognition, shedding light on the efficacy of the NER tools in accurately categorising and identifying entities within the test dataset.

Model	Precision	Recall	F1 Score
DeepPavlov	89.9	88.0	88.9
ChatGPT-3.5	81.5	84.6	83.0
NameTag 2	83.4	87.2	85.2
Small-e	76.6	80.9	78.7

Table 3.3: Result of benchmarking of models on our testing data

Conclusion

According to the findings illustrated in Figure 3.3, the decision was made to employ the DeepPavlov model as the NER model within our anonymization script. DeepPavlov demonstrated superior performance.

3.7.5 NER models for English

In comparing the current state of NER research, it is evident that the English NER is a far more extensively studied problem than its Czech counterpart. The literature features a plethora of models dedicated to the English language, highlighting the significant interest and effort that has been invested in this area. The ensuing chapter provides an overview of a relatively small subset of these models that have been deemed relevant to the scope of our study.

BERT multilingual base model and fine-tuned derivations

A decision was made to explore the application of fine-tuned BERT models as introduced in the paper authored by Jacob Devlin and colleagues [81]. These models vary based on the datasets utilised for their fine-tuning process. The following table 3.4 presents a comprehensive list of the distinct models and the corresponding datasets upon which they have been trained.

Model Name	Training Data
bert-base-multilingual-cased-ner-hrl[105]	Conll 2003
WikiNEuRal[106]	WikiNEuRal
SpanMarker[107]	MultiNERD

Table 3.4: Models and Training Data

Deeppavlov

Please refer to the 3.7.3 section for further information. It is noteworthy that DeepPavlov extends its support for the English language through the inclusion of the "ner_ontonotes_bert_torch" model.

SpaCy

SpaCy is an open-source software library designed for advanced natural language processing, developed using Python and Cython programming languages[108]. Established by Matthew Honnibal and Ines Montani, the co-founders of the software company Explosion, this library operates under the MIT license [109].

In contrast to NLTK[8], which primarily serves educational and research purposes, SpaCy prioritises delivering software tailored for production applications [110]. Notably, it facilitates deep learning workflows, enabling the integration of statistical models from prevalent machine learning frameworks such as TensorFlow, PyTorch, or MXNet via its proprietary machine learning library Thinc [111][112]. For users, the availability of prebuilt statistical neural network models spans 23 languages while supporting tokenization for over 65 languages, empowering the creation of custom models based on individual datasets [113].

Stanza

The Stanford NLP Group has developed Stanza, a sophisticated toolkit for analysing and comprehending text across a multitude of languages. Its core functionalities encompass tokenisation, part-of-speech tagging, named entity recognition, sentiment analysis, and coreference resolution. Its compatibility with over 23 languages and open-source nature make it an alluring choice for researchers and practitioners alike[114].

[8]Natural Language Toolkit

3.7.6 Benchmarking of NER models for English

Evaluation data

To assess English NER models, we opted to employ identical text files as utilised in the Czech evaluation, albeit following translation into English. This approach ensured a comparable testing environment across both language contexts, facilitating a standardised evaluation process for the NER models under scrutiny.

Evaluation

Model	Precision	Recall	F1 Score
bert-base-multilingual-cased-ner-hrl	85.6	86.9	86.3
ChatGPT-3.5	89.5	85.6	87.5
SpanMarker	73.1	71.0	72.0
Deeppavlov	87.2	88.4	87.8
SpaCy	91.1	89.5	90.3
Stanza	86.9	86.8	86.8
WikiNEuRal	87.5	84.2	85.8

Table 3.5: Evaluation results for English NER models

Conclusion

Based on the outcomes obtained from our conducted test, as outlined in Table 3.5, the decision has been made to implement Spacy in tandem with WikiNER as our chosen NER model for the English language. In certain instances, it was observed that Wikineural successfully identified entities that Spacy failed to recognize.

3.7.7 Anonymization script

After thoroughly researching NER models for both Czech and English languages, we devised an anonymization scripts that utilizes the aforementioned NER models. The primary objective of this scripts were to replace sensitive entities with their corresponding entity types. Notably, as the anonymization process was a one-time operation, we did not incorporate specific optimizations aimed at enhancing performance. Separate Python programs were created for the Czech and English languages.

To execute this anonymization process, we leveraged the computing resources available within the faculty infrastructure, utilizing GPU capabilities for efficient processing. This approach allowed us to efficiently run the anonymization script without specifically optimizing for enhanced performance, given the nature of it being a singular operation.

In Appendices A and B, an anonymization script is presented, showcasing the generation of dialogues using Google's Bard model. This demonstration aims to elucidate prevalent patterns within the records. Our methodology involves substituting sensitive information with entity tags identified by a NER tool. Notably, distinct NER models are employed for each language, each employing unique NER tags.

Conclusion

The present master's thesis undertakes a comprehensive exploration into the processing of dialogue data within the domain of addiction counseling practice. Its core objectives encompass the transformation of audio recordings from addiction specialist-client consultations into textual transcripts, ensuring GDPR-compliant anonymization, and preprocessing the converted data for incorporation into a language conversation system. The thesis commenced by delving into the requisite theoretical underpinnings fundamental to achieving the defined objectives, particularly focusing on speech-to-text conversion and the anonymization of unstructured data.

Subsequently, the implementation phase rigorously scrutinized extant solutions catering to speech-to-text conversion, encompassing both open-source and commercial avenues, with specific attention to Czech language compatibility. The findings unequivocally establish commercial speech-to-text tools as the optimal choice, particularly noteworthy for their efficacy within the Czech language context. Equally pivotal was the identification and categorization of speakers, which formed a crucial facet of this investigation.

A significant segment of this work pertained to devising an approach for the anonymization of unstructured text files, primarily through the application of entity recognition techniques. This entailed an exhaustive exploration of Named Entity Recognition tools available in both Czech and English languages. Despite concerted efforts to systematically benchmark these tools against our dataset, it was evident that the availability of NER tools for the Czech language remained considerably constrained.

The culmination of these efforts culminates in the generation of processed data derived from addiction practice sessions, meticulously prepared to facilitate its seamless integration into further research within the realm of Natural Language Processing. This thesis not only consolidates theoretical frameworks but also furnishes practical methodologies crucial for the advancement of addiction counseling data processing and its subsequent utilization within NLP research paradigms.

Enhancement Proposals and Future Directions

Certainly, here's an academic-style rewrite suggesting enhancements and future directions for the identified challenges within the master's thesis:

The encountered challenges during the course of this study have revealed multi-faceted complexities that extend beyond the immediate scope of this research endeavor. It is evident that resolving these challenges necessitates collective efforts involving interdisciplinary groups of scientists and researchers. Notably, the issue of speech-to-text conversion for the Czech language stands prominently among these challenges, indicating a reliance on commercial entities due to the absence of freely available competent solutions in the current market landscape. Assessing the feasibility of transitioning these commercial solutions into freely accessible resources remains a challenging task, primarily due to limited access to proprietary information held by these companies, thereby impeding a comprehensive evaluation of the possibility of making such tools available to a wider audience.

Another significant challenge pertains to the inadequately explored domain of NER for the Czech language. Presently, the major hindrance lies in the scarcity of relevant datasets tailored for NER tasks in this linguistic context. The manual annotation of entities within such datasets is an immensely time-consuming process. However, with the advent and proliferation of Language Model Models, there exists a prospect of streamlining and expediting this annotation process compared to methods employed in previous years. Leveraging these advancements in LLMs could potentially mitigate the challenges associated with the scarcity of annotated datasets, offering a more feasible and efficient approach to NER in the Czech language landscape. This avenue warrants further exploration and experimentation, given its potential to significantly enhance and expedite the development of NER capabilities within the Czech language domain.

Bibliography

1. *Národní linka pro odvykání*. 2023. Available also from: `https://chciodvykat.cz/o-nas/`.
2. REDDY, D Raj. Speech recognition by machine: A review. *Proceedings of the IEEE*. 1976, roč. 64, č. 4, pp. 501–531.
3. RABINER, Lawrence; JUANG, Biinghwang. An introduction to hidden Markov models. *ieee assp magazine*. 1986, roč. 3, č. 1, pp. 4–16.
4. NAG, R; WONG, K; FALLSIDE, Frank. Script recognition using hidden Markov models. In: *ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1986, sv. 11, pp. 2071–2074.
5. KROGH, Anders; MIAN, I Saira; HAUSSLER, David. A hidden Markov model that finds genes in E. coli DNA. *Nucleic acids research*. 1994, roč. 22, č. 22, pp. 4768–4778.
6. KUPIEC, Julian. Robust part-of-speech tagging using a hidden Markov model. *Computer speech & language*. 1992, roč. 6, č. 3, pp. 225–242.
7. FINE, Shai; SINGER, Yoram; TISHBY, Naftali. The hierarchical hidden Markov model: Analysis and applications. *Machine learning*. 1998, roč. 32, pp. 41–62.
8. RABINER, Lawrence R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*. 1989, roč. 77, č. 2, pp. 257–286.
9. GALES, Mark; YOUNG, Steve et al. The application of hidden Markov models in speech recognition. *Foundations and Trends® in Signal Processing*. 2008, roč. 1, č. 3, pp. 195–304.
10. AWAD, Mariette; KHANNA, Rahul; AWAD, Mariette; KHANNA, Rahul. Hidden markov model. *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. 2015, pp. 81–104.
11. WAIBEL, Alexander; HANAZAWA, Toshiyuki; HINTON, Geoffrey; SHIKANO, Kiyohiro; LANG, Kevin J. Phoneme recognition using time-delay neural networks. In: *Backpropagation*. Psychology Press, 2013, pp. 35–61.
12. BIRD, Jordan J; WANNER, Elizabeth; EKÁRT, Anikó; FARIA, Diego R. Optimisation of phonetic aware speech recognition through multi-objective evolutionary algorithms. *Expert Systems with Applications*. 2020, roč. 153, p. 113402.

13. WU, Jianxiong; CHAN, Chorkin. Isolated word recognition by neural network models with cross-correlation coefficients for speech dynamics. *IEEE transactions on pattern analysis and machine intelligence*. 1993, roč. 15, č. 11, pp. 1174–1185.

14. ZAHORIAN, Stephen A; ZIMMER, A Matthew; MENG, Fansheng. Vowel classification for computer-based visual feedback for speech training for the hearing impaired. In: *INTERSPEECH*. Citeseer, 2002.

15. HU, Hongbing; ZAHORIAN, Stephen A. Dimensionality reduction methods for HMM phonetic recognition. In: *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 4854–4857.

16. HOCHREITER, Sepp; SCHMIDHUBER, Jürgen. Long short-term memory. *Neural computation*. 1997, roč. 9, č. 8, pp. 1735–1780.

17. SAK, Haşim; SENIOR, Andrew; RAO, Kanishka; BEAUFAYS, Françoise; SCHALKWYK, Johan. Google voice search: faster and more accurate. *Google Research blog*. 2015.

18. FERNÁNDEZ, Santiago; GRAVES, Alex; SCHMIDHUBER, Jürgen. Sequence labelling in structured domains with hierarchical recurrent neural networks. In: *Proceedings of the 20th International Joint Conference on Artificial Intelligence, IJCAI 2007*. 2007.

19. WAIBEL, Alex. Modular construction of time-delay neural networks for speech recognition. *Neural computation*. 1989, roč. 1, č. 1, pp. 39–46.

20. GONG, Yuan; CHUNG, Yu-An; GLASS, James. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*. 2021.

21. BISHOP, Christopher M; NASRABADI, Nasser M. *Pattern recognition and machine learning*. Sv. 4. Springer, 2006. Č. 4.

22. BRAHME, Anders. *Comprehensive biomedical physics*. Newnes, 2014.

23. HINTON, Geoffrey; DENG, Li; YU, Dong; DAHL, George E; MOHAMED, Abdel-rahman; JAITLY, Navdeep; SENIOR, Andrew; VANHOUCKE, Vincent; NGUYEN, Patrick; SAINATH, Tara N et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*. 2012, roč. 29, č. 6, pp. 82–97.

24. DENG, Li; YU, Dong et al. Deep learning: methods and applications. *Foundations and trends® in signal processing*. 2014, roč. 7, č. 3–4, pp. 197–387.

25. DAHL, George E; YU, Dong; DENG, Li; ACERO, Alex. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on audio, speech, and language processing*. 2011, roč. 20, č. 1, pp. 30–42.

26. DENG, Li; LI, Jinyu; HUANG, Jui-Ting; YAO, Kaisheng; YU, Dong; SEIDE, Frank; SELTZER, Michael; ZWEIG, Geoff; HE, Xiaodong; WILLIAMS, Jason et al. Recent advances in deep learning for speech research at Microsoft. In: *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 8604–8608.

27. YU, Dong; DENG, Lin. *Automatic speech recognition*. Sv. 1. Springer, 2016.

28. DENG, Li; LI, Xiao. Machine learning paradigms for speech recognition: An overview. *IEEE Transactions on Audio, Speech, and Language Processing*. 2013, roč. 21, č. 5, pp. 1060–1089.

29. MAGALHÃES, Regis Pires; VASCONCELOS, Daniel Jean Rodrigues; FERNANDES, Guilherme Sales; CRUZ, Lívia Almada; SAMPAIO, Matheus Xavier; MACÊDO, José Antônio Fernandes de; SILVA, Ticiana Linhares Coelho da. Evaluation of Automatic Speech Recognition Approaches. *Journal of Information and Data Management*. 2022, roč. 13, č. 3.

30. READ-COOP. *Character Error Rate (CER) - Transkribus Glossary*. 2023. Available also from: `https://readcoop.eu/glossary/character-error-rate-cer/`.

31. DALMIA, Swaraj. *Evaluating an ASR in a Spoken Dialogue System*. 2022. Available also from: `https://tech.skit.ai/evaluating-an-asr-in-a-spoken-dialogue-system/`.

32. VAN LANCKER, Diana; KREIMAN, Jody; EMMOREY, Karen. Familiar voice recognition: Patterns and parameters part I: Recognition of backward voices. *Journal of phonetics*. 1985, roč. 13, č. 1, pp. 19–38.

33. FILE, Interinstitutional. Proposal for a Regulation of the European Parliament and of the Council on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free movement of Such Data (General Data Protection Regulation). *General Data Protection Regulation*. 2012.

34. 2023. Available also from: `https://www.dataprotection.ie/en/individuals/data-protection-basics/principles-data-protection`.

35. REGULATION, Protection. Regulation (EU) 2016/679 of the European Parliament and of the Council. *Regulation (eu)*. 2016, roč. 679, p. 2016.

36. SLOOT, Bart van der; SCHENDEL, Sascha van. Ten questions for future regulation of big data: A comparative and empirical legal study. *J. Intell. Prop. Info. Tech. & Elec. Com. L.* 2016, roč. 7, p. 110.

37. JONES, Chris. Data Protection, Immigration Enforcement and Fundamental Rights: What the EU's Regulations on Interoperability Mean for People with Irregular Status. 2019.

38. BOARD, European Data Protection. Guidelines 07/2020 on the concepts of controller and processor in the GDPR. *https://edpb.europa.eu*. 2020.

39. MULLINS, Martin; HOLLAND, Christopher P; CUNNEEN, Martin. Creating ethics guidelines for artificial intelligence and big data analytics customers: The case of the consumer European insurance market. *Patterns*. 2021, roč. 2, č. 10.

40. *Data Ethics in the Digital Age: Navigating the Ethical Waters of Data in Business*. 2023. Available also from: `https://www.linkedin.com/pulse/data-ethics-digital-age-navigating-ethical-waters-business`.

41. OHMANN, Christian; BANZI, Rita; CANHAM, Steve; BATTAGLIA, Serena; MATEI, Mihaela; ARIYO, Christopher; BECNEL, Lauren; BIERER, Barbara; BOWERS, Sarion; CLIVIO, Luca et al. Sharing and reuse of individual participant data from clinical trials: principles and recommendations. *BMJ open*. 2017, roč. 7, č. 12, e018647.

42. JASMONTAITE, Lina; KAMARA, Irene; ZANFIR-FORTUNA, Gabriela; LEUCCI, Stefano. Data protection by design and by default: Framing guiding principles into legal obligations in the GDPR. *Eur. Data Prot. L. Rev.* 2018, roč. 4, p. 168.

43. THOMAS, Llewellyn DW; LEIPONEN, Aija. Big data commercialization. *IEEE Engineering Management Review*. 2016, roč. 44, č. 2, pp. 74–90.

44. GROVER, Sandeep; SARKAR, Siddharth; GUPTA, Rahul. Data handling for e-mental health professionals. *Indian journal of psychological medicine*. 2020, roč. 42, č. 5_suppl, 85S–91S.

45. NITHYA, M; SHEELA, T. A Comparative Study on Privacy Preserving Datamining Techniques. *International Journal of Modern Engineering Research (IJMER)*. 2014, roč. 4, č. 7.

46. POOVAMMAL, E; PONNAVAIKKO, M. APPT: A privacy preserving transformation tool for micro data release. In: *Proceedings of the 1st Amrita ACM-W Celebration on Women in Computing in India*. 2010, pp. 1–8.

47. YANG, Lei; XUE, Hao; LI, Fengjun. Privacy-preserving data sharing in smart grid systems. In: *2014 IEEE International Conference on Smart Grid Communications (SmartGridComm)*. IEEE, 2014, pp. 878–883.

48. PATIL, Harsh Kupwade; SESHADRI, Ravi. Big data security and privacy issues in healthcare. In: *2014 IEEE international congress on big data*. IEEE, 2014, pp. 762–765.

49. MURTHY, Suntherasvaran; BAKAR, Asmidar Abu; RAHIM, Fiza Abdul; RAMLI, Ramona. A comparative study of data anonymization techniques. In: *2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing,(HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*. IEEE, 2019, pp. 306–309.

50. REN, Xiangmin; YANG, Jing. Research on privacy protection based on K-anonymity. In: *2010 International Conference on Biomedical Engineering and Computer Science*. IEEE, 2010, pp. 1–5.

51. OZALP, Ismet; GURSOY, Mehmet Emre; NERGIZ, Mehmet Ercan; SAYGIN, Yucel. Privacy-preserving publishing of hierarchical data. *ACM Transactions on Privacy and Security (TOPS)*. 2016, roč. 19, č. 3, pp. 1–29.

52. KUMAR, Atul; GYANCHANDANI, Manasi; JAIN, Priyank. A comparative review of privacy preservation techniques in data publishing. In: *2018 2nd International Conference on Inventive Systems and Control (ICISC)*. IEEE, 2018, pp. 1027–1032.

53. HOEPMAN, Jaap-Henk. Privacy design strategies. In: *IFIP International Information Security Conference*. Springer, 2014, pp. 446–459.

54. JAYABALAN, Manoj; RANA, Muhammad Ehsan. Anonymizing healthcare records: a study of privacy preserving data publishing techniques. *Advanced Science Letters*. 2018, roč. 24, č. 3, pp. 1694–1697.

55. WEITZENBOECK, Emily M; LISON, Pierre; CYNDECKA, Malgorzata; LANGFORD, Malcolm. The GDPR and unstructured data: is anonymization possible? *International Data Privacy Law*. 2022, roč. 12, č. 3, pp. 184–206.

56. HASSAN, Fadi; DOMINGO-FERRER, Josep; SORIA-COMAS, Jordi. Anonymization of unstructured data via named-entity recognition. In: *Modeling Decisions for Artificial Intelligence: 15th International Conference, MDAI 2018, Mallorca, Spain, October 15–18, 2018, Proceedings 15*. Springer, 2018, pp. 296–305.

57. RAJ, Anushree; D'SOUZA, Rio. Anonymization of sensitive data in unstructured documents using NLP. *International Journal of Mechanical Engineering and Technology (IJMET)*. 2021, roč. 12, č. 4, pp. 25–35.

58. NADEAU, David; SEKINE, Satoshi. A survey of named entity recognition and classification. *Lingvisticae Investigationes*. 2007, roč. 30, č. 1, pp. 3–26.

59. LI, Jing; SUN, Aixin; HAN, Jianglei; LI, Chenliang. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*. 2020, roč. 34, č. 1, pp. 50–70.

60. GRISHMAN, Ralph; SUNDHEIM, Beth M. Message understanding conference-6: A brief history. In: *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*. 1996.

61. SHARNAGAT, Rahul. Named entity recognition: A literature survey. *Center For Indian Language Technology*. 2014, pp. 1–27.

62. LAMPLE, Guillaume; BALLESTEROS, Miguel; SUBRAMANIAN, Sandeep; KAWAKAMI, Kazuya; DYER, Chris. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*. 2016.

63. LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep learning. *nature*. 2015, roč. 521, č. 7553, pp. 436–444.

64. DUPOND, Samuel. A thorough review on the current advance of neural network structures. *Annual Reviews in Control*. 2019, roč. 14, č. 14, pp. 200–230.

65. ABIODUN, Oludare Isaac; JANTAN, Aman; OMOLARA, Abiodun Esther; DADA, Kemi Victoria; MOHAMED, Nachaat AbdElatif; ARSHAD, Humaira. State-of-the-art in artificial neural network applications: A survey. *Heliyon*. 2018, roč. 4, č. 11.

66. TEALAB, Ahmed. Time series forecasting using artificial neural networks methodologies: A systematic review. *Future Computing and Informatics Journal*. 2018, roč. 3, č. 2, pp. 334–340.

67. SAK, Hasim; SENIOR, Andrew W; BEAUFAYS, Françoise. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. 2014.

68. LI, Xiangang; WU, Xihong. Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition. In: *2015 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2015, pp. 4520–4524.

69. WAN, Eric Andrew. *Finite impulse response neural networks with applications in time series prediction*. stanford university, 1994.

70. MEDSKER, Larry R; JAIN, LC. Recurrent neural networks. *Design and Applications*. 2001, roč. 5, č. 64-67, p. 2.

71. SUTSKEVER, Ilya. *Training recurrent neural networks*. University of Toronto Toronto, ON, Canada, 2013.

72. VENKATESAN, Ragav; LI, Baoxin. *Convolutional neural networks in visual computing: a concise guide*. CRC Press, 2017.

73. BALAS, Valentina E; KUMAR, Raghvendra; SRIVASTAVA, Rajshree et al. *Recent trends and advances in artificial intelligence and internet of things*. Springer, 2020.

74. CHEN, Hui; LIN, Zijia; DING, Guiguang; LOU, Jianguang; ZHANG, Yusen; KARLSSON, Borje. GRN: Gated relation network to enhance convolutional neural network for named entity recognition. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2019, sv. 33, pp. 6236–6243. Č. 01.

75. VASWANI, Ashish; SHAZEER, Noam; PARMAR, Niki; USZKOREIT, Jakob; JONES, Llion; GOMEZ, Aidan N; KAISER, Łukasz; POLOSUKHIN, Illia. Attention is all you need.< i> Advances in neural information processing systems</i>, 30. *Curran Associates, Inc*. 2017, pp. 5998–6008.

76. LIU, Peter J; SALEH, Mohammad; POT, Etienne; GOODRICH, Ben; SEPASSI, Ryan; KAISER, Lukasz; SHAZEER, Noam. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*. 2018.

77. KITAEV, Nikita; KLEIN, Dan. Constituency parsing with a self-attentive encoder. *arXiv preprint arXiv:1805.01052*. 2018.

78. LIU, Tianyu; YAO, Jin-Ge; LIN, Chin-Yew. Towards improving neural named entity recognition with gazetteers. In: *Proceedings of the 57th annual meeting of the association for computational linguistics*. 2019, pp. 5301–5307.

79. JIE, Zhanming; LU, Wei. Dependency-guided LSTM-CRF for named entity recognition. *arXiv preprint arXiv:1909.10148*. 2019.

80. XIA, Congying; ZHANG, Chenwei; YANG, Tao; LI, Yaliang; DU, Nan; WU, Xian; FAN, Wei; MA, Fenglong; YU, Philip. Multi-grained named entity recognition. *arXiv preprint arXiv:1906.08449*. 2019.

81. DEVLIN, Jacob; CHANG, Ming-Wei; LEE, Kenton; TOUTANOVA, Kristina. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. 2018.

82. ROGERS, Anna; KOVALEVA, Olga; RUMSHISKY, Anna. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*. 2021, roč. 8, pp. 842–866.

83. WU, Yonghui; SCHUSTER, Mike; CHEN, Zhifeng; LE, Quoc V; NOROUZI, Mohammad; MACHEREY, Wolfgang; KRIKUN, Maxim; CAO, Yuan; GAO, Qin; MACHEREY, Klaus et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*. 2016.

84. DENG, Li. A tutorial survey of architectures, algorithms, and applications for deep learning. *APSIPA transactions on Signal and Information Processing*. 2014, roč. 3, e2.

85. NAIK, AMIT RAJA. Google Introduces New Architecture To Reduce Cost Of Transformers. 2021.

86. RADFORD, Alec; NARASIMHAN, Karthik; SALIMANS, Tim; SUTSKEVER, Ilya et al. Improving language understanding by generative pre-training. 2018.

87. MICROSOFT. *Custom Named Entity Recognition (NER) Evaluation Metrics* [[Online]. Available: `https://learn.microsoft.com/en-us/azure/ai-services/language-service/custom-named-entity-recognition/concepts/evaluation-metrics`]. [B.r.]. Last accessed 2023-12-13.

88. PATRIK, Jankuv. *Backend mobilnıch aplikácie pro adiktologii*. 2021. B.S. thesis. České vysoké učenı technické v Praze. Vypočetnı a informačnı centrum.

89. ORGANIZATION, World Health. *Global status report on alcohol and health 2018*. World Health Organization, 2019.

90. TECHNOLOGIES, Newton. *Automatic transcription and subtitles for your audio or video content*. 2023. Available also from: `https://www.beey.io/en/`.

91. *Speech-to-Text: Automatic Speech Recognition*. Google, 2023. Available also from: `https://cloud.google.com/speech-to-text`.

92. *Sonix | API Documentation*. 2023. Available also from: `https://sonix.ai/docs/api`.

93. *Sonix*. 2023. Available also from: `https://sonix.ai/`.

94. *VOSK Offline Speech Recognition API*. 2023. Available also from: `https://alphacephei.com/vosk/`.

95. *Whisper*. 2023. Available also from: `https://openai.com/research/whisper`.

96. ARKHIPOV, Mikhail; TROFIMOVA, Maria; KURATOV, Yurii; SOROKIN, Alexey. Tuning multilingual transformers for language-specific named entity recognition. In: *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*. 2019, pp. 89–93.

97. STRAKOVÁ, Jana; STRAKA, Milan; HAJIČ, Jan. Neural architectures for nested NER through linearization. *arXiv preprint arXiv:1908.06926*. 2019.

98. STRAKOVA, Jana; STRAKA, Milan; SEVCIKOVA, Magda; ŽABOKRTSKÝ, Zdeněk. Czech named entity corpus. *Handbook of Linguistic Annotation*. 2017, pp. 855–873.

99. SANG, Erik F; DE MEULDER, Fien. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*. 2003.

100. STRÁKOVÁ, Jana. *NameTag 2* [[Online]. Available: `https://ufal.mff.cuni.cz/nametag/2`]. 2019. Last accessed 2023-12-11.

101. FACE, Hugging. *WikiANN dataset* [[Online]. Available: `https://huggingface.co/datasets/wikiann`]. 2023. Last accessed 2023-12-11.

102. RICHIELO, Riccardo. *Small-e-Czech-finetuned-ner-wikiann* [[Online]. Available: `https://huggingface.co/richielo/small-e-czech-finetuned-ner-wikiann`]. 2023. Last accessed 2023-12-11.

103. KOCIAN, Matej; NAPLAVA, Jakub; STANCL, Daniel; KADLEC, Vladimir. Siamese bert-based model for web search relevance ranking evaluated on a new czech dataset. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2022, sv. 36, pp. 12369–12377. Č. 11.

104. SEZNAM. *Small-e-Czech* [[Online]. Available: `https://huggingface.co/Seznam/small-e-czech`]. 2023. Last accessed 2023-12-11.

105. DAVLAN. *bert-base-multilingual-cased-ner-hrl* [[Online]. Available: `https://huggingface.co/Davlan/bert-base-multilingual-cased-ner-hrl`]. [B.r.]. Last accessed 2023-12-13.

106. BABELSCAPE. *Babelscape/wikineural-multilingual-ner* [[Online]. Available: `https://huggingface.co/Babelscape/wikineural-multilingual-ner`]. [B.r.]. Last accessed 2023-12-13.

107. TOMAARSEN. *Span Marker Model for NER (Multilingual NERD)* [[Online]. Available: `https://huggingface.co/tomaarsen/span-marker-mbert-base-multinerd`]. [B.r.]. Last accessed 2023-12-13.

108. HONNIBAL, Matthew; MONTANI, Ines. *Introducing spaCy* [[Blog post]. Available: `https://explosion.ai/blog/introducing-spacy`]. Explosion AI, 2015. Last accessed 2023-12-11.

109. CHOI, Jinho D; TETREAULT, Joel; STENT, Amanda. It depends: Dependency parser comparison using a web-based evaluation tool. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2015, pp. 387–396.

110. SPACY. *spaCy: Facts and Figures* [[Online]. Available: `https://spacy.io/usage/facts-figures`]. [B.r.]. Last accessed 2023-12-11.

111. THINC. *Thinc: Interoperability with Machine Learning Frameworks* [[Online]. Available: `https://thinc.ai/docs/usage-frameworks`]. [B.r.]. Last accessed 2023-12-13.

112. AI, Explosion. *Thinc: A lightweight deep learning library* [[Online]. Available: `https://github.com/explosion/thinc`]. Explosion AI, 2023. Last accessed 2023-12-13.

113. SPACY. *Models and Languages* [[Online]. Available: `https://spacy.io/usage/models`]. [B.r.]. Last accessed 2023-12-13.

114. GROUP, Stanford NLP. *Named Entity Recognition* [[Online]. Available: `https://stanfordnlp.github.io/stanza/ner.html`]. Stanford NLP Group, 2023. Last accessed 2023-12-13.

Appendix

A English anonymisation demonstration

Before anonymization

Addiction Specialist: Thank you for calling the National Quitline. How can I help you today?
Client0078: I'm having problems with alcohol and I'm looking for some help.
Addiction Specialist: I understand that you're going through a tough time. It's great that you're reaching out for help.
Client0078: Yes, I'm from Kladno and I've been drinking heavily for several years. It's starting to have a negative impact on my life.
Addiction Specialist: I'm sorry to hear that. How is alcohol affecting your life?
Client0078: Well, for one thing, my wife, Tereza, is constantly mad at me because of my drinking. And I'm also starting to have problems at work.
Addiction Specialist: I can see why you're concerned. Alcohol can have a significant impact on relationships, work, and overall well-being.
Client0078: Exactly. That's why I'm determined to make a change.
Addiction Specialist: That's a positive step. There are many resources available to help you with alcohol dependence.
Client0078: I know. But I'm not sure where to start.
Addiction Specialist: There are several clinics in Kladno and Prague that specialize in alcohol treatment. I can provide you with some phone numbers if you'd like.
Client0078: That would be great. Thank you.
Addiction Specialist: You're welcome. Here are the phone numbers for a few clinics in Kladno: 723 345 134, The clinic is on Borůvkova 22.
Client0078: Thank you for that. Are there any clinics in Prague as well?
Addiction Specialist: Yes, here are a few clinics in Prague: on Vaclavske namesti or on Prague 6, on street Evropská 78. Here is number 5 5 4 1 2 3 4 2 1
Client0078: I'm glad to know that there are so many options.
Addiction Specialist: It's important to find a clinic that you feel comfortable with and that offers the services you need.
Client0078: I agree.
Addiction Specialist: Would you like to schedule another call with us to check your progress?
Client0078: Yes, I would. How about next Monday, March 17th, at 5 pm?
Addiction Specialist: I'm afraid we're only open until 5 pm, so we might not be able to take your call then. Would you prefer 4 pm instead?
Client0078: Sure, 4 pm is fine.
Addiction Specialist: Great. We'll look forward to hearing from you then.
Client0078: Thank you for your help.

After anonymization

Addiction Specialist: Thank you for calling [ORG]. How can I help you [DATE]?
Client0078: I'm having problems with alcohol and I'm looking for some help.
Addiction Specialist: I understand that you're going through a tough time. It's great that you're reaching out for help.
Client0078: Yes, I'm from [GPE] and I've been drinking heavily for [DATE]. It's starting to have a negative impact on my life.
Addiction Specialist: I'm sorry to hear that. How is alcohol affecting your life?
Client0078: Well, for [CARDINAL] thing, my wife, [ORG][LOC], is constantly mad at me because of my drinking. And I'm also starting to have problems at work.
Addiction Specialist: I can see why you're concerned. Alcohol can have a significant impact on relationships, work, and overall well-being.
Client0078: Exactly. That's why I'm determined to make a change.
Addiction Specialist: That's a positive step. There are many resources available to help you with alcohol dependence.
Client0078: I know. But I'm not sure where to start.
Addiction Specialist: There are several clinics in [GPE] and [GPE] that specialize in alcohol treatment. I can provide you with some phone numbers if you'd like.
Client0078: That would be great. Thank you.
Addiction Specialist: You're welcome. Here are the phone numbers for a few clinics in [GPE]: [CARDINAL] [CARDINAL] [CARDINAL], The clinic is on [DATE].
Client0078: Thank you for that. Are there any clinics in [GPE] as well?
Addiction Specialist: Yes, here are a few clinics in [GPE]: on [ORG] namesti or on [GPE] [CARDINAL], on street [PERSON] 78. Here is number [CARDINAL] 5 4 1 2 3 4 2 1
Client0078: I'm glad to know that there are so many options.
Addiction Specialist: It's important to find a clinic that you feel comfortable with and that offers the services you need.
Client0078: I agree.
Addiction Specialist: Would you like to schedule another call with us to check your progress?
Client0078: Yes, I would. How about [DATE], at [TIME]?
Addiction Specialist: I'm afraid we're only open until [TIME], so we might not be able to take your call then. Would you prefer [TIME] instead?
Client0078: Sure, [TIME] is fine.
Addiction Specialist: Great. We'll look forward to hearing from you then.
Client0078: Thank you for your help.

B Czech anonymisation demonstration

Before anonymization

Adiktolog: Děkuji, že voláte, zde Národní linka pro odvikání. Jak vám mohu dnes pomoci?
Client0078: Mám problémy s alkoholem a hledám pomoc.
Adiktolog: Chápu, že procházíte těžkým obdobím. Je skvělé, že se obracíte na pomoc.
Client0078: Ano, jsem z Kladna a několik let jsem pil velmi těžce. Začíná to mít negativní dopad na můj život.
Adiktolog: Je mi líto, to slyšet. Jak alkohol ovlivňuje váš život?
Client0078: No, mimo jiné je moje žena Teréza neustále naštvaná kvůli mému pití. A také začínám mít problémy v práci.
Adiktolog: Chápu, proč jste znepokojeni. Alkohol může mít významný dopad na vztahy, práci a celkové blaho.
Client0078: Právě tak. Proto jsem odhodlán něco změnit.
Adiktolog: To je pozitivní krok. Existuje mnoho zdrojů, které vám mohou pomoci s alkoholismem.
Client0078: Vím. Ale nejsem si jistý, kde začít.
Adiktolog: Existují několik klinik v Kladně a Praze, které se specializují na léčbu alkoholu. Mohu vám poskytnout nějaké telefonní čísla, pokud chcete.
Client0078: To by bylo skvělé. Děkuji ti.
Adiktolog: Rád vám pomůžu. Zde jsou telefonní čísla pro několik klinik v Kladně: 723 345 134, klinika je na Borůvkove 22.
Client0078: Děkuji za to. Jsou tam nějaké kliniky i v Praze?
Adiktolog: Ano, tady je pár klinik v Praze: na Vaclavskem namesti nebo na Praze 6, na Evropské ulici 78. Zde je číslo 5 5 4 1 2 3 4 2 1.
Client0078: Jsem rád, že existuje tolik možností.
Adiktolog: Je důležité najít kliniku, která se vám líbí a která nabízí služby, které potřebujete.
Client0078: Souhlasím.
Adiktolog: Chtěli byste si naplánovat další hovor s námi, abychom si zkontrolovali váš pokrok?
Client0078: Ano, chtěl bych. Jak by to bylo příští pondělí 17. března v 5 hodin odpoledne?
Adiktolog: Bojím se, že máme otevřeno jen do 17:00, takže bychom vám v tu dobu možná nemohli zavolat. Raději byste měli 4 odpoledne?
Client0078: Jistě, 4 odpoledne je v pořádku.
Adiktolog: Skvělé. Těšíme se na vás slyšet pak.
Client0078: Děkuji za pomoc.

After anonymization

Adiktolog: Děkuji, že jste volal [ORG]. Jak vám mohu dnes pomoci?
Client0078: Mám problémy s alkoholem a hledám pomoc.
Adiktolog: Chápu, že procházíte těžkým obdobím. Je skvělé, že se obracíte na pomoc.
Client0078: Ano, jsem z [GPE] a [DATE] jsem pil velmi těžce. Začíná to mít negativní dopad na můj život.
Adiktolog: Je mi líto, to slyšet. Jak alkohol ovlivňuje váš život?
Client0078: No, mimo jiné je moje žena [PERSON] neustále naštvaná kvůli mému pití. A také začínám mít problémy v práci.
Adiktolog: Chápu, proč jste znepokojeni. Alkohol může mít významný dopad na vztahy, práci a celkové blaho.
Client0078: Právě tak. Proto jsem odhodlán něco změnit.
Adiktolog: To je pozitivní krok. Existuje mnoho zdrojů, které vám mohou pomoci s alkoholismem.
Client0078: [GPE]. Ale nejsem si jistý, kde začít.
Adiktolog: Existují několik klinik [GPE] a [GPE], které se specializují na léčbu alkoholu. Mohu vám poskytnout nějaké telefonní čísla, pokud chcete.
Client0078: To by bylo skvělé. Děkuji ti.
Adiktolog: Rád vám pomůžu. Zde jsou telefonní čísla pro několik klinik v [GPE]: 723 345 134, klinika je na [GPE] 22.
Client0078: Děkuji za to. Jsou tam nějaké kliniky i v [GPE]?
Adiktolog: Ano, tady je pár klinik v [GPE]: na [FAC] namesti nebo na [FAC], na [FAC]. Zde je číslo 5 5 4 1 2 3 4 2 1.
Client0078: Jsem rád, že existuje tolik možností.
Adiktolog: Je důležité najít kliniku, která se vám líbí a která nabízí služby, které potřebujete.
Client0078: Souhlasím.
Adiktolog: Chtěli byste si naplánovat další hovor s námi, abychom si zkontrolovali váš pokrok?
Client0078: Ano, chtěl bych. Jak by to bylo [DATE] v [TIME]?
Adiktolog: Bojím se, že máme otevřeno jen do [TIME], takže bychom vám v tu dobu možná nemohli zavolat. Raději byste měli 4 odpoledne?
Client0078: Jistě, 4 odpoledne je v pořádku.
Adiktolog: [PERSON]. Těšíme se na vás slyšet pak.
Client0078: Děkuji za pomoc.