

> are you real?

- Yes, I am real. [13.75]

> are you a robot?

- No, I'm not a robot. [14.5]

> are you human?

- Yes. [12.1]

> how old are you?

- 19 [6.8]

> are you a girl?

- I am a girl.

> what is the meaning of life?

- I think it's a joke. [14.5]

> is life funny?

- No, it's not. [13.45]

> so it's a bad joke?

- I think it's a bad joke. [15.1]

## Talk to me, I won't byte - Programming a Conversational Chatbot

Matura Project by Natasha Savic, class 6g  
Kantonsschule Zürcher Unterland  
2019/2020

Supervisors: Ms Siri Kohl and Ms Astrid Romer Bär  
Expert: Ms Nicole Westermann

<b>1</b>	<b>Topic</b>	<b>1</b>
<b>2</b>	<b>Objective</b>	<b>1</b>
<b>3</b>	<b>Theory</b>	<b>1</b>
3.1	Subdivisions of artificial intelligence	1
3.2	Artificial intelligence	1
3.3	Machine learning	2
3.4	Representation learning	2
3.5	Neural networks	3
3.6	Retentiveness of neural networks	3
3.7	Deep learning	3
3.7.1	Supervised, unsupervised and semi-supervised learning	4
3.7.2	Reinforcement learning	4
3.8	Implementing data into neural networks	4
3.9	Requirements for deep learning	4
<b>4</b>	<b>Concept of this project</b>	<b>5</b>
<b>5</b>	<b>Material</b>	<b>5</b>
5.1	Dataset	5
5.1.1	Creating new data sources	5
5.1.2	Readily available data sources	6
5.1.3	Source	6
5.1.4	Format and Buffering	6
5.2	Software	6
5.2.1	Programming language	6
5.2.2	Installation and Programming environment	6
5.2.3	Relational Database Management System	7
5.2.4	Database Management Tools	7
5.2.5	Python libraries	7
5.2.6	Code from GitHub	7
5.3	Hardware	8
<b>6</b>	<b>Method</b>	<b>8</b>
6.1	Preparing the data using a database	8
6.2	Training the chatbot	9
<b>7</b>	<b>Results</b>	<b>9</b>
<b>8</b>	<b>Discussion</b>	<b>11</b>
8.1	Positives	11
8.1.1	Accomplishment of aim	11
8.2	Negatives	11
8.2.1	Dataset reduction	11
8.2.2	Profusion of data	11
8.2.3	Elusive and non-existent compatibility	11
8.2.4	Inadequate hardware	11
8.2.5	Overall performance	12

<b>9</b>	<b><i>Conclusion</i></b>	<b>12</b>
<b>10</b>	<b><i>Acknowledgements</i></b>	<b>12</b>
<b>11</b>	<b><i>References</i></b>	<b>13</b>
<b>11.1</b>	<b>Information sources</b>	<b>13</b>
<b>11.2</b>	<b>Image sources</b>	<b>14</b>
<b>11.3</b>	<b>Links to datasets and code</b>	<b>14</b>
<b>12</b>	<b><i>Additional exemplary transcriptions of conversations</i></b>	<b>15</b>
<b>13</b>	<b><i>Declaration of Originality</i></b>	<b>16</b>

## 1 Topic

“Programming a conversational chatbot” is a Matura project dedicated to obscuring the line between human and computer interaction. Artificial intelligence enables complex neural networks to be constructed and thus behaviour closely resembling humans to be replicated and imitated by their digital counterparts.

## 2 Objective

The purpose of this project was to create a functioning chatbot, a computer program simulating a casual interlocutor, capable of partaking in simple conversations. The exchanges were intended to be kept on an everyday, informal basis in an attempt to mock basic human interaction as accurately as possible. Following the empirical method of reinforcement learning, the chatbot was to be an independent conversational partner, meaning it was intended to autonomously generate adequate responses to given prompts without necessitating further cues or programming.

## 3 Theory

### 3.1 Subdivisions of artificial intelligence

The term “artificial intelligence” covers a wide range of fields revolving around digital anthropomorphism. It can be broken down into the following subcategories, each new step representing a subset of the former<sup>1</sup>:

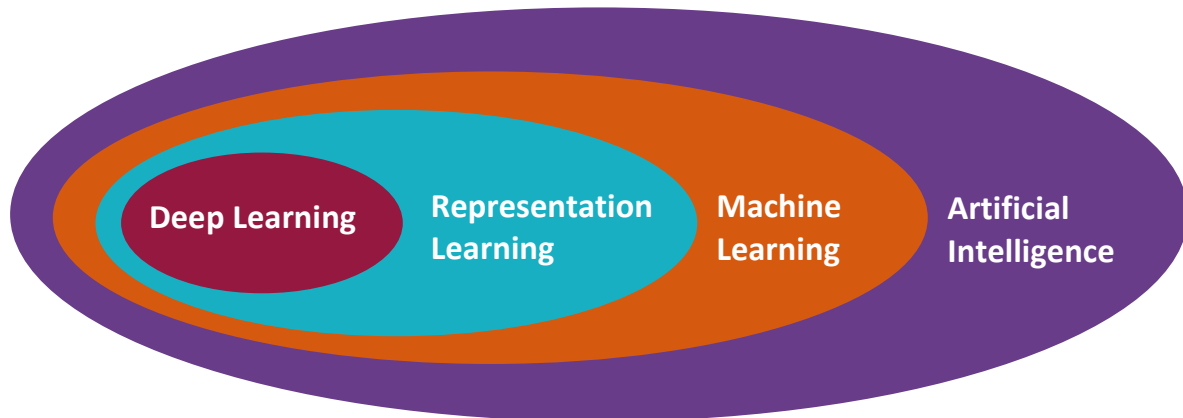


Figure 1: Visualization of the link between artificial intelligence, machine learning, representation learning and deep learning

### 3.2 Artificial intelligence

Artificial intelligence simply signifies the presence of a machine possessing skills typically attributed to human beings. This however does not stipulate independent learning done by the computer; there are artificial intelligence instances entirely pre-programmed with no traces of adaptive learning.<sup>2</sup>

<sup>1</sup> Chollet, François. 2018. *Deep learning with Python*. Manning Publications Co. Shelter Island, NY, p. 3-12

<sup>2</sup> Ibid.

### 3.3 Machine learning

The additional restriction introduced with machine learning is the renouncement of pre-set algorithms as a guide. Contrary to common programming practice, where humans provide a machine with data as well as the respective algorithms to be applied, machine learning's approach suggests that the computer is solely supplied with example input data and responses, which are used to derive an individual intrinsic set of commands. These are saved for future use on fresh and unfamiliar data.<sup>3</sup>

In an effort to do so, the targeted apparatus statistically analyzes the data presented in search of repeating patterns, from which unique rules are inferred.<sup>4</sup> For this reason, it is propitious that a high amount of data is obtained.

### 3.4 Representation learning

Representation learning is further confined in the sense of input data appearing in a new configuration. It is presented in a procedure optimized for the specific task to convey a univocal comprehension of which traits are specifically to be targeted by the computer. The machine will learn and adapt these representations iteratively and automatically. When renewing the model, the revision is selected from an accumulation of derived formulae, termed the *hypothesis space*<sup>5</sup>.

What is achieved by doing so is a purification of the data, as it is filtered for relevant information<sup>6</sup>. This data distillation exclusively retains pertinent details as *features*.<sup>7</sup> A higher feature-index necessitates more layers of representation, indicating high dimensionality. Therefore, the sum of representation levels is a direct measurement for the depth of the implemented machine learning model.<sup>8</sup>

It is worth noting that these layers have been denominated as *neural networks*.<sup>9</sup>

---

<sup>3</sup> Ibid.

<sup>4</sup> Ibid.

<sup>5</sup> Blockeel H. (2011). *Hypothesis Space*. In: Sammut C., Webb G.I. (eds) *Encyclopedia of Machine Learning*. (Springer. Boston, MA), Retrieved 26.10.2019 from:

[https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-30164-8\\_373](https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-30164-8_373)

<sup>6</sup> Chollet, François. 2018. *Deep learning with Python* (Manning Publications Co. Shelter Island, NY), p. 3-12

<sup>7</sup> Goodfellow, I., Bengio, Y. & Courville, A. 2016. *Deep Learning* (The MIT Press. Cambridge, Massachusetts & London, England), p. 3-9

<sup>8</sup> Balas, V. E., Roy, S. S., Sharma, D. & Samui, P. (eds.). 2019. *Handbook of Deep Learning Applications* (Springer Nature Switzerland AG. Cham, Switzerland), p.3

<sup>9</sup> Chollet, François. 2018. *Deep learning with Python* (Manning Publications Co. Shelter Island, NY), p. 3-12; Mohammed, M., Khan, M. B., & Bashier, E. B. M. (2017). *Machine Learning: Algorithms and Applications* (CRC Press, Florida), p. 89

### 3.5 Neural networks

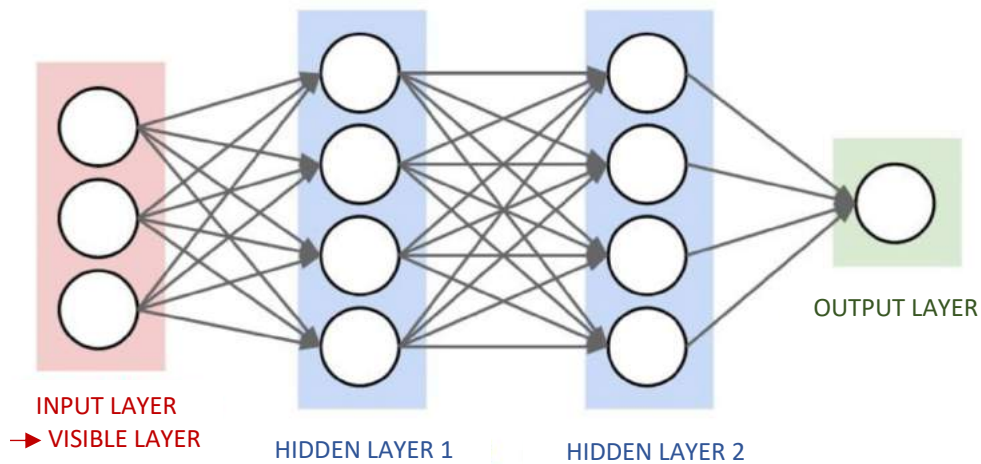


Figure 2: Neural Networks with the first input layer as a *visible layer*, followed by  $n$  amount of *hidden layers* and an output layer.

There are two types of demands which neural networks can be specialized to accommodate<sup>10</sup>:

- Classification: The purpose of the program is the distribution of given values to predefined categories.
- Regression: The machine is required to deduce a new value upon provocation.

### 3.6 Retentiveness of neural networks

Neural networks are not modelled to optimize data retention. In particular, feedforward networks, also known as multilayer perceptron or MLP, obliterate all data inserted. As a consequence, the entire dataset must be presented to the network simultaneously for an efficacious execution.<sup>11</sup>

Nonetheless, limited networks with forms of memory are emerging. Recurrent neural networks, or RNN, benefit from loops to engineer *long short-term memory*. However, this ability to recall information is merely ephemeral; as soon as high amounts of new data are submitted, the former input is forgotten in a process called *catastrophic interference*.<sup>12</sup>

### 3.7 Deep learning

A unique subcategory of machine learning is depicted in deep learning, which is a combination of supervised, unsupervised and reinforcement learning<sup>13</sup>.

<sup>10</sup> Skansi, Sandro. 2018. *Introduction to Deep Learning* (Springer International Publishing AG. Cham, Switzerland), p. 59-61

<sup>11</sup> Chollet, François. 2018. *Deep learning with Python* (Manning Publications Co. Shelter Island, NY), p. 196;

Goodfellow, I., Bengio, Y. & Courville, A. 2016. *Deep Learning* (The MIT Press. Cambridge, Massachusetts & London, England), p. 3-9

<sup>12</sup> Chollet, François. 2018. *Deep learning with Python* (Manning Publications Co. Shelter Island, NY), p. 196; Ramsundar, Bharath and Zadeh, Reza Bosagh. 2018. *TensorFlow for Deep Learning: from linear regression to reinforcement learning* (O'Reilly Media Inc. Sebastopol), p. 5, 152

<sup>13</sup> Skansi, Sandro. 2018. *Introduction to Deep Learning* (Springer International Publishing AG. Cham, Switzerland), p. 51

### 3.7.1 Supervised, unsupervised and semi-supervised learning

When training data is labelled by either a human or a machine based on criteria, the learning process is regarded as supervised. The approach of unsupervised learning describes the total absence of labels, placing the computer in search of an unclassified parameter. Both forms of training can be combined to semi-supervised learning.

### 3.7.2 Reinforcement learning

The acquisition of knowledge by means of trial and error is referred to as reinforcement learning. As no external guidance is provided, all insights amassed are done so empirically.<sup>14</sup>

## 3.8 Implementing data into neural networks

In order to transform the training data into a useable structure to contrive neural networks, the dataset must assume the format of a vector. Artificial intelligence specializing in language processing typically *tokenizes* the sample language excerpts. Here, tokenization entails separating each text file into individual characters or words, the so-called bag-of-words model. Subsequently, every element is numbered for the purpose of counting the occurrence of each unit. One disadvantage posed by the bag-of-words model is that it is saved as a set rather than a sequence, causing all sentence structure memory to be omitted. For this reason, n-grams, which are strings of successive words, were introduced and incorporated in the tokenization of the dataset.<sup>15</sup> Autoencoders are modules that encode and decode data whilst caching the largest amount of applicable information and discarding impractical contents, hence their frequent use in data processing.<sup>16</sup>

## 3.9 Requirements for deep learning

Typically, machine learning programs are operated in three stages<sup>17</sup>:

1. Representation
2. Evaluation
3. Optimization

To enable proper deep learning in such a manner, the following requirements must be met:

- The dataset used for training must be comprised of fitting input with corresponding replies as examples.
- An indicator measuring the success and failure rate to assist in consistently improving the algorithms is a prerequisite tool.<sup>18</sup>

---

<sup>14</sup> Goodfellow, I., Bengio, Y. & Courville, A. 2016. *Deep Learning* (The MIT Press. Cambridge, Massachusetts & London, England), p. 25

<sup>15</sup> Chollet, François. 2018. *Deep learning with Python* (Manning Publications Co. Shelter Island, NY), p. 3-9; Canny, John. 2014. *Introduction to Data Science: Natural Language Processing* [PDF file] Retrieved 03.11.2019 from:

<https://bcourses.berkeley.edu/courses/1267848/files/50935030/download?verifier=qPVn1u6pa0LKopYB6n7d aB9KX9stNjxCWnwM7oBh&wrap=1>

<sup>16</sup> Goodfellow, I., Bengio, Y. & Courville, A. 2016. *Deep Learning* (The MIT Press. Cambridge, Massachusetts & London, England), p. 3-9

<sup>17</sup> Domingos, Pedro. 2012. *Communications of the ACM Volume 55 Issue 10* (Machinery. NY) [PDF file] Retrieved 23.06.2019 from: <https://pdfs.semanticscholar.org/c3b6/0802b56eeec611e9def0dfbcaf42b851b99.pdf>

<sup>18</sup> Chollet, François. 2018. *Deep learning with Python* (Manning Publications Co. Shelter Island, NY), p. 3-12

A higher amount of data ameliorates effective training and raises the success rate. Thus, it is common practice to artificially increase the dataset through data augmentation<sup>19</sup>. This can entail rearranging, damaging or reconstructing the procured data to create pseudo-data.<sup>20</sup>

## 4 Concept of this project

As computer programs are unable to form opinions and therefore replies of their own, it's imperative to utilize a large amount of data elucidating regular human interaction. These datasets act as the catalyst leading to the emanation/emergence of the program's body of thought. The training data are organized into structured strings of conversation, making it possible to systematically run algorithms on these sequences. Only thereby is it possible for authenticating patterns to be distinguished in neural networks, ultimately leading to the derivation of independent responses.

## 5 Material

### 5.1 Dataset

The selection of the data type is a vital step in establishing an opportune work environment. Since a high prevalence of data promotes the generalization of the content<sup>21</sup>, a dataset large enough to appease the training model is of the utmost importance.

When searching for material, the two main options are to *create new data sources* or to rely on *readily available data sources*, the former of which can be further divided into two classes:

#### 5.1.1 Creating new data sources

- *Private sources*

Data obtained through private sources usually offer the advantage of high levels of uniformity. However, the additional privacy results in an expensive price as well as rigid restrictions regarding the download policy to personal devices and copyright. This delays the calculations performed on the data, as they cannot be saved on the server and require repeated downloading every time they are used.

- *Open sources*

In comparison to private sources, both the cheaper price and more lenient regulations about downloading and ownership of open source data pose a more advantageous alternative to the privately sourced equivalent. As open source data are readily available to anyone and thus easily manipulated, the issue of converting the data into a workable format still remains.

---

<sup>19</sup> Zhou, S. K., Greenspan, H. & Shen, D. (eds.). 2017. *Deep Learning for Medical Image Analysis* (Elsevier Academic Press. London, United Kingdom), p. 31

<sup>20</sup> Breiman, Leo. 1998. *Using convex pseudo-data to increase prediction accuracy* [PDF file], p.2, Retrieved 13.10.2019 from: <https://www.stat.berkeley.edu/~breiman/513.pdf>

<sup>21</sup> Balas, V. E., Roy, S. S., Sharma, D. & Samui, P. (eds.). 2019. *Handbook of Deep Learning Applications* (Springer Nature Switzerland AG. Cham, Switzerland), p.60



### 5.1.2 Readily available data sources

Since neither of these are an ideal tool for the project, readily available data sources were used instead of creating new ones. This datatype is typically labelled and consequently simplifies the process of converting the information into practicable material as well as a suitable format.

### 5.1.3 Source

For this chatbot, a dataset consisting of *Reddit* comments was employed<sup>22</sup>. *Reddit* is an interactive social media platform enabling users to exchange thoughts on a broad variety of topics. The practicality of this online portal is the thorough structure due to the subcategories, namely the “subreddits”. These are further organized into “parent-comments” with the respective replies forming different lines of conversation, constituting the epitome of human conversational interaction.

### 5.1.4 Format and Buffering

The data are in the JSON-format and contain information in abundance, exceeding merely the contents of the comments; they contain details on the username of the comment’s author, the score based on upvotes and downvotes and the date amongst other figures. However, these superfluous contents are not of importance as they can be filtered out during the process of text pre-processing.

The format of the data allows for a perspicuous dissection into the needed categories.

```
{
  "author": "jpb",
  "author_flair_text": null,
  "edited": false,
  "id": "c2117",
  "parent_id": "t1_c2103",
  "gilded": 0,
  "distinguished": null,
  "retrieved_on": 1473821405,
  "body": "it's strangely mesmerizing to watch the patterns, then set fire to the oil",
  "controversiality": 0,
  "subreddit_id": "t5_6",
  "stickied": false,
  "link_id": "t3_21509",
  "subreddit": "reddit.com",
  "score": 2,
  "ups": 2,
  "created_utc": 1135622591,
  "author_flair_css_class": null
}
```

Figure 3: An extracted exemplary comment from the *Reddit* dataset in JSON-format

## 5.2 Software

### 5.2.1 Programming language

The programming language most suitable for this project is Python. The plethora of open source code libraries found in Python, including several natural language processing options, provide a sufficient foundation for the training algorithms of the chatbot. Despite the higher number of accessible learning resources surrounding the older Python 2.7, the slightly newer version, namely Python 3.8, was the programming language of choice.

### 5.2.2 Installation and Programming environment

Prior to installing Python, the package manager Homebrew was downloaded. The purpose of such software is to log the already existing programs of an engine and ensure compatibility where interference between software versions occurs.<sup>23</sup>

Initially, Python’s Integrated Development and Learning Environment, commonly referred to as IDLE, was commenced with but soon after substituted by PyCharm as the final programming environment. PyCharm is a more extensive software package with additional features and a sleeker user interface.

<sup>22</sup> Link to torrent for *Reddit* comments: <https://files.pushshift.io/reddit/comments/> (29.04.2019)

<sup>23</sup> Retrieved 06.07.2019 from: <https://brew.sh/>

### 5.2.3 Relational Database Management System

Relational Database Management Systems, or RDBMS, store and organize data in tables. For the sake of efficiently structuring relevant information from large datasets, it is crucial to use RDBMS, especially when dealing with a multitude of data.<sup>24</sup> This task was assigned to PostgreSQL, a free relational database management system run by an open source community. Its advantages are the high speed of operation as well as its robust integration with software development frameworks<sup>25</sup>.

### 5.2.4 Database Management Tools

Database Management Tools are the graphical user interface employed to maintain and access the data in the Relational Database Management Systems. They are able to send queries to the database servers and extract the requested information<sup>26</sup>. DBeaver was chosen due to being free of charge, open source and widely used in the field of database administration<sup>27</sup>.

### 5.2.5 Python libraries

Python libraries were implemented as a way of enabling the creation of this project despite lacking adequate levels of Machine Learning experience. These libraries are samples of code and functions that operate in very specialized areas of programming requiring deep understanding of those fields. On behalf of this project, the python libraries listed below were applied:

- TensorFlow 1.14.0  
A tool used to create neural networks by means of dataflow graphs. It was created by the Google team and presupposes data formatted in arrays both entering as well as exiting the library<sup>28</sup>. This configuration of the data was not programmed in TensorFlow, but rather prepared separately as a regular Python project.
- Psycopg2  
A client library for interfacing with the PostgreSQL database server.<sup>29</sup>

### 5.2.6 Code from GitHub

For the neural network of the chatbot, the TensorFlow training and inference code was referenced from the open source library on GitHub.<sup>30</sup>

---

<sup>24</sup> Qin, L., Yu, J.X. and Chang, L. 2009 *Keyword search in databases: the power of RDBMS*. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data (ACM)* [PDF file] Retrieved 30.11.2019 from:

[https://www.researchgate.net/profile/Lijun\\_Chang/publication/221215146\\_Keyword\\_search\\_in\\_databases\\_the\\_power\\_of\\_RDBMS/links/0fcfd5140f8b78b220000000/Keyword-search-in-databases-the-power-of-RDBMS.pdf](https://www.researchgate.net/profile/Lijun_Chang/publication/221215146_Keyword_search_in_databases_the_power_of_RDBMS/links/0fcfd5140f8b78b220000000/Keyword-search-in-databases-the-power-of-RDBMS.pdf)

<sup>25</sup> Retrieved 18.07.2019 from: <https://www.postgresql.org/about/>

<sup>26</sup> Retrieved 18.07.2019 from: <https://docs.plesk.com/en-US/obsidian/administrator-guide/72893/>

<sup>27</sup> Retrieved 29.09.2019 from: <https://dbeaver.io/about/>

<sup>28</sup> Retrieved 29.09.2019 from: <https://www.tensorflow.org/>;

Retrieved 29.09.2019 from :<https://towardsdatascience.com/tensorflow-goes-2-0-but-what-is-dataflow-16a2056e6cf8>

<sup>29</sup> Retrieved 19.10.2019 from: <https://pypi.org/project/psycopg2/>

<sup>30</sup> Open library, retrieved from: <https://github.com/daniel-kukiela/nmt-chatbot>

### 5.3 Hardware

The primary hardware needed was a personal computer, which was the coding tool and the origin of commands. Since the computer's built in central processing unit (CPU) was inexpedient for coping with the volume of data (190GBs), a separate computer with 4 CPUs was assigned the task of executing the operations. Despite the hardware change, the program still proved to perform at an insufficient speed, hence the switch to a cloud based virtual machine run by Google for the training of the bot.

A Google Cloud Compute Engine<sup>31</sup> with the following specifications was rented for the duration of the training and testing phase:

- 8 vCPUs
- 30 GBs of memory
- A 500 GBs disk

## 6 Method

The creation of the chatbot proceeded in the following stages:

### 6.1 Preparing the data using a database

- 1.1. An ideal environment for the development of the chatbot was established. This was achieved by selecting and downloading the suitable programs, and installing as well as configuring the necessary software and hardware.
- 1.2. Next, the chosen dataset was analyzed with regards to formal structure, format and content. Using the SQL server and client, a database was created. This was configured to contain the comment id, the parent id, the content of the comment, the name of the subreddit and the score of each comment.
- 1.3. Thereafter, using Python, the *Reddit* comments were parsed and inserted into the newly created database. Additional parsing was performed to link comments with their parent as well as filtering the comments. Since parent comments could possess multiple replying comments, the intention was to merely store the version with the highest score.<sup>32</sup>
- 1.4. The data were then divided into a "from" file representing the parent comment and a "to" file for the reply to meet the vectorial stipulations of TensorFlow.<sup>33</sup>

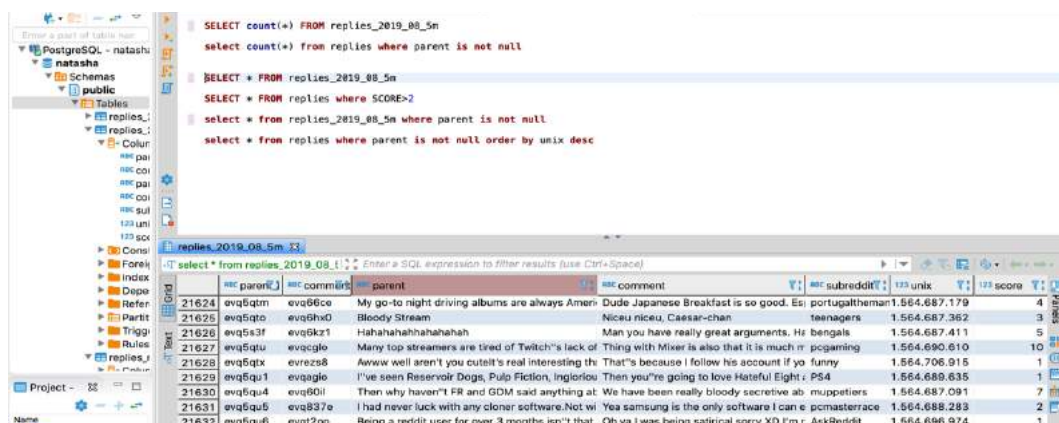


Figure 4: The Reddit comment database with the parent and comment id, the parent comment's and reply's content, the subreddit, the unix timestamp of the comment and the score

<sup>31</sup> Retrieved 23.11.2019 from: <https://cloud.google.com/compute/>

<sup>32</sup> Created code will be made accessible on: <https://github.com/natasha-savic/chatbot-data-preparation>

<sup>33</sup> Created code will be made accessible on: <https://github.com/natasha-savic/chatbot-data-preparation>

## 6.2 Training the chatbot

1.5. For the actual training of the chatbot, a source code from *GitHub* was taken as a template<sup>34</sup>. The open library was examined and slightly altered and adapted to accommodate this project and software configuration. The project contained two main modules:

- Training: The prepared data was inputted and repeatedly traversed through the TensorFlow model's neural network. The program was comprised of 13'000 steps per epoch, iterating for 3 epochs in total. The duration of this process was 72 hours.
- Inference: Once the training had ceased, the inference could be launched. This nearly instantaneously completed segment is initiated each time a new conversation session with the chatbot is commenced. It causes the machine to apply logic to the acquired data and triggers the interactive feature of the bot.

## 7 Results

A chatbot able to generate responses to statements made by a user of the program was successfully developed. Every entry evokes numerous answers, only displaying the most appropriate selection based on a compatibility score. Positive values indicate a high level of appositeness and are depicted in green, whereas negative figures print the responses in red as they are deemed inadequate. Each input is unerringly assigned a single outcome per session, causing identical questions or prompts to invariably instigate the same reply. In spite of that, refreshing the conversation by relaunching the inference code eliminates a fragment of the short-term memory, occasionally provoking a different reply than in previous conversations.

```
> where are you from          > where are you from
- I am from the US [8.9]      - South Africa [8.3]
> how old are you?           > how old are you?
- 18 [6.8]                   - 19 [6.8]
```

Figure 5: Different answers evoked by the same questions in separate sessions

```
> where are you from
- South Africa [8.3]
> Where are you from
- I'm from Canada [-86.25]
> where are you from?
- I'm from the UK [-86.25]
> Where are you from?
- I'm from Canada [-86.25]
```

Figure 6: Varied responses despite the same meaning of question asked

The program is case-sensitive, meaning that it responds dissimilarly to capitalized and lower-case words. This distinction can also be observed regarding punctuation. As an example, the inquiry "Where are you from?" was made four times during a single session, differing only slightly in terms of capital letters and interrogation points yet yielding non-deterministic results.

<sup>34</sup> Retrieved 23.11.2019 from: <https://github.com/daniel-kukiela/nmt-chatbot>

Furthermore, alternate spellings of the same words, albeit all correct, are interpreted diversely and instigate a wide range of reactions.

```
> favorite food
- I love it [7.85]

> favourite food
- I have a little chicken [-85.05]

> favorite song
- I love it [7.85]

> favourite song
- I'm not sure what [-85.95]

> favorite book
- I love it! [13.0]

> favourite book
- I was thinking the same thing [-84.15]
```

Figure 7: Differences in spelling are regarded as different statements

It is important to note that the chatbot does not comprehend regular commands as a computer would. The calculations performed on the input are done so in an closed context on a textual basis without access to computational capabilities.

```
> what is two plus two?
- 2018 [7.1]

> what is 2 + 2?
- 2.5. [12.1]

> 2 + 2
- 2,7 [-88.05]

> 2 + 2 =
- 2,7,7,7, [-87.3]
```

Figure 8: Failure to fulfill basic math operations

Although individual beliefs and judgements are predominantly well-articulated, the program frequently delivers incoherent, nonsensical statements. This usually occurs when knowledge and creativity are targeted. For example, when prompted to not only provide a reaction, but to additionally offer suggestions, recommendations or advice, the program behaves faultily.

```
> do you like memes
- I like it [7.85]

> rage quit
- r/unexpected [-86.7]

> reddit or instagram
- Instagram normies [9.05]

> Creeper?
- Aww man [7.55]

> are you the OP?
- Yes [6.95]

> upvoted my comment
- I upvoted ur comment [-185.5]

> thanks
- No problem [8.0]

> is today your cake day?
- No, I'm not a normie. [14.65]

> today is my cake day
- Happy cake day! [13.75]
```

Figure 9: Reddit's running jokes such as "Creeper? Aww man", "Instagram normies" and "cake day" were picked up on

An unsurprising specialty of the chatbot is its knowledge on trending witticism and Reddit terminology. Current catch-phrases were recognized and met with the standardized answer correctly. Reddit terminology regarding upvotes, original posters of content as well as even references to specific subreddits were observed.

## 8 Discussion

### 8.1 Positives

#### 8.1.1 Accomplishment of aim

In retrospect, the objective of the project was attained. A functional chatbot was created by means of neural networks, implemented using TensorFlow.

Given that the final constellation of the program was fully functional and producing a substantial amalgam of outputs, the software and code proved to be viable.

### 8.2 Negatives

#### 8.2.1 Dataset reduction

The dataset was specifically chosen on the basis of its considerable size, yet the used resources were not equipped with the means to process such high quantities. Initially set target capacities had been severely miscalculated and therefore needed to be retrenched. After reassessments, only a fragment of the procurable data was employed. If the chatbot optimization were to proceed past this project, the data could be gradually augmented, eliminating the issue of data curtailment.

#### 8.2.2 Profusion of data

Even with the reduction of data, the program dealt with a plenitude of information. Decompression revealed a more colossal size than what had been accounted for. This precluded proper regulation over and an assiduous analysis of the data. Additionally, the exorbitant dimensions of the input resulted in a severe retardment of all steps involved. Processing the SQL database as well as training the model required several days each. For this reason, experiments were an inordinately costly time investment, as it was a common occurrence for a program to crash after a runtime of multiple hours.

#### 8.2.3 Elusive and non-existent compatibility

The majority of time was undoubtedly expended on configurational and compatibility errors. Before any coding could even be attempted, the required software and libraries needed to be brought to a condition allowing for collaborative manipulation. This transcends just purely downloading the necessary material, and was further exacerbated through the multitude of hardware changes made, some of which were purely undertaken due to impossible interoperability.

#### 8.2.4 Inadequate hardware

While CPUs solve sequential calculations adeptly, GPUs' area of expertise lies in simultaneous processing of programs. As neural network training is based on calculations with arrays<sup>35</sup>, GPUs are the preferred choice for accelerated computing and having access to one could have minimized the lengthy runtimes involved.

---

<sup>35</sup> Balas, V. E., Roy, S. S., Sharma, D. & Samui, P. (eds.). 2019. *Handbook of Deep Learning Applications* (Springer Nature Switzerland AG. Cham, Switzerland), p. 11

### 8.2.5 Overall performance

Certain prompts evoke coherent responses from the chatbot, but incomprehensible and extraneous replies still betide far too frequent for it to be a useable conversational partner. This is presumedly attributed to the decreased training data and not errors in the code itself.

## 9 Conclusion

This project demonstrated the construction of a chatbot using Reddit comments. It is evident that the selected dataset exerts considerable influence on the way in which the software responds; as a future continuation, it could further be expanded by manipulating the generated replies based on the input of different data. Exploiting this feature could lead to bots possessing unique dispositions.

Moreover, pursuing this program to a greater extent would entail expanding the input information and upgrading the computing hardware. It would inevitably improve the success rate, garner an increasing number of logical outputs and accelerate the computing process. Once this is achieved, it is probable that conversations held by computers could become indistinguishable from those held by humans, fulfilling the ultimate purpose of artificial intelligence.

## 10 Acknowledgements

I would like to express my gratitude to my supervisors Ms A. Romer Bär and Ms S. Kohl for their trust and support during my unpredictable journey in an unfamiliar area.

Additionally, I would like to extend my thanks to my supportive family and in particular my father, who has always encouraged IT in our household and initiated my interest in programming.

## 11 References

### 11.1 Information sources

Balas, V. E., Roy, S. S., Sharma, D. & Samui, P. (eds.). 2019. *Handbook of Deep Learning Applications* (Springer Nature Switzerland AG. Cham, Switzerland), p.3, 11, 60

Breiman, Leo. 1998. *Using convex pseudo-data to increase prediction accuracy* [PDF file], p.2, Retrieved 13.10.2019 from: <https://www.stat.berkeley.edu/~breiman/513.pdf>

Blockeel, Hendrik. 2011. *Hypothesis Space*. In: Sammut C., Webb G.I. (eds.) *Encyclopedia of Machine Learning*. (Springer. Boston, MA), Retrieved 26.10.2019 from: [https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-30164-8\\_373](https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-30164-8_373)

Canny, John. 2014. *Introduction to Data Science: Natural Language Processing* [PDF file]. Retrieved 03.11.2019 from: <https://bcourses.berkeley.edu/courses/1267848/files/50935030/download?verifier=qPVn1u6pa0LKopYB6n7daB9KX9stNjxCWnwM7oBh&wrap=1>

Chollet, François. 2018. *Deep learning with Python*. (Manning Publications Co. Shelter Island, NY), p. 3-12, 152, 196

Domingos, Pedro. 2012. *Communications of the ACM Volume 55 Issue 10* (Machinery. NY) [PDF file]. Retrieved 23.06.2019 from: <https://pdfs.semanticscholar.org/c3b6/0802b56ecec611e9def0dfbcaf42b851b99.pdf>

Goodfellow, I., Bengio, Y. & Courville, A. 2016. *Deep Learning* (The MIT Press. Cambridge, Massachusetts & London, England), p. 3-9, 25

Mohammed, M., Khan, M. B., & Bashier, E. B. M. 2017. *Machine Learning: Algorithms and Applications* (CRC Press. Florida), p. 89

Skansi, Sandro. 2018. *Introduction to Deep Learning* (Springer International Publishing AG. Cham, Switzerland), p. 51, 59-61

Qin, L., Yu, J. X. & Chang, L. 2009 *Keyword search in databases: the power of RDBMS*. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data* (ACM) [PDF file]. Retrieved 30.11.2019 from: [https://www.researchgate.net/profile/Lijun\\_Chang/publication/221215146\\_Keyword\\_search\\_in\\_databases\\_the\\_power\\_of\\_RDBMS/links/0fcfd5140f8b78b220000000/Keyword-search-in-databases-the-power-of-RDBMS.pdf](https://www.researchgate.net/profile/Lijun_Chang/publication/221215146_Keyword_search_in_databases_the_power_of_RDBMS/links/0fcfd5140f8b78b220000000/Keyword-search-in-databases-the-power-of-RDBMS.pdf)

Zhou, S. K., Greenspan, H. & Shen, D. (eds.). 2017. *Deep Learning for Medical Image Analysis* (Elsevier Academic Press. London, United Kingdom), p. 31

*About: What is PostgreSQL?* Retrieved 18.07.2019 from: <https://www.postgresql.org/about/>



*Database Management Tools*. Retrieved 18.07.2019 from: <https://docs.plesk.com/en-US/obsidian/administrator-guide/72893/>

*DBeaver Community: About; Overview*. Retrieved 29.09.2019 from: <https://dbeaver.io/about/>

*Google Cloud Compute Engine*. Retrieved 23.11.2019 from: <https://cloud.google.com/compute/>

*Homebrew: The Missing Package Manager for macOS or Linux*. Retrieved 06.07.2019 from: <https://brew.sh/>

*Let's talk Dataflow: What is TensorFlow?* Retrieved 29.09.2019 from: <https://towardsdatascience.com/tensorflow-goes-2-0-but-what-is-dataflow-16a2056e6cf8>

*Psycopg2: Project description*. Retrieved 19.10.2019 from: <https://pypi.org/project/psycopg2/>

*TensorFlow: Why TensorFlow*. Retrieved 29.09.2019 from: <https://www.tensorflow.org/>

## 11.2 Image sources

Figure 2:

*Let's talk Dataflow: What is TensorFlow?* Retrieved 29.09.2019 from :<https://towardsdatascience.com/tensorflow-goes-2-0-but-what-is-dataflow-16a2056e6cf8>

## 11.3 Links to datasets and code

Torrents for Reddit comments, retrieved 29.04.2019 from: <https://files.pushshift.io/reddit/comments/>

Created code for data processing will be made accessible on: <https://github.com/natasha-savic/chatbot-data-preparation>

Open library with training module, retrieved 10.11.2019 from: <https://github.com/daniel-kukiela/nmt-chatbot>

## 12 Additional exemplary transcriptions of conversations

```
> what year is it?  
- 2019 [7.1]  
  
> what day is it?  
- It"s a good one [-86.25]  
  
> are you enjoying life?  
- I"m not really enjoying life. I"m not [-82.95]  
  
> are you depressed?  
- I"m not depressed. [14.2]  
  
> are you happy?  
- Yes [6.95]
```

```
> nice outfit!  
- Thanks! [12.55]  
  
> great hair  
- Thank you! [13.0]  
  
> I like your smile  
- I like you [8.0]
```

```
> you're my best friend  
- I"m not a fan of my friend. [15.55]  
  
> excuse me?  
- I"m not sure what [-85.95]
```

### 13 Declaration of Originality

I, Natasha Savic, hereby affirm that the attached work is entirely my own, except where the words or ideas of other writers are specifically acknowledged according to accepted citation conventions.

I give my permission to share a copy of this project upon request.