BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF FOREIGN LANGUAGES
ÚSTAV JAZYKŮ

NATURAL LANGUAGE PROCESSING: ANALYSIS OF INFORMATION TECHNOLOGY STUDENTS' SPOKEN LANGUAGE
ZPRACOVÁNÍ PŘIROZENÉHO JAZYKA: ANALÝZA MLUVENÉHO JAZYKA STUDENTŮ OBORU INFORMAČNÍ TECHNOLOGIE

BACHELOR'S THESIS
BAKALÁŘSKÁ PRÁCE

AUTHOR — **Aleksandar Stankovic**
AUTOR PRÁCE

SUPERVISOR — **Mgr. Ing. Eva Ellederová, Ph.D.**
VEDOUCÍ PRÁCE

BRNO 2021

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Bakalářská práce

bakalářský studijní program **Angličtina v elektrotechnice a informatice**
obor Angličtina v elektrotechnice a informatice
Ústav jazyků

Student: Aleksandar Stanković ***ID:*** 185914
Ročník: 3 ***Akademický rok:*** 2020/21

NÁZEV TÉMATU:

Zpracování přirozeného jazyka: analýza mluveného jazyka studentů oboru informační technologie

POKYNY PRO VYPRACOVÁNÍ:

Analyzujte mluvený jazyk studentů oboru informační technologie prostřednictvím metody zpracování přirozeného jazyka.

DOPORUČENÁ LITERATURA:

1) Dale, R., Moisl, H., & Somers, H. (2000). Handbook of natural language processing. New York: Marcel Dekker, Inc.
2) Lane, H., Hapke, H. M., & Howard, C. (2019). Natural language processing in action. Understanding, analyzing, and generating text with Python. Shelter Island: Manning Publications Co.
3) Beysolow II, T. (2018). Applied natural language processing with Python. New York. Apress.

Termín zadání: 31. 1. 2021 ***Termín odevzdání:*** 31. 5. 2021

Vedoucí práce: Mgr. Ing. Eva Ellederová, Ph.D.

doc. PhDr. Milena Krhutová, Ph.D.
předseda oborové rady

Abstract

This bachelor's thesis deals with the issue of new artificial intelligence technologies in natural language processing. The thesis consists of a theoretical part and an analytical part. The theoretical part approaches the issue by dividing it into three chapters: artificial intelligence and statistics, natural language processing, and IBM Watson Natural Language Understanding. Each of these chapters is elaborated on by using at least one example from the real world. In the first chapter, the main aim is to frame the theoretical framework of artificial intelligence and its practices, while in the second chapter, natural language processing and its primary functions are explained as well as its relation to artificial intelligence itself. The aim of the third chapter is to introduce natural language understanding as the primary tool for analysis which is done in the analytical part. The analytical part deals with the analysis of students' spoken language using various methods. Collected video samples are transcribed by means of a machine translator as a natural language processing application, while the textual output is analysed through a natural language understanding engine. The applied knowledge from the theoretical part is used in the analytical part that includes the description of research methodology, presentation and interpretation of research results.

Key words

artificial intelligence, new technologies, natural language processing, IBM, data, analysis, score, application, Kubernetes, IBM Cognos Analytics, natural language understanding

Abstrakt

Tato bakalářská práce se zabývá problematikou nových technologií umělé inteligence při zpracování přirozeného jazyka. Práce je rozdělena na teoretickou a analytickou část. Teoretická část přistupuje k problému rozdělením do tří kapitol: umělá inteligence a statistika, zpracování přirozeného jazyka a IBM Watson Natural Language Understanding. Každá z těchto kapitol je rozpracována včetně uvedení alespoň jednoho příkladu z praxe. V první kapitole je hlavním cílem vymezit teoretický rámec umělé inteligence a jejích postupů, zatímco ve druhé kapitole je vysvětlena problematika zpracování přirozeného jazyka a jeho primární funkce včetně jeho vztahu k samotné umělé inteligenci. Cílem třetí kapitoly je představit porozumění přirozenému jazyku jako primární nástroj pro analýzu, která je realizována v analytické části práce. Analytická část se zabývá analýzou mluveného jazyka studentů prostřednictvím různých metod. Transkripce shromážděných vzorků videí je provedena strojovým překladem jako aplikací zpracování přirozeného jazyka, zatímco textový výstup je analyzován prostřednictvím nástroje porozumění přirozenému jazyku. V analytické části, která popisuje výzkumnou metodologii, prezentuje a interpretuje výsledky výzkumu, jsou využívány aplikované znalosti z teoretické části práce.

Klíčová slova

umělá inteligence, nové technologie, zpracování přirozeného jazyka, IBM, data, analýza, skóre, aplikace, Kubernetes, IBM Cognos Analytics, porozumění přirozenému jazyku

Stankovic, A. (2021). *Zpracování přirozeného jazyka: analýza mluveného jazyka studentů oboru informační technologie*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií. 55 s.
Vedoucí bakalářské práce: Mgr. Ing. Eva Ellederová, Ph.D.

Prohlášení

Prohlašuji, že bakalářskou práci na téma *Zpracování přirozeného jazyka: analýza mluveného jazyka studentů oboru informační technologie* jsem vypracoval samostatně pod vedením vedoucí bakalářské práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené bakalářské práce dále prohlašuji, že v souvislosti s vytvořením této práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědom následků porušení ustanovení § 11 a následujících zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

V Brně dne 31.5. 2021 .....

Aleksandar Stankovic

Acknowledgements

I would like to express gratitude to my supervisor Mgr. Ing. Eva Ellederová, Ph.D., and I would like to thank her for the effort, trust, and support during the writing of my thesis. Also, I would like to thank for the help of my managers Dalibor Zavacky and Ondřej Zmund for their support in my career path and IBM. Finally, I would like to express my gratitude to my brother and my mother. Without their tremendous understanding and encouragement in the past few years, it would be impossible for me to complete my study.

Table of Contents

Introduction

Not many decades ago, as technology started to develop rapidly, the demand for data exploration significantly increased. Corporations, governments, and other institutions struggled through the 20th century because at that time computational technologies were not developed enough from the performance perspective. To handle the data and analysis, a new discipline emerged, natural language processing (hereinafter referred to as NLP). NLP is a combination of linguistics, computer science, and artificial intelligence (hereinafter referred to as AI). Its main concern is the interaction between computers and human language (Dale, Moisl, & Somers, 2000).

Firstly, this bachelor's thesis will deal with how AI progressed over time, its adaptation to a modern society and its impact on the world of technologies. The chapters will lead through the various disciplines and applications of AI. Machine learning, as the widely used discipline in the industry environment, will be tested in the educational sphere. Thus, the modern tools of AI should help future generations to overcome highly stressful tasks. On the other hand, the time invested in some research thanks to the tools from IBM Watson Cloud should reduce workload. The disadvantage of entrusting to AI is its unreliability. Models trained by AI have a scale of confidence which rates from 0 to 100%, thus this metric should be an indicator if the general user of the product should bring the business decision based on the score or if it should be omitted.

Secondly, the thesis will explore NLP and its concept. Each stage of its analysis will be explained, and in the analytical part of the thesis, a machine translator and IBM Watson Natural Language Understanding application will be used to analyse 53 sample videos of persuasive presentations made by students. The NLP tools are open source, but their use is limited to a certain extent. Research results will be used to evaluate the results of NLU performance and its predictions.

The main goal of this thesis is to integrate my knowledge gained in the study programme focused on the language of electrical engineering and information technology and my experience with new emerging technologies of NLP.

THEORETICAL PART

1 Artificial Intelligence and Statistics

Inspired by reverse engineering, AI is designed to simulate operations in the human brain or the way how the neurons work. Our brain consists of small units called neurons and the cluster of neurons is called the neural network. Furthermore, AI models are considered as the outputs, designed to support intelligent solutions (Zsolt, 2019; Vibhor, May 26, 2018). AI as a discipline was founded in 1955 (Press, December 30, 2016). However, the hardware and software capabilities were weak, so it was impossible to make independent robots. The optimism for developing AI has made a comeback a couple of decades later. Design and modelling of AI have a very strong connection with other disciplines which are outside the AI spectrum such as (Zsolt, 2019):

1) robotics – manipulation of objects in space;
2) theory of algorithms – designing the effective algorithms;
3) statistics – core discipline, computations of results, future predictions, and analysis of the past;
4) psychology – fundamentals of human brain operations;
5) software engineering – the creation of efficient software;
6) software programming – implementation of the application in real-time;
7) mathematics – advance computations;
8) control theory – feed-forward and feedback systems modelling;
9) information theory – encoding and decoding;
10) graph-theory – modelling and optimization of the parameters;
11) physics – real-world modelling;
12) GPU – image and video processing.

According to Zsolt (2019), all the above-mentioned disciplines have contributed to the fundamental knowledge to establish the AI discipline.

1.1 Subfields of AI

To distinguish the disciplines within AI, scientists have defined six sub-fields of AI where each sub-field has its own specialists who conduct research and developments. These subfields (see Figure 1) are mostly combined to create an ideal AI solution for the business. Recently, companies have been the primary drivers of AI development. The usual implementation of these AI solutions is observed in data science, healthcare, customer experience, and security (Zsolt, 2019).

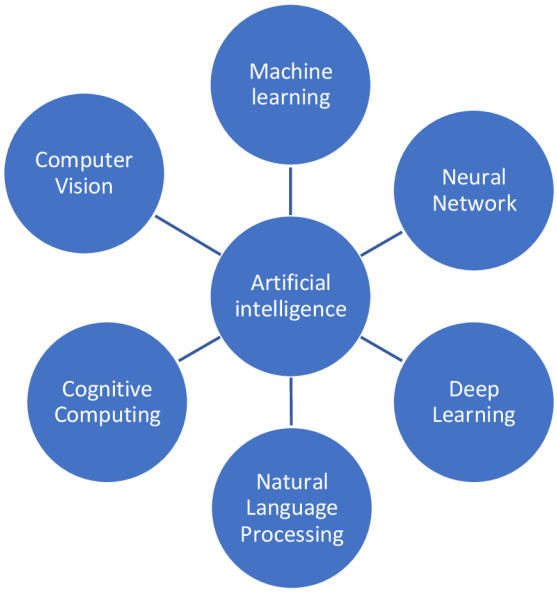


Figure 1. Subfields of AI. Adapted from Software Testing Help (November 13, 2020).

1.1.1 Machine Learning

Machine learning has been used for a long time and it is considered the most common feature of AI in business. Machine learning operations technology and practices largely depend on the gathering of raw data and its capability to learn from different cases. The goal of machine learning models is to predict the output. Zsolt (2019) and Software Testing Help (November 13, 2020) define pattern recognition as the subcategory of machine learning that helps to automate the recognition of the blueprint of the raw data by using computer algorithms. Examples of the pattern could be a persistent series of data over time which is used for forecasting trends or sequence events, identifying objects, recurring a combination of words and sentences, and others.

According to Software Testing Help (November 13, 2020), there are five steps to clarify pattern recognition (see the Figure 2 below):

1) data acquisition – a collection of raw data from databases;
2) pre-processing of input data – filters are applied for removing ~~of~~ noise in data or unwanted variables;
3) feature extraction – various algorithms are put into action to acquire the most accurate one;
4) classification – based on the output of the algorithm, the class is assigned to the pattern;
5) post-processing – the values of the output are assured with the needed standardized result.

Figure 2. Pattern recognition. Adapted from Software Testing Help (November 13, 2020).

1.1.2 Deep Learning

In comparison to machine learning, deep learning is considered the most advanced and most precise subfield. Deep learning, as Schmidhuber (2015) describes, is a series of supervised, semi-supervised, and unsupervised machine learning methods using the deep neural network. Zsolt (2019) defines a deep neural network as a neural network that has often more than two layers and it uses special mathematical approaches to create a deep learning model. The networks try to automatically extract feature sets from the data, therefore, they are mostly used in data types where the feature selection process is e.g., when analysing unstructured datasets, such as image data, videos, sound, and text. As Agarwal (November 22, 2020) notes, in software development, the TensorFlow library is widely used to conduct pieces of training on deep learning models. TensorFlow is an open-source library created by Google which is available in many languages.

1.1.3 Neural Networks

So far, neural networks have been the latest subfield of AI. Inspired by the human brain,

scientists Warren McCulloch[1] and Walter Pitts[2] (1943) came up with an idea to mimic neurons and the way how the human brain resolves problems. The most fundamental unit of a deep *neural network* is an artificial neuron called *perceptron*[3]. The neural networks consist of an input layer where each perceptron has its weight, the hidden layer where the calculation is made, and an output layer where, in the end, the success of the process is verified. According to Agarwal (November 22, 2020), the most common types of neural networks architecture are:

1) *convolutional neural network* – deep learning algorithm designed mostly for image processing/recognition,
2) *recurrent neural network* – widely used in text processing and numerical computation,
3) *autoencoders* – the algorithm used to learn from data in an unsupervised manner.

1.1.4 Natural Language Processing

Zsolt (2019) points out that NLP is one of the most used sub-field in the business where the various AI models are used to create a chatbot, watchdogs, speech recognition software, translation applications, and many more. Due to the constant need for NLP applications, companies are achieving success when concerning NLP implementation. The precise goal of NLP applications in business is saving a budget and workload reduction.

1.1.5 Cognitive Computing

Cognitive computing comparing to AI has a different approach to problem resolving. While AI is mainly focused on finding the best algorithm, cognitive computing is mimicking human behaviour, human intelligence, and wisdom by analysing a series of factors (Makadia, May 10, 2019). David Kenny, the General Manager of IBM Watson technologies, states that "AI can only be as smart as the people teaching it." However, due to technical advantages in cognitive computing, the world will undergo a completely new experience

---

[1] Warren McCulloch was an American neurophysiologist and cybernetician who worked on the theory of the human brain.
[2] Walter Pitts was an American logician who worked in the field of computational neuroscience.
[3] The first model of a perceptron was developed in 1958 at the Cornell Aeronautical Laboratory by Frank Rosenblatt.

with technology. Therefore, to understand cognitive computing better, cognitive computing uses a blend of AI, neural networks, machine learning, NLP, sentiment analysis, and contextual awareness (Makadia, May 10, 2019). This is so far the most advanced system which learns at scale and its main purpose is interaction with humans in natural form.

1.1.6 Computer Vision

Often abbreviated as CV, computer vision is a subfield of AI whose main purpose is to enable computers to see. Nowadays, applications of CV can be found on any smartphone. Its main purpose is to distinguish objects from humans and recognize them. Also, the latest car producers are investing in the development of CV such as TESLA where their primary goal is to use the autopilot in their cars. IBM uses CV solutions for air companies to tackle the technical issues with aircraft by discovering the anomalies within the surface of the airplane. However, the companies and governments should respect the ethical rules for using CV technology since using cameras to detect faces and tracking people, such as by the Chinese government, is considered an unethical approach. CV is still imperfect, and it is crucial in the process of using algorithms and training data. Also, as Mihajlovic (April 25, 2019) observes, data poisoning in the AI sphere is equivalent to viruses in operating systems. Its purpose is to inject the false positives in CV application to cause malfunctions. Data poisoning is not only a problem of CV but also in many other subfields as well.

1.2 Statistics and probability

To build highly accurate AI models, Nabi (January 7, 2019) pointed out that data scientists with knowledge of advanced mathematics should use statistics and probability techniques to achieve that goal. *Descriptive statistics* is referred to as a method for summarizing and organizing information within a data set.

This means that we can have all the scores available to review and decide if the data could be used to train the machine learning model or not. If the data has missing values in the quantitative attribute section such as income, the missing values can be easily replaced with the average or mean of all numerical data within that attribute. Replacement is suggested in case of more than 30% missing values as the data will be biased and the accuracy score to train a model would be far worse than expected. Data consists of elements and variables.

Elements, also called “cases” or “subjects”, could be understood as entities for which information is collected, while variables, also known as “attributes”, are the characteristic of the element.

Nabi (January 7, 2019) defines *probability* in data science as an event or set of outcomes of an experiment to which a probability is assigned. If E stands for an event, then P can be understood as the probability that E will occur. A situation where E might happen or not is called a trial. For example, a student is attending an exam and there is a high chance of probability that the student will pass because he prepared well, however, if he did not prepare for an exam, that means that probability would drop significantly. Kane (2020) claims that the quality of data is the most important factor in the creation of AI models.

1.2.1 Fundamentals of Data Types

According to Kane’s (2020) explanation of the theory of the data types, there are three data types such as numerical, categorical, and ordinal. Numerical data represents some sort of quantitative measurement which could be the height of people, stock prices, etc. *Numerical data* has subcategories such as discrete data and continuous data. To clarify these subcategories and their meaning, discrete data is mostly integer-based. They often represent a certain number of events (for example, how many purchases have been made through a web shop), while continuous data has an infinite number of possible values. These kinds of data are usually organized into graphs for better understanding. The second type of data is categorical. *Categorical data* or *qualitative data* has no inherent mathematical meaning. Gender, yes/no (binary data), race, and state of residence are categorical data. Even if we assign numbers to these categories, we still have a categorical type of data, not numerical. The third and last data type is a mixture of numerical and categorical. Categorical data has a mathematical meaning, and they are different from the classic categorical type of data. Usually, *ordinal data* is used as the scores of the movies (e. g. movie ratings can vary from 1 to 5, and based on the average score, we can determine if the movie is good or not).

1.2.2 Descriptive Statistics

According to Mann (1995), descriptive statistics is used to describe the basic features of the data in a research study. It provides simple summaries about the sample and the observations that have been made. Together with simple graphics analysis, it forms the

basis of virtually every quantitative analysis of data. Descriptive statistics is used to present quantitative descriptions in a manageable form.

Nabi (January 7, 2019) defines the basic mathematical terms in descriptive statistics:

Data:

Elements – entities for which information is collected and stored.

Variables – characteristics of an element is called a variable.

Measures of centre:

Mean – an arithmetic average of a data set that usually belongs to the numerical type of data.

Median – a middle data value.

Mode – a data value that occurs with the greatest frequency. The Mode can occur in both numerical and categorical data.

Mid-range – an average of the maximum and minimum values in the data set.

Measures of variability:

Range – the difference between maximum and minimum values.

Variance – population variance is defined as the average of the squared differences from the mean.

Standard deviation – a bunch of numbers which indicates how much the individual numbers tend to differ from the mean.

Measures of position:

Percentile – the percentile of a data set is the data value such that percent of the values in data are represented in a range of 0 to 100%.

Percentile rank – a percentile rank of a data value equals the percentage of value in the data set that are at or below the value.

Interquartile range or IQR – measures the difference between quartile data.

Z–score – represents how many standard deviations the data value lies above or below the mean

After learning the fundamentals of mathematics audience or data analysts can see data in a more meaningful way. Types of a graph in use are histograms, pie charts, scatter plots, box plots, time series plots, etc.

1.2.3 Probability

Machine learning refers to making predictions. Kane (2020) explained that various types of machine learning models can be found across industries, healthcare, and retail services. In this case, predictions are outputs of the input data. These predictions are not to be taken into consideration as the real outcome, but at least to help give an idea about a possible solution to the user. For example, the bank offers you a loan, but some conditions must be met to get it. To predict whether the person has eligibility to take a loan or not, the special scoring system is capable based on the inputs of some values to provide an output. According to Kane (2020), the machine learning model called the *binary classification* is one of the most used models among the other applications of modelling and it is the best way to start with understanding probability. Fundamentals in mathematics is an important step to start with in order to understand the concepts of machine learning or AI in general.

Types of probability adapted from Nabi (January 7, 2019) are:

Empirical probability – the number of times an event occurs divided by the total number of the observed incidents.

Theoretical probability – given by the number of possibilities the particular event could occur divided by the total number of possible outcomes.

Joint probability – applies only when the values A and B are independent, which means A should not change the probability of B and vice versa.

Conditional Probability – applies only when the values A and B are not independent, which means that the best way is to compute the conditional probability.

1.2.4 Bayes' Theorem

Kane (2020) explains that Bayes' theorem, named after the 18th-century British mathematician Thomas Bayes, is a mathematical formula for determining conditional probability. We can say that conditional probability is the likelihood of an outcome occurring, based on a previous outcome occurring. Bayes' theorem provides a way to revise existing predictions or theories (updated probabilities) given new or additional evidence. This theorem is applied mostly in models for credit risk scoring.

The formula for Bayes' theorem adapted from Nabi's (January 7, 2019) research article is the following

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) * P(B|A)}{P(B)}$$

where

P(A) is the probability of A occurring

P(B) is the probability of B occurring

P(A|B) is the probability of A given B

P(B|A) is the probability of B given A

P(A∩ B) is the probability of both A and B occurring

Summary of Chapter 1

My experience with data analytics has brought me to the field of AI and machine learning. To achieve my objective, I have spent countless hours trying to understand logic behind AI. In this chapter, I included theoretical knowledge and experience that I had gained while working in the data analytics sector of IBM. The basic concepts of AI are crucial to understand in order to move on to the analytical part of this bachelor's thesis that will demonstrate NLP in action using data acquired from IT students' presentations. The next chapter will focus on the concept of NLP so that the reader could understand its processes and applications.

2 Natural Language Processing

NLP is considered the core discipline regarding the interpretation of data gathered by corporations, government, and other organizations. NLP itself is a discipline which has three backgrounds such as linguistics, computer science, and AI. These three backgrounds empower the data analysts to gain valuable insights by means of transforming ordinary text or unstructured data to structured. Through decades, many techniques have been developed and made contributions such as the voice in text processing, natural language translators, and chatbots.

Text mining and data fetching play an important role, especially in finding the most frequent terms and keywords where all the findings are used to improve the search applications. Linguistic rules are crucial in NLP. Nowadays, machines can define the semantic roles and syntactic functions in the text and determine whether the test is positive or negative if there is a terrorist threat and even create spam filters for emails or other communicational platforms. In the following chapters, the concept of NLP will be described in more detail.

2.1 Natural Language Processing: Stages of Analysis

The main applications are developed in the way that the desired input is transformed to the desired output, which means that the process of transformation has already established a convention where the text is assessed through five stages of analysis. Nowadays, the deep learning algorithms have already-defined variables, which has simplified the whole process, so the overall experience is much more accurate. The chart flow in Figure 3 illustrates the stages of analysis.

Figure 3. Stages of analysis of NLP. Adapted from Dale, Moisl and Somers (2000, p. 3).

2.1.1 Tokenization

We can understand a *tokenization process* (also known as *segmentation*) in NLP as breaking up the sequence of characters in a text into individual paragraphs, sentences, phrases, words (referred to as *tokens*), punctuation marks, and numbers (see e. g. Lane, Howard & Hapke,

2019; Dale et al., 2000). As Lane et al. (2019) note, a *tokenizer* (often called a *scanner* or *lexer*) "breaks unstructured data, natural language text, into chunks of information that can be counted as discrete elements" (p. 33). Only then the following procedure of determining whether the word is a verb, a noun, an adjective, etc., can be initiated.

2.1.2 Lexical Analysis

Lexical analysis or tokenization is the process of converting a sequence of characters into a sequence of tokens (strings with an assigned and thus identified meaning). Usually, this stage is considered computational and in computer programming, it can perform a statistical analysis and reveal keywords (Lane at al., 2019). Regarding the keywords, the most frequent technique is keyword counting where we can see the most used words within the text input. For example, social media or marketing departments use this technique of keyword counting to determine the word frequency and to reveal what the audience or users usually want.

2.1.3 Syntactic Analysis (Parsing)

This stage is usually related to the grammatical rules. NLP needs to be programmed in a way where it should recognize the grammatical structure. Therefore, there are a few possibilities. In machine learning applications, the most used small databases of examples are called *corpora*. A set of words with their rules allows the machine learning application to distinguish whether the word is a noun, a verb, or an adverb (Tutorials Point, 2019). The more data means much better accuracy. With large data sets, the application can recognize an ambiguous word, and depending on the content of the sentence, it can easily determine its true meaning. During syntactic analysis or parsing, the data structure is built in the form of a tree and it is interpreted from top to bottom and vice versa. According to Tutorials Point (2019), the main roles of parsing are:

- to report any syntax errors,
- to recover from commonly occurring error so that the processing of the remainder of the program can be continued,
- to create a parse tree,
- to create a symbol table,
- to produce intermediate representations (IR).

2.1.4 Semantic Analysis

Dale et al. (2000, p. 93) stress the importance of understanding an utterance, which is a complex process that depends on "the result of parsing, as well as on lexical information, context, and common-sense reasoning," and results in what they call *semantic interpretation in context*. The purpose of semantic analysis is to draw the exact meaning, or you can say the dictionary meaning from a text. Tutorials Point (2019) define the work of the *semantic analyzer* as checking the text for meaningfulness. Lexical analysis is based on smaller tokens while semantic analysis focuses on larger chunks. That is why semantic analysis can be divided into the following two parts (Turorials Point, 2019):

1) Studying the meaning of individual words: It is the first part of the semantic analysis in which the study of the meaning of individual words is performed. This part is called lexical semantics.

2) Studying the combination of individual words: In the second part, the individual words will be combined to provide meaning in sentences.

The most important task of semantic analysis is to get the proper meaning of the sentence. For example, analyse the sentence "Ram is great." In this sentence, the speaker is talking either about Lord Ram or about a person whose name is Ram. That is why the job, to get the proper meaning of the sentence, of the semantic analyzer is important. Figure 4 illustrates the flow of semantic analysis.

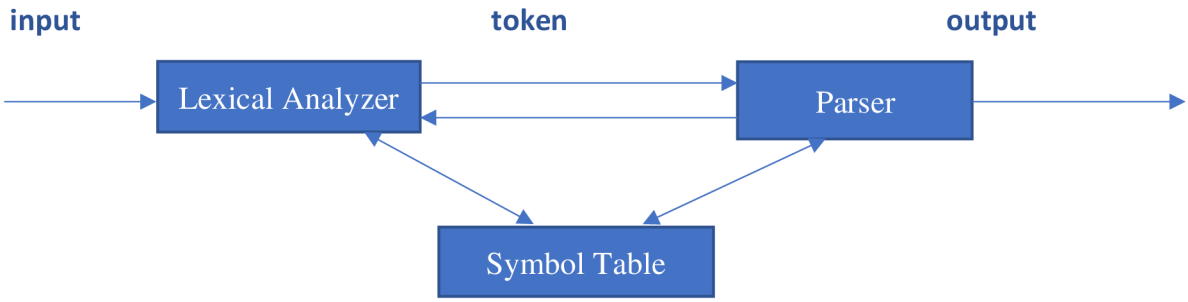


Figure 4. Semantic analysis. Adapted from Tutorials Point (2019).

2.1.5 Pragmatic Analysis

Pragmatic analysis is part of the process of extracting information from a text. Specifically, it is the stage that focuses on selecting a structured set of text and figuring out what its actual

meaning is. Pragmatic analysis is one of the objectives of pragmatics – the subfield of linguistics that studies how context contributes to meaning.

Many text meanings are related to the context in which it was said/written. Ambiguity, in particular limiting ambiguity, is at the core of NLP, so pragmatic analysis is quite crucial concerning extracting the meaning or information (Dale at al., 2000).

Pragmatic analysis that focuses on what was described is reinterpreted by what it meant, deriving the various aspects of language that require real-world knowledge.

2.2 Applications in Natural Language Processing

Thanks to the latest innovations, NLP applications can run independently as web applications where human interaction is not necessary.AI or deep learning algorithms are very helpful in terms of unsupervised learning. Lane et al. (2019) explain that unsupervised learning refers to the use of AI algorithms to identify patterns in data sets containing data points that are neither classified nor labelled. The algorithms are thus allowed to classify, label and/or group the data points contained within the data sets without having any external guidance (human interaction) in performing that task. There is a portion of data fed at the beginning of the process and occasionally an algorithm is capable of training itself and let AI learn from its own mistakes. Through continual learning, algorithms gradually improve and can be interpreted in a confidence score. According to Keckl (May 29, 2019), the *confidence score*, or *classification threshold*, indicates how certain the NLP Service/Machine Learning Model is that the respective intent has been assigned correctly. The score can have a value between 0 and 1, depending on how the neural networks work. In general, a score is calculated for each intent for each user input and the one with the highest value is returned as the result. If the confidence level falls below a pre-defined limit, a fallback intent is output.

For example, if you are searching for something on google such as “what is the weather today” it will give you a list of the answers, the most confident answer will be on the top. In mathematics, the scoring is understood through normalized numbers or the numbers from 0 to 1 (Sharma, July 8, 2020). In the very beginning, those applications were not possible to be trusted to unsupervised learning which means that the whole process has been done through the supervised method of learning. Supervised learning, human interaction all the time, and tunning of the parameters are of great importance

2.2.1 Sentiment Analysis

Sentiment analysis is considered one of the greatest achievements in the field of AI and NLP. The true meaning is the analysis of the speakers' feelings. Each word in the sentence has a score and based on these scores, we can calculate and determine whether the text is positive or negative through computation techniques. The range of the score is represented in normalized numbers and its range is from -1 to 1 or from 0 to 1. The computational formula to obtain the overall score is simplified and it should be understood as follows:

$$(P - N)/AP =$$
$$\text{Or}$$
$$((P + NN)-N)/AP =$$

P = positive words, N = negative words, AP = all words, NN = neutral words

Another important analytical approach to reveal the human's intention on the Internet is called intent analysis. Gupta (January 7, 2018) states: "Intent Analysis refers to the user's intention behind a message and identifies whether the message relates to an opinion, news, marketing, complaint, suggestion, appreciation, or query."

Gupta (January 7, 2018) also adds that sentiment analysis is very crucial for marketing departments that, based on this analysis, can determine whether they have positive or negative opinions about a product, mostly depending on an overall user's comments.

2.2.2 Chatbots

Large companies with many customers frequently face the inability to serve each of them if they have the increased workload within the customer's centre teams. For example, call centres are usually over-saturated with daily calls and requests. The users call these call centres just to verify some transactions or mainly to report minor problems. To overcome these daily problems, chatbots are one of the effective solutions and it has turned out that the careful implementation could help and replace the operators, so the operators can focus more on the medium or major problems. Chatbots or virtual assistants are designed to provide you with an answer without human interaction. It is usually programmed in a way that it recognizes the keywords and, depending on your question or answer, it can give you feedback. If a chatbot does not understand the question, it will try to navigate you to stay in the scope of the questions referring to the topic (Zia, March 7, 2018; IBM, 2019).

IBM (2019) report that one of the highly developed chatbots is TOBi from Vodafone empowered with the latest enhanced IBM Watson AI experience. Moreover, TOBi is the first chatbot capable of completing the transaction from the beginning to the end without human interaction.

2.2.3 Text Classification

Zia (2018) explains that text classification or text tagging can be understood as the categorization of a text into organized groups. He further remarks that NLP applications or text classifiers can automatically analyse the text and then assign pre-defined tags based on its content. Usually, text classifiers are focused on unstructured data such as emails, chat conversations, websites, and social media. To process all the information, a human would need time to extract the keywords, determine whether the text is scientific or not, and even determine which language is used. Thanks to the NLP text classifiers, this process is accelerated, and the user can analyse the key features of any document very quickly. Common examples of automated text classifiers are sentiment analysis, topic detection, and language detection. The following chapters will focus on text classification and its properties.

Later, the Natural Language Understanding Project by IBM Watson AI technology will be introduced.

2.2.4 Machine Translators

Machine translators automatically convert one natural language into another, and their best example is Google Translate. Machine translating is considered one of the oldest subfields of AI and the recent shifts towards large-scale empirical techniques have led to significant improvements in the quality of translation, which means less confusion about ambiguous words and more accurate translation. However, the problem with translation to every language will persist and definitely, it might be one of the disadvantages. Machine translators have two approaches to converting the input text to output. The first is a shallow approach where the application does not know the text (Machine Translation, 2012). The statistical methods are applied and based on the scores, the application will provide the output, while a deep approach has comprehensive knowledge of the word, but it is not as effective as the shallow approach.

2.2.5 Text Extraction

Text extraction or text mining is widely used in knowledge-driven organizations. Therefore, text mining is the process of examining large sets of documents and its main purpose is to investigate and discover new information or to help answer particular research questions. Text mining identifies facts, relationships, and assertions that would otherwise remain buried in the mass of big textual data. A powerful tool called Hadoop is capable of comprehending an extremely large set of data, identifying patterns, and converting them into a structured form (Laxmi, Kollati, & Amaranatha, January 2018). Once data conversion is carried out, it is much easier to conduct further analysis. The structured data created by text mining can be integrated into databases and different types of data analysis can be performed, such as descriptive, prescriptive, or predictive, which is mostly based on the demands of the companies (Kobayashi, Mol, & Berkers, August 2017).

Summary of Chapter 2

The aim of Chapter 2 was to explain how NLP works, describe its stages of analysis and applications. The analytical part of the bachelor's thesis will demonstrate its capabilities, in particular speech to text conversion. Besides, the machine learning model will be used to recognize the key words. Also, IBM Watson Natural Language Understanding[4] will prove the significant power of recognizing the text's concept, its sentiment, and syntactic roles. Then, the error rate will be measured to see if IBM Watson Natural Language Understanding is reliable to use for academic purposes.

[4] IBM Watson Natural Language Understanding is a cloud native product that uses deep learning to extract metadata from text such as entities, keywords, categories, sentiment, emotion, relations, and syntax.

3 IBM Watson Natural Language Understanding

This chapter will introduce IBM Watson Natural Language Understanding (hereinafter referred to as NLU), a powerful NLP engine for textual analytics. Empowered by deep learning algorithms, NLU has enhanced capabilities to extract metadata and provide the user with a useful insight. Big data driven companies face challenges and backlashes of constant growth of the data, especially textual data such as documents, transcripts, books, comments on social media and others. These extracts may help improve business operations, reduce the costs of analytics, safe time, and maintain better customer engagement.

Another advantage of NLU is the possibility of customizing text mining possibilities such as entity detection and classifications. Thanks to the IBM Watson Knowledge Studio[5], users can make their custom models and infuse it with NLU. These cases can be found, for example in engineering, healthcare, military. As the standard pre-trained product, NLU may not recognize specific discourse used in documents. The IBM research called Framework for Applying AI in Enterprise revealed the benefits of data science and machine learning, less than 50% have been deployed to production environment[6]. The slow development is because of insufficient data or the fact that about 85% of the business leaders improve data, while the remaining 15% have a proper use of data.

IBM Cloud (2017, February 27) infrastructure helps business clients to identify the pain points related to the proper use of data, data organization and text mining. Also, with this offer, there is a possibility of deployment to a live production framework. With a few of API references, generated with deployment, the user can integrate with their application and it can run engines of Watson within their domain. Scepticism about AI is everywhere, most problems might occur in the domain of training and testing machine learning models, especially in accuracy where some trained models can assign a positive meaning to the word “chaos”, thus confusing analysts, which inevitably results in additional tunning and retraining of the models.

NLU will be tested on 53 samples of presentations given by students of information technology (IT). The results should bring a conclusion and answer to the question whether

---

[5] IBM Watson Knowledge Studio is a cloud-based application that enables users to train Watson to understand the linguistic nuances of a specific industry domain.

[6] Production environment refers to a real-time setting where programs are run, and hardware setups are installed and relied on for organization or commercial daily operations.

AI could be reliable and helpful in education, in terms of saving time and workload reduction.

3.1 Principles of NLU Operation

In order to avoid complex coding of NLP scripts, IBM offers a unique framework where the input data are easily processed and delivered in the output form of JSON script[7]. In JSON script, the user can find valuable information about keywords and the confidence score. The confidence score is an indicator of certainty, where numbers are shown to the user in normalized output form. Normalized output is range from 0 to 1. There are three ways of input formats and those are a uniform resource locator or URL (must be publicly accessible), plain text and hypertext markup language (see Figure 5).

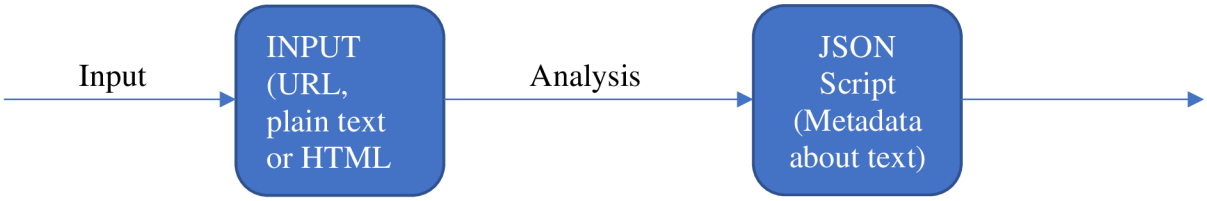


Figure 5. NLU Analysis. Adapted from Vergara et al. (2017, p. 2).

3.1.1 API Calls and Supported Programming Interfaces

The following setups are crucial for service to operate within a local device. NLU supports three basic programming interfaces such as Java, Python, and cURL[8]. Codes are organized in a manner where the user can actually trigger any function shown in documentation. In order to work, the user should have already obtained its credentials. These credentials are needed to be verified upon an application programming interface call (or an API call). The API call is a process where some functions are called to perform an operation. In this case, an NLU engine is called to perform an extraction of metadata from input data. Programming interfaces will perform an action and it will display an output in the command window.

---

[7] JSON is an abbreviation for JavaScript Object Notation used for storing information in an organized, easy-to-access manner.

[8] cURL is a command-line tool for getting or sending data including files using URL syntax.

3.1.2 Supported features

NLU consists of 10 unique features, where the user invokes commands through the API and executes the scripts through NLU, providing the desired output. These features are Sentiment, Semantic roles, Relations, Metadata, Keywords, Entities, Emotion, Concepts, and Categories. With these features, the user may have a more accurate picture of the analysed text, therefore, it could be employed to scan thousands and thousands text documents. Figure 6 illustrates how the text is processed with NLU and how the output is shown after the completion.

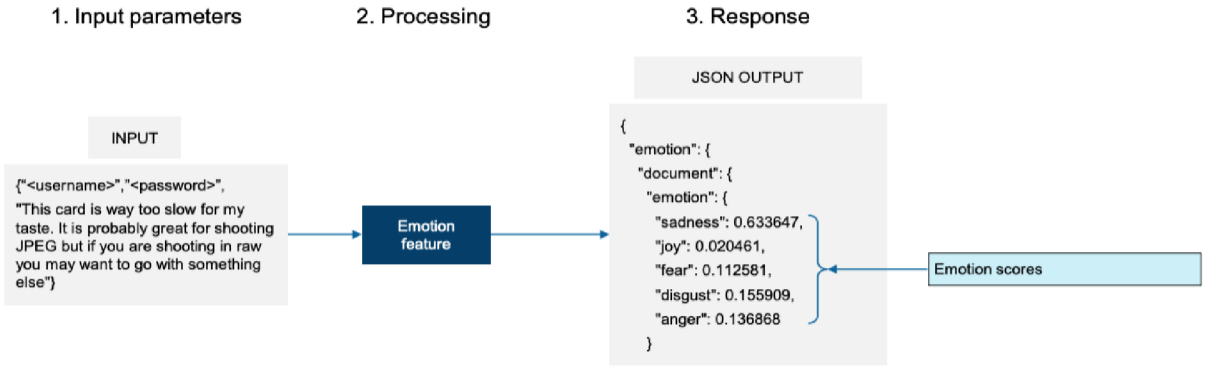


Figure 6. NLU emotion feature. Reprinted from Vergara et al. (2017, p. 16).

In this example, the plain text is inserted in quotas, so the NLU can recognize the primary target for analysis. After the execution, the JSON code is generated with emotion feature results. The confidence score of "sadness" is prevailing in results, so the reader can obviously realize that the person is dealing with dissatisfaction in a passive manner. In an analytical part of this bachelor's thesis, complete demonstration will be performed on all features and its functionality.

3.1.3 Integration into Third Party Apps

When it comes to integration, IBM Watson makes sure that the necessary codes are available to be used for integration into any framework. A user with a fundamental skill of CSS and HTML can design a framework where the raw input and output can be transformed into the application interface with text box and graphs. According to IBM Cloud (2017, February 27), there is a sample code which can be used for the integration. In this case, the primary

focus will be on the Python programming language and the demonstration will be shown in the example of the code snippet (see Figure 7).

```
pip install --upgrade "ibm-watson>=5.1.0".  #CODE FOR COMMAND LINE INTERFACE
```

```
import json
from ibm_watson import NaturalLanguageUnderstandingV1
from ibm_cloud_sdk_core.authenticators import IAMAuthenticator
from ibm_watson.natural_language_understanding_v1
    import Features, CategoriesOptions

authenticator = IAMAuthenticator('{apikey}')
natural_language_understanding = NaturalLanguageUnderstandingV1(
    version='2020-08-01',
    authenticator=authenticator
)

natural_language_understanding.set_service_url('{url}')

response = natural_language_understanding.analyze(
    url='www.ibm.com',
    features=Features(categories=CategoriesOptions(limit=3))).get_result()

print(json.dumps(response, indent=2))
```

Figure 7. API for Python. Adapted from IBM Cloud (2017, February 27).

An important step in this code snippet is installation of IBM Watson which is annotated with a hashtag. These are library setups for NLU to work within an environment. On the web page of IBM Cloud, the API key, a unique key given to the user upon registration, can be generated.

3.2 Blueprint of the NLU Project

NLU works with text only, therefore the conversion of speech to text will be done. The data will be in the textual unstructured form, and once the conversion ends, NLU can be used for the further analysis. Unfortunately, NLU is not capable of performing spoken language analysis, therefore conversion is unavoidable process.

3.2.1 Speech to Text Algorithm in Microsoft Word

Microsoft has recently released a new feature in their Microsoft Word product which is speech to text conversion. Empowered by deep learning algorithms, speech to text or dictation can be used within the Word application, thus it can help with accelerating the scripting process. The only requirements to use dictation is a functional microphone. For better performance, microphones with better quality should reduce noises in the background, and they are primarily focused on the incoming sound wave. Under these conditions, the Word application should be able to operate at a low error rate.

3.2.2 Blueprint

Blueprint will serve as a guide to the project described in the analytical part of the thesis. It is important to mention the phases within the project. In Figure 8, the phases are illustrated in the workflow scheme.

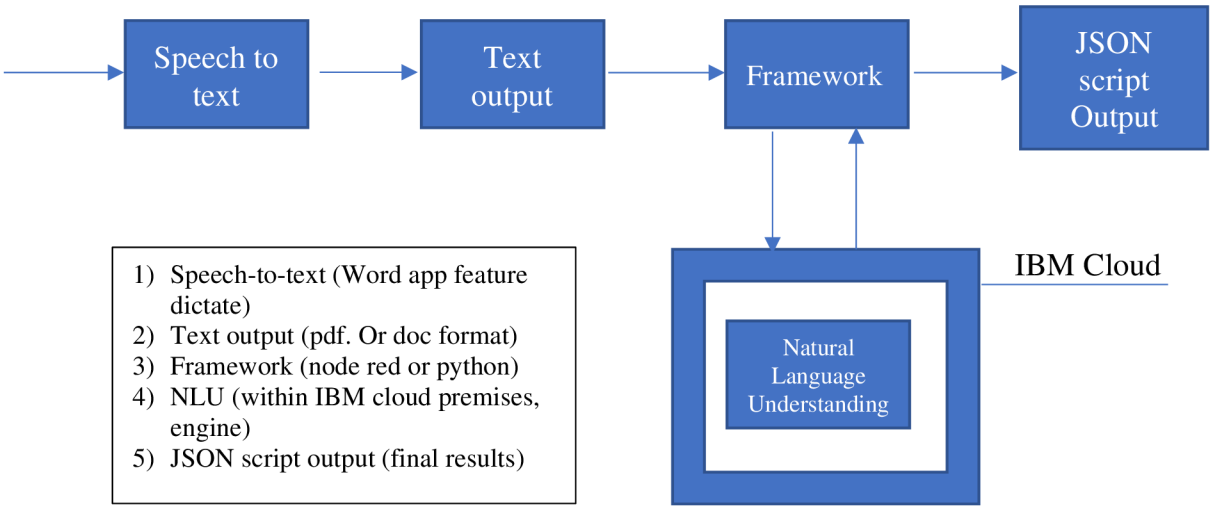


Figure 8. NLU workflow scheme. Adapted from IBM Cloud (2017, February 27).

Summary of Chapter 3

This chapter dealt with integrating the NLU engine into the framework for further progress of the bachelor's thesis project. First, the chapter described the product itself and examples where it can be found. The following subchapter discussed NLU in greater detail, in particular the API call function and communication between the user and the server. Besides, the example of functionality was given, with a sample text and output. The NLU emotion feature was demonstrated to see whether NLU could recognize the context of the text or not. Furthermore, the blueprint for the project and the analysis was drawn up. The blueprint should serve as a guide to the analysis of IT students' spoken language that will be described in the analytical part of this thesis.

ANALYTICAL PART

4 Introduction to the Analytical Part

The analytical part aims to give a detailed description of the analysis of the IT students' spoken language in their persuasive presentations on various IT products. This analysis was done by the NLP method where the metadata was extracted from the content of their presentations. Each video recording was transformed into textual data. A speech-to-text algorithm within the Word application was used to transcribe the recordings.

Chapters 4.1 and 4.2 will describe the scope of IT student's persuasive presentations, provide the number of samples collected and the overall length of the presentations used for NLP analysis. Chapter 5 will deal with the methodology and the results of the analysis. Finally, conclusions about the effectiveness of NLU will be drawn.

4.1 Description of the Research Sample

The research sample consisted of 53 videos of the students of the Faculty of Information Technology and the Faculty of Electrical Engineering and Communication at Brno University of Technology. The video samples were collected by Mgr. Ing. Eva Ellederova, Ph.D., a teacher of English for IT. The primary purpose of the assignment was to evaluate students' persuasive strategies used when making sales presentations for hardware or software products. The instructions were given to the students to help them make their presentations more efficiently.

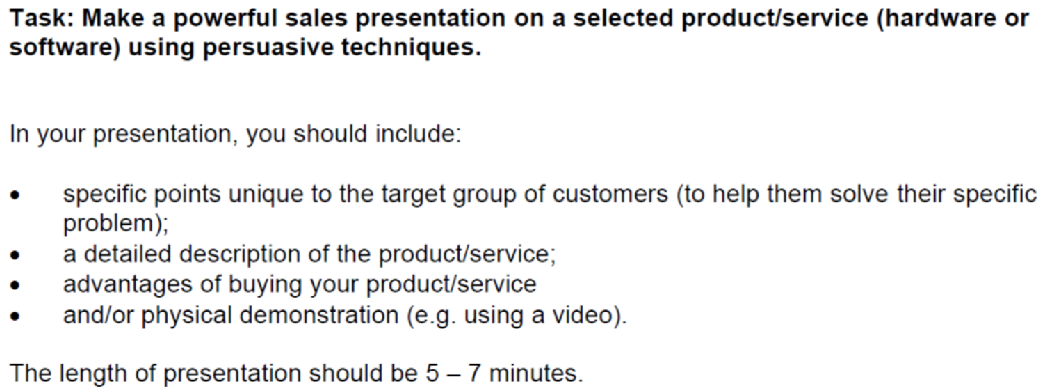
The videos were recorded by mutual agreement between the teacher who took part in the research and the selected students. The overall length of the presentations is 5 hrs and 51 minutes.

4.2 Students' Assignment

In this subchapter, more details about students' presentations will be provided. The previous subchapter gave the basic information related to the total amount of samples and the overall length of presentations. The main goal for each student was to make a persuasive presentation that will be 5–7 minutes long. Some recordings were more than 7 minutes long; however, they were accepted by the teacher because some students

wanted to provide more detailed information about the particular product they were presenting. This resulted in exceeding the length of some samples up to 10 minutes.

The assignments were mostly IT-oriented. There were two categories to choose from: IT and electrical engineering. Students were offered interesting topics to choose from. The deadline for giving presentations was set in advance, so the students had some time to prepare their presentations. Some students had an opportunity to present a unique product which could be fictional. Unfortunately, the pandemic of COVID-19 resulted in online learning and some students had to record their videos at home using their own equipment. Regarding the guidelines of the assignment, Figure 9 illustrates the instructions given by the teacher.



Task: Make a powerful sales presentation on a selected product/service (hardware or software) using persuasive techniques.

In your presentation, you should include:

- specific points unique to the target group of customers (to help them solve their specific problem);
- a detailed description of the product/service;
- advantages of buying your product/service
- and/or physical demonstration (e.g. using a video).

The length of presentation should be 5 – 7 minutes.

Figure 9. Task assigned to students (Ellederová, 2020, p. 127).

Summary of Chapter 4

Chapter 4 provided essential information about the research samples to be analysed. First, the chapter described the size and content of the samples and outlined the process of analysis. Then the main goal of the assignment, to deliver a powerful persuasive presentation on the selected products, was set including the instructions for making the presentations. In the following chapter, an NLU analysis of the samples will be elaborated in greater detail.

5 Analysis of IT Students' Spoken Language with NLU

This chapter should reveal the results of the analysis of the student's spoken language. First, the chapter on methodology will describe the whole process of transcription of the students' presentation. Then, the next stage, the NLU analysis will follow. The deployment of the NLU application is currently available online on the public server. The application itself is open source, which means that there is no limitation in the license use, thus it can be used for the further research. The downside of the deployed application in Kubernetes[9]. The container offers the service for only one month. The following subchapters will discuss the research results and offer the insights about the analysis of the persuasive presentations. The tabular data taken directly from the NLU application are presented too. Furthermore, statistical data converted into graphs will present a clear picture of the overall project. For these findings, IBM Cognos Analytics 11.1.x[10] which is empowered by AI, was used to simplify the analysis process.

5.1 Methodology

After acquiring the video samples, one of the challenges was to design the flow of the project. There were two methods available, such as a speech-to-text service on IBM Cloud and dictation in the Microsoft Word application. The original idea was to simulate the process where the first thing that comes to the human's mind after hearing or reading something is what the information is about. This cognitive process is also present in AI, such as NLU which is empowered with a deep learning algorithm. However, a speech-to-text service turned out ineffective due to the problem with a variety of English accents spoken by non-native speakers of English. The second option such as Microsoft Word dictation was ideal for transcribing. In this case, I used speakers from one device and a microphone in another device. The sound was loud, so the microphone could record the words. Auto-punctuation is a feature in dictation whose role is to identify full stops or commas in spoken sentences in a video recording. The first 23 videos did not have this feature, while the rest of the videos did. A great challenge in the process of transcription was pronunciation of certain words, so the data received from videos were not altered at all. It happens even with devices that

---

[9] Kubernetes is an open-source system for automating deployment, scaling, and management of containerized applications.

[10] IBM Cognos Analytics is a web-based integrated business intelligence suite by IBM.

include voice recognition software such as Android mobile phones with enabled Google services. Often a spoken sentence could be misheard from a device and then unexpected formulations of the sentences could be offered.

The same approach was adopted in the process of transcription. 53 videos were transcribed, but 2 videos were missing, and 5 videos were excluded from the transcription process because of the bad sound quality. Overall, 87.28% of all samples was used in an NLU analysis.

After the transcription process was completed, an analysis by means of NLU began. In the introduction to Chapter 5, I mentioned the deployment of the NLU application on the IBM Cloud premises. In Figure 10, the workflow demonstrates the interaction of the user and the cloud-based application.

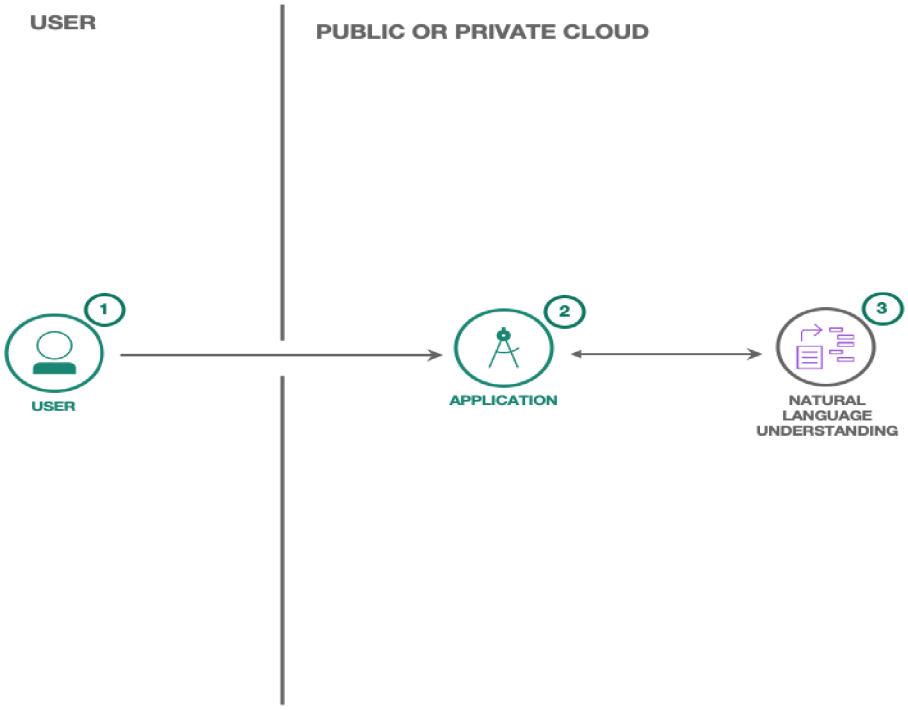


Figure 10. Workflow of NLU in IBM Cloud.

A user sends a message to the application located in IBM Cloud. The application forwards the input to the NLU engine and the output is retrieved. A source code and testing are available in IBM Cloud as a lite plan version which can be used for integration to any other framework. In Figure 11, the environment of NLU for testing is shown.

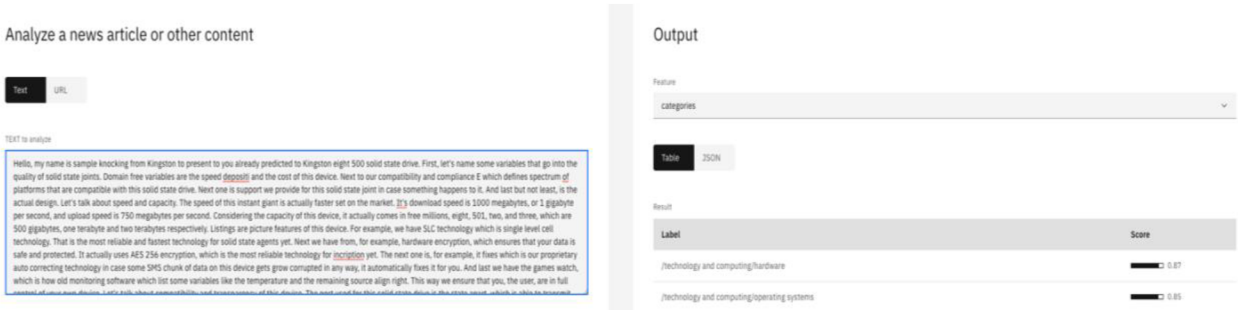


Figure 11. Persuasive presentation analysis using NLU.

As Figure 11 shows, the interface is simplified – on the left side there are two input possibilities. One is a text box and another one is a webpage with a URL input. Once the input text is inserted, it is automatically sorted and after clicking on "Analyze" on the right side, the output information is displayed.

In the demonstration of the sample analysis, there was the confidence score as one category (see Appendix 1). This text was about the solid-state drive, so NLU recognized that the text was about technology and computing/hardware. One of the advantages of NLU is that if the terminology in an input text is related to scientific discourse, such as an operating system, NLU can increase the precision of the findings and it can calculate the confidence score as 1. However, the downsides of NLU are that the emotion is too complex for AI to understand and the certainty levels in the analysis are usually low. This method was used for all transcribed samples. The output data are presented in Appendix 1 and interpreted in the next subchapter.

5.2 Research results

The research results (see Appendix 1) were recorded as tabular data and they were divided into 11 columns as a combination of the strings and integers. For each generated output, the result was recorded in an Excel table. After the recording to the Excel table, the data were presented in the bachelor's thesis as the research results data. The data were analysed in IBM Cognos Analytics to generate the results in the form of statistical data. In Appendix 1, you can find the results of the analysis.

The general properties of the research results are as follows:

- 87.28% of data were analysed;

- approximately 33,187 words were analysed with NLU with an average of 626.16 words per presentation;
- 34 videos were recorded by students and 19 videos were recorded by the teacher. This had a significant impact on the quality of the data.

In the next chapter, the collected data will be interpreted.

5.3 Data Insights

This part will deal with data management and analysis using IBM Cognos Analytics. The accuracy score tells us how much NLU is certain about the analysed text. In the following phase, the primary target will be score analysis such as “Category Score”, “Emotion”, “Emotion Score” “Sentiment Score”, “Category Count” and “Concept” (for more details see Appendix 1).

5.3.1 Category Score Results

The category score tells us about the level of certainty in the analysed text. For example, the output for this category of the analysed text will include “technology and computing/operating system,” which tells us that the text is about operating systems. In Figure 12, the sentiment score is grouped into 5 groups. However, these scores are considered as the good ones if they range from 0.50 and above.

The bar chart in Figure 12 reveals that the average score of accuracy for all samples is 0.95. 30 samples were rated 0.95 and above. 15 samples were rated with a score of 1 with the confidence score based on the concept of the text. The rest of the samples are satisfying the minimum requirement which is 0.50 and above.

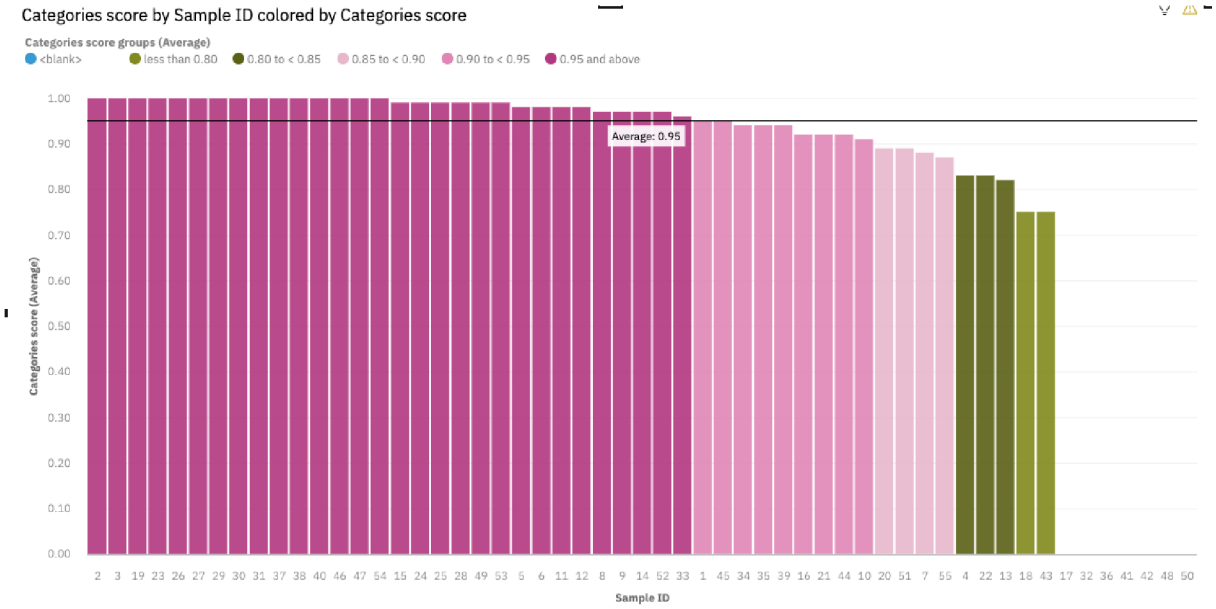


Figure 12. Bar chart of the category score.

5.3.2 Sentiment Score Results

The research results revealed the sentiment text and emotion. These two attributes should not be considered as the same. The emotion feature serves more about recognizing emotional presence within the text, such as “anger”, “joy”, “fear”, “sadness”, while the sentiment score refers to the classification of emotion in terms of positivity, negativity, and neutrality. In this case, 0 to 1 is considered as the level of certainty related to positive emotion. If the results are below 0, the emotion is considered as negative, while 0 means neutral emotion (see Figure 13).

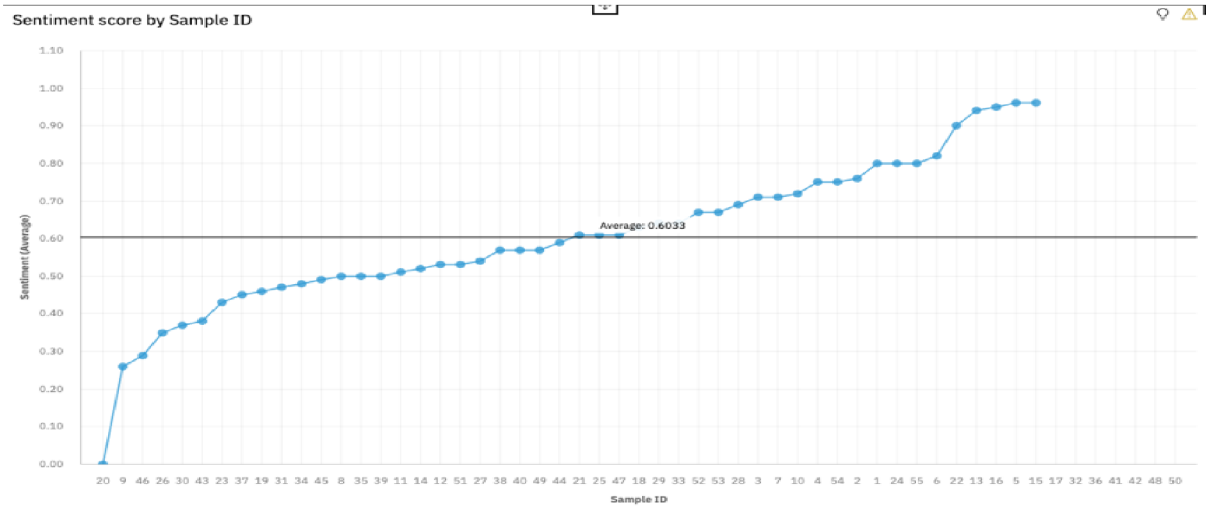


Figure 13. Sentiment score sorted in ascending order.

The average sentiment score for analysed samples was 0.6033, which tells us that all analysed samples had positive sentiment presented in the text. However, Sample 20 (see also Appendix 1) has a neutral score based on the analysis. This can happen if NLU fails to understand the complete context from the data source, and if the text was modified, it would increase a chance to recognize it. The samples such as 5 and 15 have the sentiment score of 0.95 which is high in comparison to the rest of the analysed samples (in Appendix 1).

5.3.3 Emotion Count and Emotion Score

The emotion count will indicate prevalent emotion within the sample. There are 4 emotions present in all samples as "anger", "fear", "sadness" and "joy". These emotions are sorted in the pie chart in Figure 14.

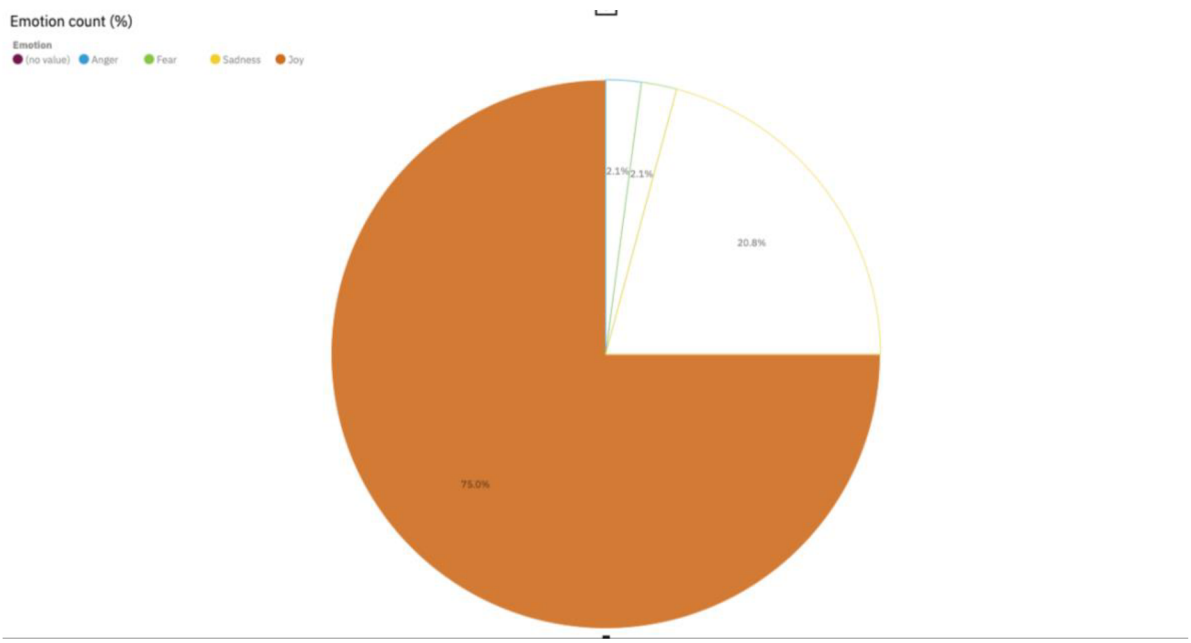


Figure 14. Emotion count (%) pie chart.

The emotion count tells us that prevalent emotion in the samples is "joy" with 75%. The second prevalent emotion is "sadness" with 20.8%, while the "anger" and "fear" have 2.1%. The following analysis will focus on "joy" and its certainty score.

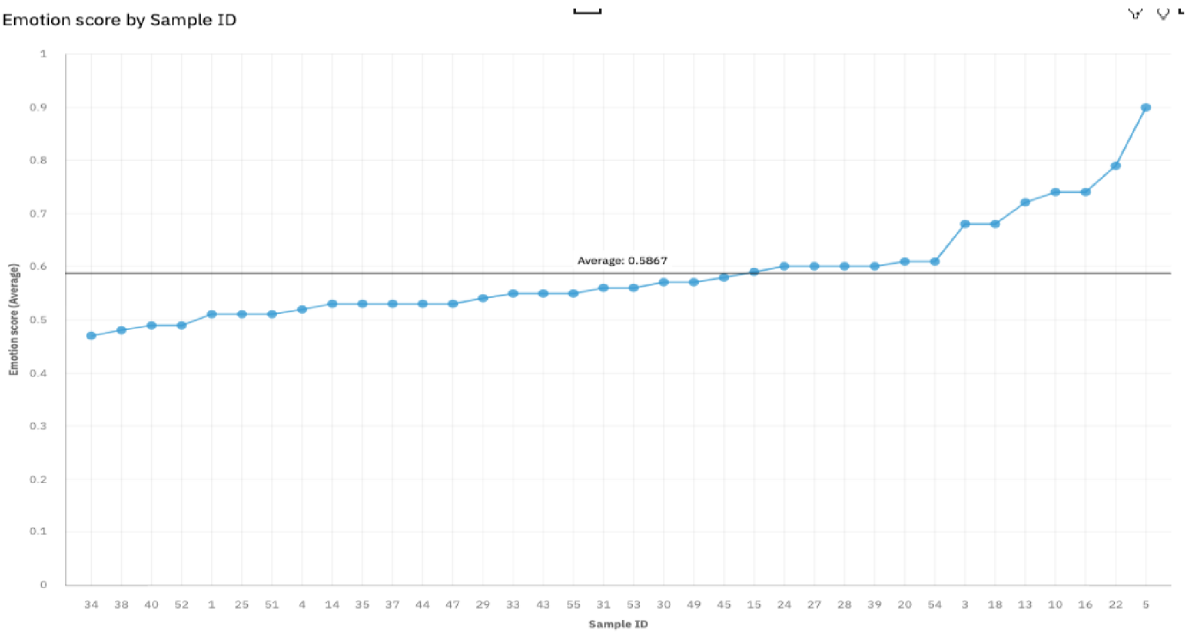


Figure 15. Emotion score sorted in ascending order.

The graph in Figure 15 indicates that the average score of all analysed samples classified as "joy" is 0.5867. Sample 5 reached the highest score of 0.9, while most samples reached the median line.

5.3.4 Category Count

In the category count, the numerical count of the analysed topics is presented. This time, there are the top counts as shown in Figure 16.

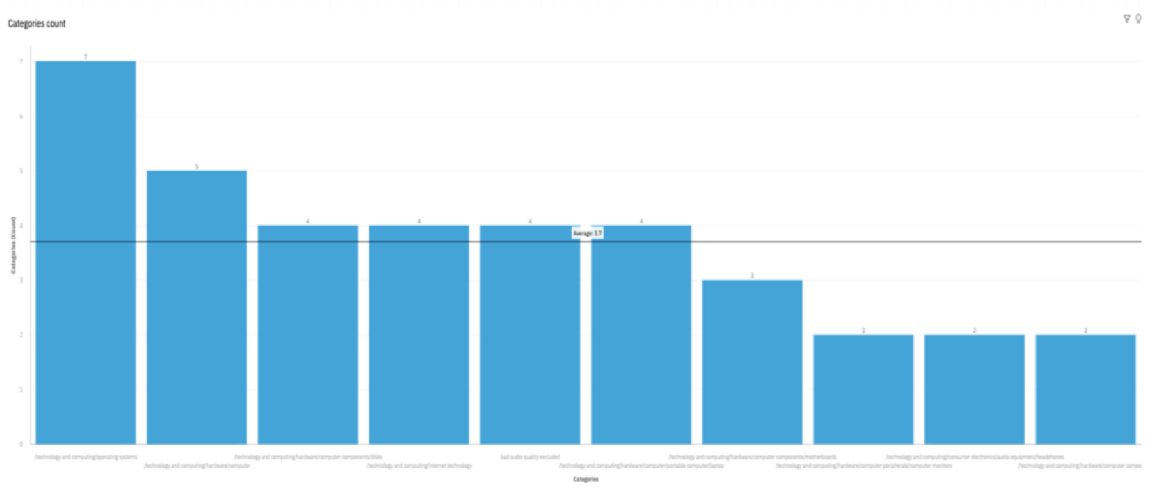


Figure 16. Top counts based on the category count.

The presence of the category of the operating system is highest in the analysed samples. However, two persuasive presentations are about the mainframes, so NLU classified them as the operating system category. All in all, the category is not 100% accurate in this case, but it gives an idea of what the text is about.

5.3.5 Concept Word Cloud and Concept Score Analysis

The concept analysis was done by NLU and its capability to recognize one or more concepts of the text. This can help us learn more about the text which I analysed. The word cloud method, as illustrated in Figure 17, displays all available concepts within a sample.

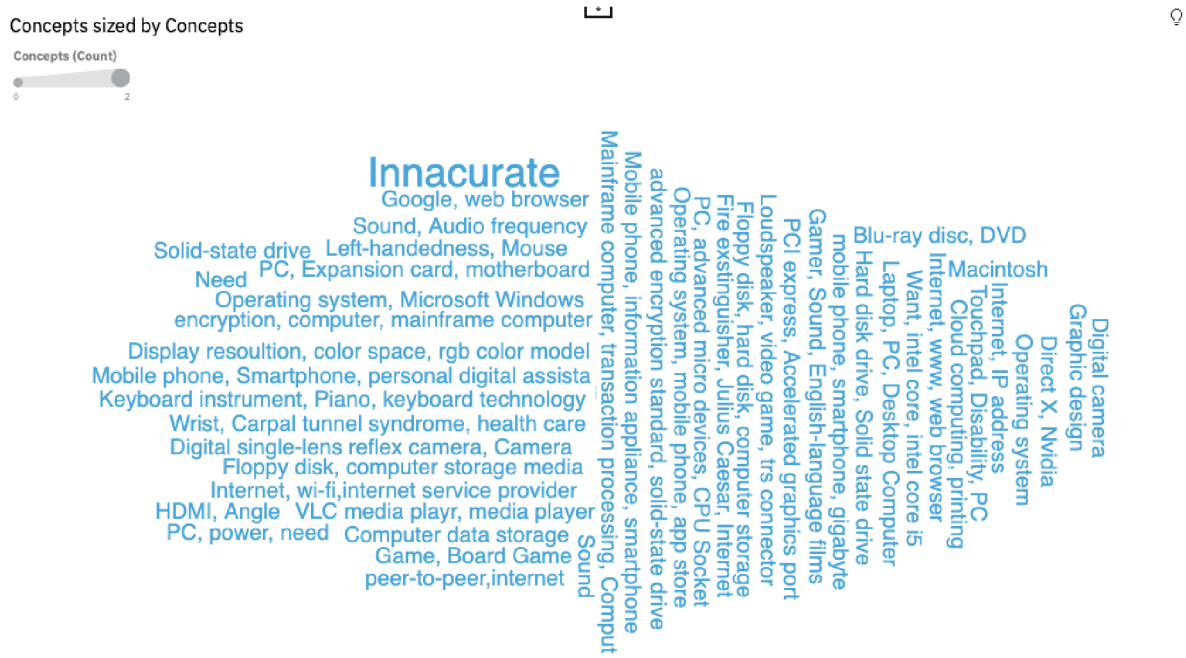


Figure 17. Word cloud of the concepts found in the persuasive presentations.

Scoring of these concepts can be found in the graph in Figure 18 demonstrating the accuracy score of certainty in NLU.

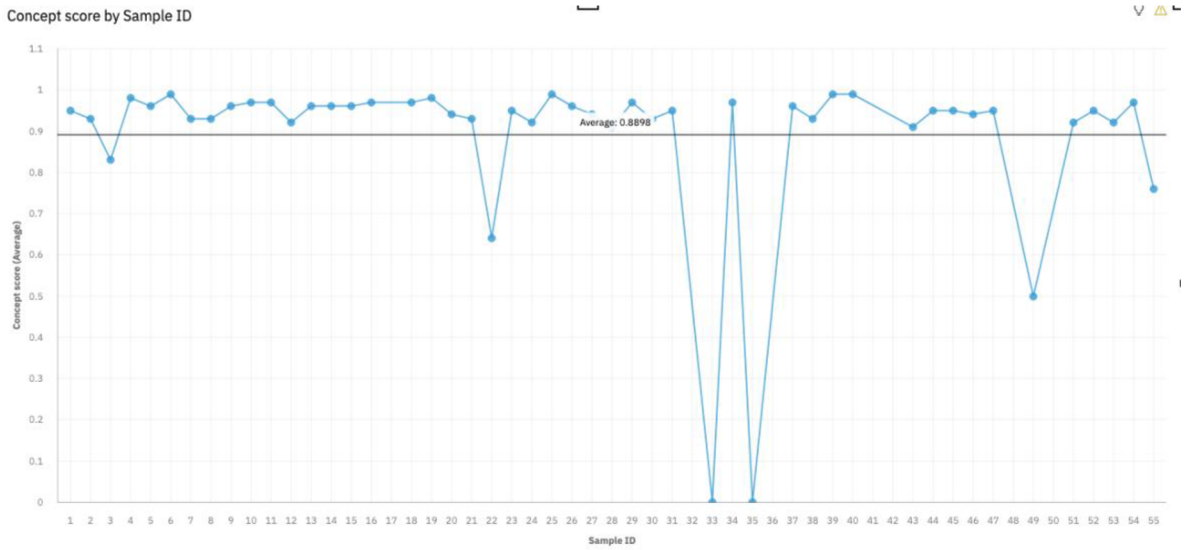


Figure 18. Concept score accuracy.

NLU demonstrated that it has a good capability of guessing the data analysed, however the concept of some samples, such as 33 and 35 failed to be recognized. The average score was 0.8898, which is a good indicator showing if the overall analysis was effective or not.

Summary of Chapter 5

In Chapter 5, the methodology for analysing the persuasive presentations was provided. Each step was described, and one demo of the NLU application was shown as a screenshot. Then the research results were presented in the form of tabular data and graphs, and consequently interpreted. However, some samples were excluded due to the bad quality or missing data. The research results served for the further analysis, such as calculating the accuracy score for the overall data. Also, the category count was presented to demonstrate the occurrences within the analysed data.

6 Discussion

This chapter aims to discuss the research findings based on the analysis of IT students' spoken language. In comparison to a human, AI does not appear as a totally reliable solution. In this case, the overall process was done in the simplest steps possible. Videos of persuasive presentations were transcribed, and the tool used for transcription was Microsoft Word. Microsoft Word has interesting features such as dictation or speech-to-text algorithm which converts spoken language into writing.

The challenging part was analysis of the accuracy of capturing the words during the presentations. For example, pronunciation played an essential part in this case. The students' English pronunciation is not perfect compared to native speakers, which is a natural thing. The speech-to-text algorithm was challenging to understand but, in the end, overall text was recorded for the following transcription. Also, some samples had good pronunciation, but somehow dictation failed to recognize the spoken sentence. A possible solution to this problem would be a controlled environment with a microphone situated near the speaker.

The next phases were the NLU analysis and research results presentation. From my point of view, the outcome of the analysis was better than expected. The structure of the text was questionable, but NLU was capable to analyse and classify its features. In the category score, there is a big group of samples whose accuracy score ranged between 0.85 and 1. However, in two samples such as 18 and 39, we can see that NLU classified them as an art event and health and fitness category. An explanation of this phenomenon is simple. Sample 39 used medical terms and the presentation mentioned a "carpal tunnel disorder". This presentation was about the prevention of this disorder by using a proper mouse. The reason for NLU failure was misunderstanding the concept, so these medical terms influenced its decision in analysis. A similar issue occurred with Sample 18 where NLU with the feature of the concept analyzer assigned the board games to the closest category of "arts and entertainment". The average score for determining the category was 0.95.

The other interesting findings emerged from the sentiment analysis. The sentiment score ranges between 0 and 1, and it tells us how much the meaning of the text is positive or not. The average score reached 0.60, however, there was a score of 0 for Sample 20. This might sound confusing when the emotion score and the sentiment score were different. Sample 20 delivered a piece of information about the product, however the way the text is structured

simply resulted in a wrong analysis. The error rate in samples was 1 in 48. This could be improved by providing a well-structured text with correct word order and punctuation.

Emotion and emotion score features proved to work well. A prevalent emotion label was chosen among other emotion features for the purposes of analysis. This should tell us how the samples were rated in the analysis. The average score was 0.58 for "joy", however, there were some cases in the samples where the prevalent emotion was "sadness" and "anger". For example, Sample 26 expressed the emotion "anger" but in the transcript, I could not find any negative meanings which would determine "anger" as a prevalent emotion. This type of error comes from the structuring of the text. Speech-to-text algorithms failed to capture some words and sentences properly, so this failure led to a domino effect, and it reflected negatively on the NLU output. The presence of "sadness" reached 20% but it does not mean that the texts expressed sadness. The feature does not work well with the text which does not have properly structured sentences. Regarding the concept and category classification, it turned out that NLU was successful in identifying them. There were two samples 33 and 35 in which NLU failed to understand the concept, but it recognized their category. The overall score for detection is quite high – it reached 0.88. This failure does not occur in a professional text, so the improvement would result from a better-formatted text, which applies to other cases as well.

Conclusion

Before I summarise the topic of my bachelor's thesis, I would like to recapitulate the stated objectives. The primary objective of the bachelor's thesis was to explain the foundations of AI and NLP. It should present a clear picture of the modern technological achievements and tools which are used to replace the old methods of checking the data insights. The main objective of the thesis was to analyse information technology students' spoken language. The methods of speech to text conversion and text analytic tools were applied to prove the efficiency of AI.

I faced many challenges regarding the sources I was selecting for the literature review. The spectrum of examples and theory was incredibly wide, so I carefully selected the most relevant sources. The analytical part of the bachelor's thesis was based on the theoretical framework established in the theoretical part.

The analytical part described the analysis of fifty-three samples of oral presentations given by students. Some of the samples were excluded because of their unavailability or bad quality. I applied the NLP techniques and analysis to convert the spoken language to text. The initial phase included cleaning the textual data such as misheard words so that consistency was maintained, and human bias were minimised. The final stage was focused on the NLP analysis with the IBM Watson NLU engine designed for NLP. The output was measured, and consistency ratings are available in the chapter dealing with research results. This engine should be implemented in linguistic-oriented English courses to help students improve or create a new NLP algorithm.

The NLU engine performed the required analysis and, from my point of view, with additional modification by the IBM Watson Knowledge Studio service, NLU would be capable of understanding text much better with a minimum of coding. Students of the study programme English in Electrical Engineering and Informatics may start discovering the NLU principles and their operation, which will help improve their data mining skills and bring other benefits to their professional career. We should bear in mind that AI will never be perfect without humans.

Rozšířený abstrakt

Tato bakalářská práce se zabývá problematikou nových technologií umělé inteligence při zpracování přirozeného jazyka. Práce přistupuje k dané problematice rozdělením na dvě části: část teoretickou a část analytickou. Každá z těchto částí je rozpracována včetně uvedení alespoň jednoho příkladu z praxe. V první části je hlavním cílem vymezit teoretický rámec umělé inteligence a jejích postupů, zpracování přirozeného jazyka a produktu IBM Watson Natural Language Understanding, zatímco analytická část se zabývá popisem výzkumu, výsledky výzkumu a cennými poznatky o datech.

Teoretická část představuje problematiku umělé inteligence a jejích disciplín, které zahrnují strojové učení, zpracování přirozeného jazyka, neuronové sítě, hluboké učení, počítačové vidění a kognitivní výpočty. Každá disciplína má vyhrazenou podkapitolu, která představuje výhody jejího použití ve světě moderních technologií. Pochopení těchto výhod je také užitečné pro porozumění logice a fungování určitých navržených modelů umělé inteligence, jako jsou chatboty nebo počítačové vidění, které využívá kameru k detekci rozpoznávání obličeje. Strojové učení a hluboké učení jsou běžně využívané disciplíny pro řešení náročných problémů v podnikání. Strojové učení a hluboké učení se těsně vztahují ke zpracování přirozeného jazyka. Důvodem je použití algoritmů pro zdokonalení modelů zpracování přirozeného jazyka při zjišťování struktury predikce vět a slov. Disciplíny umělé inteligence jsou vzájemně propojeny zejména při vytváření řešení pro lepší obchodní praktiky.

Následující kapitola se věnuje statistice a pravděpodobnosti. Je rozdělena do čtyř podkapitol, kde každá z nich vysvětluje základní statistické koncepty a jejich použití. První podkapitola prezentuje základní datové typy a slouží jako úvod do hlavních tří kategorií datových typů, jako jsou číselné datové typy, kategorické datové typy a pořadové číslo. Každý datový typ má své vlastní vlastnosti a způsoby, jak je s nimi manipulováno. Podkapitola popisné statistiky se skládá z matematických prvků a jejich použití při určování platnosti statistických údajů. Například shromážděná data mohou mít některé chybějící vstupy, ke kterým může snadno dojít lidskou chybou. Znalost popisné statistiky může analytikům dat pomoci překonat tyto potíže nahrazením chybějících vstupů průměrnými čísly, pokud jde o číselná data. Poslední podkapitola se zabývá teorií pravděpodobnosti včetně Bayesovy věty, která se úzce vztahuje k postupům strojového učení a dalším disciplínám. Tato podkapitola se dále zabývá čtyřmi typy pravděpodobnosti podle Nabiho (7. ledna, 2019), a to empirickou pravděpodobností, teoretickou pravděpodobností, společnou pravděpodobností a

podmíněnou pravděpodobností. Pravděpodobnost je zásadní při určování, zda model funguje podle očekávání, nebo zda je nutné provést zkoušku. Zatímco Bayesova věta se používá jako aktivační funkce, nechvalně známý algoritmus strojového učení pro skóre kreditní karty.

Kapitola 2 v teoretické části se zabývá zpracováním přirozeného jazyka a jeho procesy a technikami. Je rozdělena do dvou částí, kde první část pojednává převážně o fázích analýzy ve zpracování přirozeného jazyka, zatímco druhá část nás informuje o aplikacích zpracování přirozeného jazyka a jeho výhodách. Fází analýz je celkem pět, a tyto fáze a jejich role jsou popsány v jednotlivých podkapitolách. Tokenizace, lexikální analýza, syntaktická analýza, sémantická analýza a pragmatická analýza jsou klíčové komponenty pro práci se slovy ve větách. Tento přístup poprvé zavedli v roce 2003 Dale, Moisl a Sommers, kteří ve své knize *Handbook of natural language processing* (*Příručka zpracování přirozeného jazyka*) hovoří o hlavních principech fungování každé sekvence ve fázích analýzy. Následující kapitola věnovaná aplikacím zpracování přirozeného jazyka je rozdělena do pěti podkapitol, a každá obsahuje informace o aplikaci zpracování přirozeného jazyka a jeho použití v moderním světě. Analýza sentimentu přichází s podrobnostmi o tom, jak se používá a kde ji můžeme najít. Touto technologií se většinou zabývají marketingové služby, aby analyzovaly zpětnou vazbu uživatelů. Chatboty jsou rozvíjející se aplikací v oblasti zpracování přirozeného jazyka ve velkých podnicích a slouží operátorům při provádění snadných úkolů, zatímco operátoři se mohou soustředit na ty složitější. Jako příklad chatbotu je uveden TOBi od společnosti Vodafone. Klasifikace textu a extrakce textu mají mnoho společného, nicméně každé z těchto aplikací je věnována jedna podkapitola. Tyto aplikace jsou považovány za aplikace pro dolování z textu a jejich hlavním účelem je objevit vzory v datech, urychlit transformaci z nestrukturovaných textových dat na strukturovaná a přinést smysluplné informace o shromážděných nezpracovaných datech. V neposlední řadě je v aplikacích zpracování přirozeného jazyka uveden Machine Translator, včetně příkladů rozpoznávání řeči nebo jazykových překladačů. Rozpoznávání řeči se rychle vyvíjí a lze jej nyní najít v každém chytrém zařízení jako Google Assistant nebo Siri. Výroba strojových překladačů je založena na dvou metodách, povrchní a statistické metodě, které jsou v teoretické části práce popsány.

Poslední kapitolou teoretické části je kapitola o porozumění přirozenému jazyku od IBM Watson, která přináší více informací o produktu vytvořeném principy zpracování přirozeného jazyka a hloubkovým učením. Je rozdělena do dvou podkapitol. První podkapitola řeší principy fungování zpracování přirozeného jazyka. Začíná popisem toho,

jak zpracování přirozeného jazyka funguje. Produkt je zjednodušen na věc, kde existuje pouze jeden vstup ve formě textu a devět různých výstupů, které závisí na zvolené funkci. Nástroj porozumění přirozenému jazyku umožňuje extrakci metadat textu, který je v dnešní době klíčem v každém podnikání. Následující podkapitola popisuje podporovanou funkci a její integraci do jiných rámců. Porozumění přirozenému jazyku je velmi flexibilní s integrací API a programovacími jazyky. Podporuje nejpoužívanější programovací jazyky jako Python a Java. Tento produkt lze také použít společně se studiem Watson Knowledge Studio, kde lze extrakci metadat upravit podle potřeb uživatele. Poslední podkapitola teoretické části pojednává o plánu projektu a zabývá se procesem analýzy. Představuje příklad strojového překladače nebo algoritmu převodu řeči na text v dokumentu Microsoft Word, který se používá k přepisu videí. Celkový proces a aplikace plánu jsou podrobněji rozpracovány v analytické části bakalářské práce.

Analytická část práce se skládá z kapitol, které popisují výzkumný vzorek, výzkumnou metodologii, prezentují a interpretují výsledky výzkumu. Studenti z Fakulty informačních technologií a Fakulty elektrotechniky a komunikačních technologií VUT v Brně dostali za úkol připravit persuasivní prezentace produktů z oblasti IT. Videozáznamy byly shromážděny Mgr. Ing. Evou Ellederovou, Ph.D., učitelkou předmětu Angličtina pro IT. Tyto videozáznamy byly přepsány strojovým překladačem. Další kapitola se zabývá samotnou analýzou mluveného jazyka studentů IT prostřednictvím porozumění přirozenému jazyku. Popisuje postupný proces realizace analýzy a prezentuje její výsledky, které jsou uvedeny v Příloze 1. Po shromáždění výsledků výzkumu byla analyzována přesnost predikce analýzy porozumění přirozenému jazyku. Kapitola s názvem Data Insights se skládá z pěti podkapitol, které interpretují výsledky výzkumu. Ke generování přehledů o datech byl použit produkt IBM Cognos Analytics. Poslední kapitola bakalářské práce je věnována diskusi o výzkumných zjištěních prostřednictvím analýzy pomocí nástroje umělé inteligence. Celý proces zpracování bakalářské práce mě zaujal a umožnil mi hlubší vhled do řešené problematiky včetně seznámení s různými autory, kteří se zabývali tématikou zpracování přirozeného jazyka a umělé inteligence. To mi umožnilo získat nové zkušenosti, kterými se zabývám v poslední kapitole v rámci diskuze.

List of Figures

List of References

Agarwal, S. (November 22, 2020). Is this the end for convolutional neural networks? Retrieved from https://towardsdatascience.com/tagged/deep-learning

Dale, R., Moisl, H., & Somers, H. (2000). *Handbook of natural language processing.* New York: Marcel Dekker, Inc.

Ellederová, E. (2020). *English for information technology.* Brno: Vysoké učení technické, Fakulta elektrotechniky a komunikačních technologií.

Gupta, S. (January 7, 2018). Sentiment analysis: concept, analysis and applications. Retrieved from https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17

Hornick, M. F., Marcade E., & Venkayala S. (2007). *Java data mining: Strategy, standard and practice.* San Francisco: Elsevier Publishing.

IBM. (2019). Vodafone GmbH. Retrieved from https://www.ibm.com/case-studies/vodafone

IBM Cloud. (2017, February 27). Natural language understanding API documentation. Retrieved from https://cloud.ibm.com/docs/natural-language-understanding?topic=natural-language-understanding-getting-started

Kane, F. (2020). Machine learning, data science and deep learning with Python. Udemy course. Retrieved from https://ibm-learning.udemy.com/course/data-science-and-machine-learning-with-python-hands-on/learn/lecture/4020062#overview

Keckl, M. (May 29, 2019). Confidence score/confidence level. Retrieved from https://botfriends.de/en/botwiki/confidence-score/

Kobayashi, V. Mol, S. & Berkers H. (August 2017). Text mining in organizational research. Retrieved from https://www.researchgate.net/publication/318114889_Text_Mining_in_Organizational_Research

Lane, H., Howard, C. & Hapke, H. M. (2019). *Natural language processing in action. Understanding, analyzing, and generating text with Python.* Shelter Island: Manning Publications Co.

Laxmi, L., Kollati, K. & Amaranatha P. (January 2018). Text mining with Hadoop. Retrieved from https://www.researchgate.net/publication/332028341_Text_Mining_with_Hadoop_Document_Clustering_with_TF_IDF_and_Measuring_Distance_Using_Euclidean

Makadia, M. (May 10, 2019). What is Cognitive Computing? How are enterprises benefitting from Cognitive Technology? Retrieved from https://towardsdatascience.com/what-is-cognitive-computing-how-are-enterprises-benefitting-from-cognitive-technology-6441d0c9067b

Machine Translation (2012). In *Wikipedia.* Retrieved from https://en.wikipedia.org/wiki/Machine_translation

Mann, P. S. (1995). *Introductory statistics - student study guide.* Hoboken: Wiley.

McCulloch, W., & Pitts, W. (1943). *A logical calculus of the ideas immanent in nervous activity.* Illinois: Pergamon Press.

Mihajlovic, I. (April 25, 2019). Everything you ever wanted to know about computer vision. Retrieved from https://towardsdatascience.com/everything-you-ever-wanted-to-know-about-computer-vision-heres-a-look-why-it-s-so-awesome-e8a58dfb641e

Nabi, J. (January 7, 2019). Machine learning – probability & statistics. Retrieved from https://towardsdatascience.com/machine-learning-probability-statistics-f830f8c09326

Press, G. (December 30, 2016). A very short history of artificial intelligence (AI). Retrieved from https://www.forbes.com/sites/gilpress/2016/12/30/a-very-short-history-of-artificial-intelligence-ai/?sh=c53314d6fba2

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks, 61*, 85–117.

Sharma, A. (July 8, 2020). Top 10 applications of natural language processing. Retrieved from https://www.analyticsvidhya.com/blog/2020/07/top-10-applications-of-natural-language-processing-nlp/

Software Testing Help. (November 13, 2020). What is artificial intelligence? Retrieved from https://www.softwaretestinghelp.com/what-is-artificial-intelligence/

Tutorials Point. (2019). Natural language processing. Retrieved from https://www.tutorialspoint.com/natural_language_processing/natural_language_processing_syntactic_analysis.htm

Vergara, S., El-Khouly, M., El-Tantawi, M., Marla, S., & Sri, L. (June, 2017) Building cognitive applications with IBM Watson services: Volume 7 natural language understanding

Vibhor, N. (May 26, 2018) Artificial intelligence! What is it actually? Retrieved from https://towardsdatascience.com/artificial-intelligence-what-is-it-actually-7733032083b1

Zia, M. (March 7, 2018). Watson natural language classifier best practices. Retrieved from https://medium.com/ibm-watson/watson-natural-language-classifier-fb66206be6de

Zsolt, N. (2019). *Osnove vestacke inteligencije I masinskog ucenja.* Birmingham: Packt Publishing.

Appendices

Appendix 1: Research Results Table

Sample ID	Categories	Category score	Concept	Concept score	Emotion	Emotion score	Entities	Entity score	Keywords	Sentiment score
1	technology and computing/hardware/ computer	0.95	Computer data storage	0.95	Joy	0.51	Technical University of Brno as an organization	0.95	optical storage devices	0.80
2	technology and computing/ computer security/ antivirus and malware	1.00	Need	0.93	Sadness	0.61	ESET as an organization	0.96	basic user needs	0.76
3	technology and computing/ operating systems	1.00	Operating system	0.83	Joy	0.68	Microsoft as a company	0.96	newest version, GUI	0.71
4	technology and computing/ consumer electronics/ audio equipment/ headphones	0.83	Sound	0.98	Joy	0.52	Samsung as a company	1.00	Samsung galaxy buds, dynamic speaker	0.75
5	/technology and computing/hardware/ computer components/ graphics cards	0.98	Direct X, Nvidia	0.96	Joy	0.90	NVIDIA as a company	0.96	much power, next generation games Microsoft	0.96
6	technology and computing/hardware/ computer components/ disks	0.98	Solid-state drive	0.99	Sadness	0.42	Kingstone as a company, 2000 MB as quantity	0.95	SSD disk	0.82
7	technology and computing/software	0.88	VLC media player, media player	0.93	Sadness	0.6	Apple as a company	0.98	nice individual video, media player	0.71
8	technology and computing/hardware/ computer components/ disks	0.97	Hard disk drive, solid state drive	0.93	Sadness	0.59	240 gigabytes as quantity	0.95	new fast SSD series	0.50

9	technology and computing/operating systems/ Mac OS	0.97	Macintosh	0.96	Sadness	0.43	Apple & Acer as companies	1.00	Mac OS, MacBook Pro	0.26
10	technology and computing/Internet technology/web search/people search	0.91	Google, Web browser	0.97	Joy	0.74	Google as a company	1.00	Google chrome, web browsers	0.72
11	technology and computing/hardware/computer components/ disks	0.98	Floppy disk, computer storage media	0.97	Sadness	0.5	Peter Junak as a person	0.95	hard drive, magnetic storages	0.51
12	technology and computing/Internet technology	0.98	Internet, WWW, Web browser	0.92	Sadness	0.69	YouTube as a company	0.98	new feature, social media	0.53
13	technology and computing/hardware/ computer components	0.82	Sound, audio frequency	0.96	Joy	0.72	100% quantity, 30,000 Hz quantity	0.95	get user input dual speakers	0.94
14	/technology and computing/Internet technology	0.97	peer-to-peer, Internet	0.96	Joy	0.53	$0.99	0.53	ambitious project, access point	0.52
15	technology and computing/ operating systems	0.99	encryption, computer, mainframe computer	0.96	Joy	0.59	IBM as a company	0.99	business world, multi-frame mainframe, business transaction	0.96
16	technology and computing/consumer electronics/ camera and photo equipment/ cameras and camcorders/ camera batteries	0.92	mobile phone, smartphone, gigabyte	0.97	Joy	0.74	74.5 centimetres as quantity	0.95	Samsung galaxy S20 5G, single charge H18	0.95
17	bad audio quality → excluded	–	–	–	–	–	–	–	–	–
18	art and entertainment/ shows and events	0.75	Game, board game	0.97	Joy	0.68	YouTube as a company	0.96	war games today, ball games	0.63

19	technology and computing/hardware/ computer/ portable computer/ laptop	1.00	Laptop, PC, desktop Computer	0.98	Sadness	0.85	one terabyte as quantity	0.95	good price, bigger disk space	0.46
20	technology and computing/operating systems	0.89	Mobile phone, information appliance, smartphone	0.94	Joy	0.61	N/A	0.00	use smartphones, people today, new product	0.00
21	technology and computing/hardware/ computer peripherals/ printers, copiers and fax/ printers	0.92	Cloud computing, printing	0.93	Sadness	0.78	Canon as a company, Drop Box as a company	0.95	piece of paper, best possible way	0.61
22	technology and computing/hardware/co mputer components	0.83	Left-handedness, mouse	0.64	Joy	0.79	200 inches, 7.7 as quantity	0.95	left-handed gaming mouse, best performance	0.90
23	/technology and computing/hardware/co mputer networking/ router	1.00	Internet, wi-fi, Internet service provider	0.95	Sadness	0.67	one gigabit, five gigabits, sentinel	0.95	Internet users, big network, ISP	0.43
24	technology and computing/consumer electronics/audio equipment/headphones	0.99	Gamer, sound, English-language films	0.92	Joy	0.6	25,000 Hz as quantity	0.96	Next thing impedance, best headset	0.80
25	technology and computing/operating systems	0.99	Mainframe computer, transaction processing, computer	0.99	Joy	0.51	99% as quantity, IBM	0.96	Huge companies, IBM Mainframe Z	0.61
26	technology and computing/hardware/co mputer	1.00	Floppy disk, hard disk, computer storage	0.96	Anger	0.59	44 megabytes, 3.5 inch	0.95	Floppy disk, early computers, soft magnetic device	0.35
27	/technology and computing/hardware/ computer components/ motherboards	1.00	PC, expansion card, motherboard	0.94	Joy	0.6	32 Gigabytes, 128 Gigabytes as quantities	0.95	complicated structure, brilliant series of processor Ryzen	0.54

28	technology and computing/hardware /computer/desktop computer	0.99	PC, power, need	0.91	Joy	0.6	Microsoft as company	0.77	real games, best ways, desktop computer	0.69
29	technology and computing/operating systems	1.00	Operating system, mobile phone, app store	0.97	Joy	0.54	Nanosoft, 15 gigabytes	0.96	new operating system, new invention	0.64
30	technology and computing/Internet technology	1.00	Fire extinguisher, Julius Caesar, Internet	0.93	Joy	0.57	Tony Johnosn as person, polish church	0.95	good connection, Tony, rural town	0.37
31	technology and computing/hardware/co mputer	1.00	Want, Intel core, Intel core i5	0.95	Joy	0.56	Bryce, green bay (geographical feature)	0.8	Different version of this computer	0.47
32	bad audio quality → excluded	–	–	–	–	–	–	–	–	–
33	technology and computing/hardware/ computer/portable computer/laptop	0.96	Inaccurate	0.00	Joy	0.55	Intel, AMD as company	0.51	best things, own ideal laptop	0.64
34	technology and computing/operating systems	0.94	Mobile phone, smartphone, personal digital assistant	0.97	Joy	0.47	Steve Jobs, IBM, Samsung	0.95	category of mobile phone, regular cell phones	0.48
35	technology and computing/hardware/ computer/portable computer/laptop	0.94	Inaccurate	0	Joy	0.53	6 gigabytes, 16 gigabytes	0.95	good game, best screen, best graphics	0.50
36	bad audio quality → excluded	–	–	–	–	–	–	–	–	–
37	technology and computing/hardware/ computer components/ motherboards	1.00	PCI express, accelerated graphics port	0.96	Joy	0.53	RAMs, ISU	0.45	integrated number, small board	0.45
38	technology and computing/hardware/ computer components/ motherboards	1.00	PC, advanced micro devices, CPU Socket	0.93	Joy	0.48	AMD, Intel	0.96	advanced filtering system, filtering system	0.57

39	health and fitness/disorders	0.94	Wrist, carpal tunnel syndrome, health care	0.99	Joy	0.6	carpal tunnel disorder, Microsoft	0.95	today US, important thing	0.50
40	technology and computing/consumer electronics/camera and photo equipment/ cameras and camcorders	1.00	Digital single-lens reflex camera, camera	0.99	Joy	0.49	N/A	0.00	high end, audio technology, wide selection	0.57
41	Missing sample	–	–	–	–	–	–	–	–	–
42	Missing sample	–	–	–	–	–	–	–	–	–
43	automotive and vehicles/vehicle brands	0.75	Graphic design	0.91	Joy	0.55	Medicard	0.96	terrific last year	0.38
44	technology and computing/hardware/ computer/portable computer/laptop	0.92	Digital camera	0.95	Joy	0.53	Duncan	0.98	traffic stop, picture of the exam,	0.59
45	technology and computing/hardware/ computer	0.95	Keyboard instrument, piano, keyboard technology	0.95	Joy	0.58	Huntsman elite	0.95	best newest feature, little risk	0.49
46	technology and computing/hardware/ computer	1.00	Touchpad, disability, PC	0.94	Fear	0.58	Stephen Hawking	0.09	Stephen hawking, final step	0.29
47	technology and computing/Internet technology	1.00	Internet, IP address	0.95	Joy	0.53	Central Europe as location	0.98	Internet activity, download movies, IP address	0.61
48	bad audio quality → excluded									
49	/technology and computing/hardware/co mputer peripherals/computer monitors	0.99	Display resolution, colour space, RGB colour model	0.5	Joy	0.57	100 %, adobe	0.95	professional monitors, graphic designers, high end	0.57
50	excluded because of technical difficulties	–	–	–	–	–	–	–	–	–

51	technology and computing/hardware/ computer components/ sound cards	0.89	Loudspeaker, video game, TRS connector	0.92	Joy	0.51	Europe as location	0.2	limited time offer, sound output devices	0.53
52	technology and computing/hardware/ computer components/ disks	0.97	Blu-ray disc, DVD	0.95	Joy	0.49	Blu ray as company,	0.96	Blu-ray, optical storage devices, higher storage capacity	0.67
53	technology and computing/hardware/ computer peripherals/ computer monitors	0.99	HDMI, angle	0.92	Joy	0.56	55 centimetres, 27 inches as quantity	0.95	pixel response time, colour performance	0.67
54	technology and computing/ operating systems	1.00	Operating system, Microsoft Windows	0.97	Joy	0.61	One MB as quantity, 64 bits as quantity	0.95	Evolution of the previous Windows, Internet explorer	0.75
55	technology and computing/hardware	0.87	advanced encryption standard, solid-state drive	0.76	Joy	0.55	Kingstone as a company, 1000 megabytes	0.95	cost of this device, solid state drive	0.80

Appendix 2: Transcriptions of Students' Presentations (in the attached electronic file)