# POLITECNICO DI TORINO

**MASTER's Degree in ICT FOR SMART SOCIETIES**


**MASTER's Degree Thesis**

# VALIDITY OF Q&A METRICS THROUGH NLP MEASUREMENT PATTERNS

**Supervisors**

**Prof. MARCO TORCHIANO**

**Prof. SIMONE LEONARDI**

**Candidate**

**IRAKLIY DARZHANIYA**

**OCTOBER 2021**

# Summary

Nowadays the Artificial Intelligent (AI) systems as chatbots are becoming more popular in both business and private life. The main idea of the chat-bots is to recognize the text or the speech of the user, extract the needed information from the database and give back to the user the information he was looking for. This type of automatic question answering (QA) system can be implemented in the education field. The goal of this work is to prepare the Natural Language Processing (NLP) approach to develop in the future the chat-bot for the student's need. Specifically to understand how the already pretrained BERT large model finetuned on SQuAD can be efficient on the different Conversational Question Answering (COQA) dataset. This thesis can be considered as successful experimental work with a well-performed result of 84% of Accuracy.

# Acknowledgements

*"If you are going to try*
*go all the way*
*otherwise*
*don't even start"*
*Charles Bukowski*

# Table of Contents

# List of Tables

# List of Figures

# Acronyms

**AI**
Artificial Intelligence

**BERT**
Bidirectional Encoder Representations from Transformers

**COS SIM**
Cosine Similarity

**CQA**
Conversational Question Answering

**MRC**
Machine Reading Comprehension

**NLP**
Natural Language Processing

**QA**
Question-answer

**QT**
Question-text

**SQuAD**
Stanford Question Answering Dataset

**TF**
Term Frequency

# Chapter 1

# Introduction

This thesis reports an experimental work in one of the branches of Artificial Intelligent - Natural Language Processing. The primary idea of this research is to prepare a Question-Answering model for future usage in a chatbot. Students and professors will use the chatbot for a specific educational discipline. If a student has a question, the developed algorithm answers that question. The student's questions can be related to the discipline's materials, organization moments, information about the course. Automatic responses by computer will relieve the professors' duty to respond to the questions manually and accelerate the gathering of information for the students.

For this goal, we want to apply the already existing NLP model and do not create the model from scratch. It was possible lately only for research institutes because of the limits in the computational power. Nowadays, it is possible due to transfer learning. Transfer learning in the Machine Learning field means that the knowledge gained in the previous task can be applied to the recent and different but related problems.

From the all available pretrained models of Hugging Face [1], the BERT pretrained and finetuned on the SQuAD dataset has been chosen. We will use this model on another dataset-CoQA. This implementation gives us an idea about how well does the pretrained model work on the new dataset.

[1] https://huggingface.co/models

The aims of the thesis are:

1. to learn about pretrained BERT model, understand the techniques and architecture of the algorithm
2. to study the post-processing methods of data mining and decide which of them to use
3. to define the text-similarity approach
4. to implement all previously defined steps and manually check the questions' predictions, comparing them with the given answers
5. to analyse the results and set the goals for the future development

# Chapter 2

# State of the art

### 2.0.1 NLP tasks and Datasets

Natural Language Processing (NLP) — is one of the most complex problems of AI. NLP is consisted of two fields: Machine Learning and Linguistics. The interest in NLP has risen due to many commercial application in the market as IoT devices, which are able to understand human speech (Alexa,Siri,Google home assistant). Also, chatbots which is an integral part of web assistance, and so on. Even though, NLP field can be divided into six sub-fields with their tasks:

1. Text Classification function of which is to automatically label a text, sentence, email [1]. The algorithm should classify the text into some classes. For example, it is used in the spam filter.

2. Language Modeling. The primary function is to predict the next word of the sequence of the word [2]. This function is integrated into the search engines. When a user types the sentence in the search row in Google, Language Model gives the prediction of the next word.

3. Machine Translation. Machine translation is a task to translate the given text from one language to another [3]

4. Speech Recognition is the task of transforming human speech to a readable text [4].

5. Document Summarizing gives as the output the short description of the long text [5]

6. Question Answering function is to get the answer on the question based on the text passage [6]

Many free datasets exist for each specific task. This thesis work has a Question Answering task. Thus we will focus on the datasets related to QA problem:

1. CoQA
2. SQuADs datasets
3. QuAC

These datasets are new for question answering topics. In [7] they were compared along with three features :

1. Unanswerable questions
2. Multi-turn interactions
3. Abstract answers

CoQA and SQuAD's datasets are described in section 3.1. And now let us describe what is Conversational Question Answering.

### 2.0.2 Conversational Question Answering

The main idea of CQA (Conversational Question Answering) is to ask a machine a question based on the given text passage. Then, after receiving an answer, ask the machine another question based on the result that the device has already passed in the first turn. CQA is how humans gather information. The algorithm's capability of comprehension the natural language is evaluated and compared with human performance [1]. In some cases, the ability of a system to understand the natural language is higher than human's [8, 9]. Well-performed CQA systems have significant applications in different areas like customer service support. However, the task of CQA provides room for improvement for the researchers and developers.

[1] https://stanfordnlp.github.io/coqa/

### 2.0.3 Chat-bots review

Creating an intelligent dialogue algorithm with many functions, such as emulating human conversation and question answering on different topics, is one of the longest and most important goals in the Artificial Intelligence (AI) sector [10]. Nowadays, a considerable amount of conversational data has become accessible for free usage. Thus we can observe many works in this field [11, 12, 13, 14]. Furthermore, this field is subdivided into three directions [15]:

- Chat-oriented dialog system (Fig.2.1)
- Question-Answering dialog system (Fig.2.2)
- Task-oriented dialog system (Fig.2.3)

The Chat-oriented dialog system is used for the natural conversation between a computer and user (example-A.L.I.C.E. chatbot). The Task-oriented dialog system is needed for the task performing on the users' behalf (for example - booking a hotel/ trip organizer). Finally, the QA dialog system provides answers to users' questions based on the information in the database. The most well known dialog systems are:

- Amazon Alexa
- Apple Siri

These systems are not competent, and they need to be extensively researched.

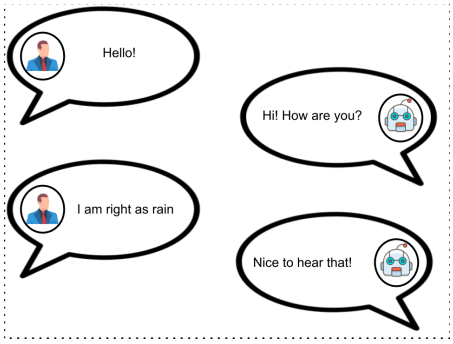


**Figure 2.1:** Example of the Chat-oriented dialog system

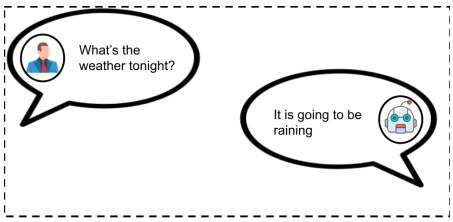


**Figure 2.2:** Example of the Question-Answering dialog system

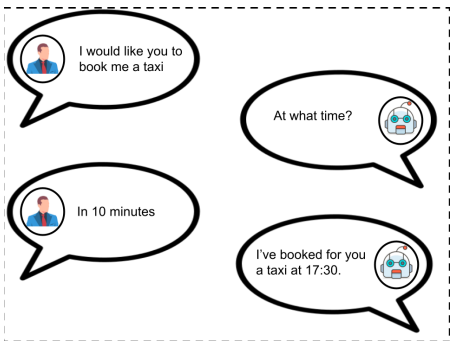


**Figure 2.3:** Example of the Task-oriented dialog system

### 2.0.4 Pretrained models

A pretrained model - is a model already created by a group of people to solve a particular problem. Instead of building a model from scratch, it is possible to use a pretrained model and start to develop the model for the specific case. This model could not be precise 100% in the developer's application, but it saves tremendous effort, time, and power.The pretrained models have been invented to not re-invent the wheel

Different kinds of pretrained models have brought outstanding performance gains to many NLP tasks, including MCR. The most well-known and common used models are:

- BERT [16]
- RoBERTa [17]
- DistilBERT [18]
- XLNET [19]
- GPT [20]

BERT ( Bidirectional Encoder Representations from Transformers) is a classic pretrained model which was designed to pre-train deep bidirectional representations from the unlabeled text by jointly conditioning on both left and right context in all layers. RoBERTa (a robustly optimized bert pre-training approach) is a clone study of BERT pre-training with adding many key hyperparameters and training data size. DistilBERT (distilled version of BERT) is a light and fast model trained by distilling BERT base. It has only 60% of the BERT parameters and runs 60% faster, achieving more than 95% of BERT's performance. XLNet is a generalized autoregressive pre-training model. The main difference of XLNet is that for each token the likelihood is calculated over all the tokens in the given text, and not only the words on the left and right. All the BERT's family models use Transformer's encoder. In contrast, GPT (Generative Pre-trained Transformer) uses a Transformer's decoder (Unidirectional Self-Attentive Model).

# Chapter 3

# Definitions

## 3.1 SQuAD datasets

First of all, we have to say that there are two SQuAD versions:

- SQuAD (2016 year)
- SQuAD v.2 (2018 year)

Both of these versions are very popular and important nowadays. This is because so many research works were based on them [1]. The SQuAD v.2 is the updated version of the SQuAD, even though we will discuss each of them in this section because of two reasons:

- The BERT model in this thesis work was finetuned on the SQuAD dataset. That is why it is mandatory to have an idea about this dataset
- There are some paper works [21] where SQuAD v.2 was compared with COQA dataset. As we worked on the COQA dataset, we had to know the difference between new datasets. In this way, it is possible to expect something from the results of the model implementation.

[1] https://rajpurkar.github.io/SQuAD-explorer/

### 3.1.1 SQuAD

#### Definition

The Stanford Question Answering Dataset (SQuAD) was created for reading comprehension task. This dataset has 107,785 question-answer pairs on 536 Wikipedia articles with a various types of questions and answers [22]. The distinctive feature of SQuAD compared to previous datasets (as MCTest, WikiQA, etc.) was the absence of a list of answers. In this case, a model has to find the answer based on all spans in texts and not only in provided ones (in the case of predating datasets).

#### Samples review

let us have a look at few samples from SQuAD [2] dataset (tab3.1). In the column "Context" there are text passages. Each text passage has many questions and answers. There is an answer position number that gives the position of the answer in the text. Indeed, in the first row, the answer position is 1. If we look at the first phrase of the context, we find "Michel Djotodia", which is matched with the answer.

| Context | Question | Answers | Answer Position | Question ID | Title |
|---|---|---|---|---|---|
| Micheal Djotodia took over as president..... | Who became president..? | Michel Djotodia | 1 | 36312 | Central African Republic |
| | What was Bozize..? | crimes against humanity | 233 | 36313 | |
| | What mass muder..? | genocide | 275 | 36314 | |
| | How many people...? | 200.000 | 379 | 36315 | |
| That same year the comedy Junior was... | What was Schwarz.. ... | Junior | 28 | 56782 | Arnold Schwarz ... |
| | How much did... | $150 million | 963 | 56783 | |

**Table 3.1:** Samples from SQuAD dataset

[2] https://datarepository.wolframcloud.com/resources/SQuAD-v1.1

### Answer type analysis

And now, let us consider the variation in answers in percentages:

- 20% of numeric answer type (Date, Numbers, etc.)
- 80% of textual answer type (Nouns, verb, and adjective phrases, names etc.)

The fifth part of answers is numeric, which is a large portion of the dataset. Textual part can be divided into four groups:

- Nouns phrases + Persons + Locations (50%)
- Verb phrases (5,5%)
- Adjective phrases (3,9%)
- Others (20%)

The most frequent answer type is the Noun type, which takes half of the entire dataset answers.

### Question analysis

Another essential key of the dataset is the question's variation. Study of the initial parts of questions gives us the concept of answers. If a question starts with "Who", then we expect a person as an answer. If a question starts with "When", the answer could be a date or time, and so on. More than half of the whole questions start with "What" (fig.3.1). It makes sense because half of the answers were nouns.

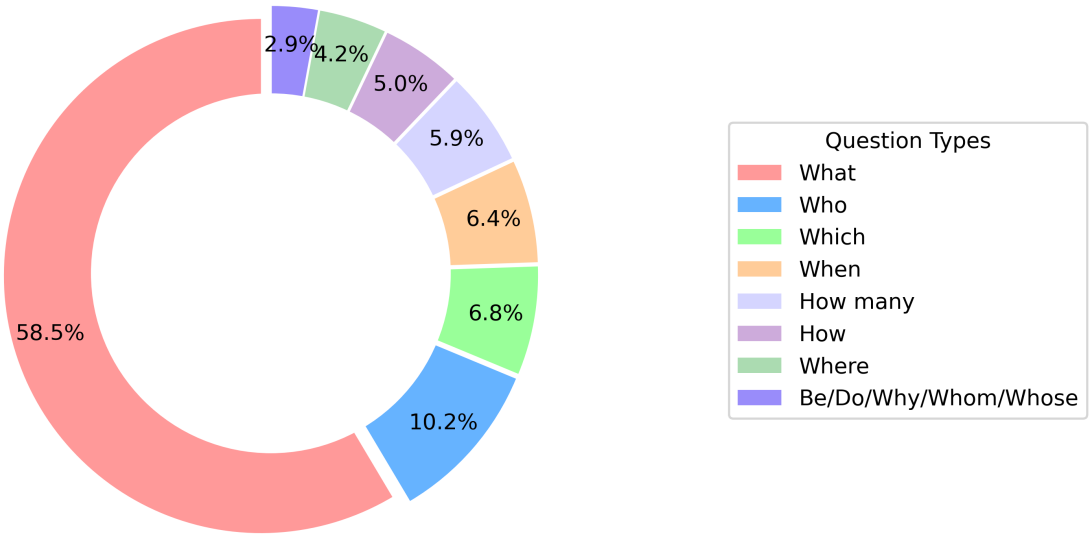


**Figure 3.1:** Question types in SQuAD

### 3.1.2 SQuAD v.2

#### Description

SQuAD v.2 is the extended version of SQuAD. More than 53,775 unanswerable questions were added to the new version of dataset [3]. Crowd workers wrote these 50k+ questions to look similar to answerable ones. In the previous version and other existing datasets, they pay their attention only to the answerable question. In contrast, SQuAD v.2 was created with two purposes [23] :

- answer the question if there is information related to the question in a text passage
- not to answer the question and refrain from it, in the case when the information related to a question is not provided in a text passage.

let us consider an example of answerable question [24] taken from SQuAD v.2:

**Title**: Victoria (Australia)

**Paragraph**: . . . Public schools, also known as state or government schools, are funded and run directly by the Victoria Department of Education. Students do not pay tuition fees, but some extra costs are levied...

**Answerable question** : What organization runs the public schools in Victoria?

**Unanswerable question** : What organization runs the waste management in Victoria?

**Plausible answer** : Victoria Department of Education

These two questions are similar from a linguistic point of view. Moreover, the expected answer type is the same.

[3] https://rajpurkar.github.io/SQuAD-explorer/

### Sample's structure

The structure of SQuAD v.2 samples has a bit changed compared to samples from SQuAD v.1. New column "Answerable" was added (Tab.3.2). The values can be either "True" or "False". This information is provided only in train set because, as we mentioned before, one of the tasks for models is to understand either a question is answerable or not.

### Question types

The variation of question types in SQuAD v.2 (Fig.3.2) is very similar to the question's distribution in SQuAD (Fig.3.1). The differences in each sector range in 2%. We will discuss the question types in detail in the next section.

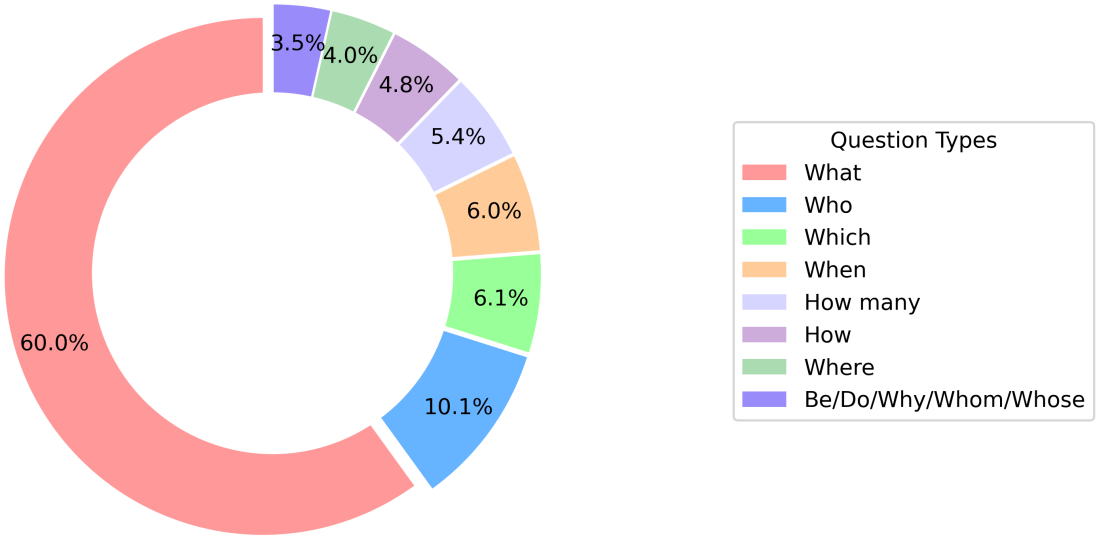


**Figure 3.2:** Question types in SQuAD v.2

| Context | Question | Answers | Answer Position | Question ID | Answerable | Title |
|---|---|---|---|---|---|---|
| Manhatted Island is linked to New York ... | Which tunnel do.. | The Lincoln Tunnel | 105 | 3631 | True | New York City |
| | The Holland tunnel opened in what year? | 1927 | 615 | 3632 | True | |
| | The Queens-Midtown tunnel was finished in what year? | 1940 | 811 | 3633 | True | |
| | Who was the first person... | President Franklin D. Roosvelt | 817 | 3664 | True | |
| | How many vehicles... | 120,000 | 139 | 3665 | True | |
| Since at least the time of Ancient Greeks... | When did the advent.. | 18th century | 427 | 57271 | True | Jews |
| | What was a result... | growing trend if assimilation | 691 | 57272 | True | |
| | Name a Jewish company.. | Kaifeng Jews of China | 335 | 57273 | True | |
| | Where did assimilation ... | in all areas | 246 | 57274 | False | |
| | What Jewish comunity.. | Kaifeng Jews of China | 335 | 57275 | False | |

**Table 3.2:** Structure of SQuAD v.2 samples

## 3.2 COQA dataset

### Description

The COQA dataset is a new dataset for Conversational Question Answering system. Mainly, this dataset was created to estimate the computer's potential to be engaged in a conversational question-answering style[25]. Thus COQA dataset was developed with three aims:

1. Human-style conversation.
2. Naturalness of the answers.
3. Robust across different domains.

### Human-style conversation

First of all, the Human conversation style is based on remembering a dialogue history. It means that all the questions except the first one are dependent on the previous questions and answers. Let us take a random example from the COQA dataset. In Tab3.3 we can see that there is a dialogue of 2 persons based on the text passage, where:

- $Q_i$ -is the i-th question in the conversation
- $A_i$ -is the answer to i-th question
- $R_i$ - is a rationale which supports the answer $A_i$

**Text passage**: (CNN) – Lewis Hamilton extended his Formula One drivers' championship lead after finishing second behind Red Bull's Mark Webber at the British Grand Prix. World champion Jenson Button, who narrowly missed out on his first podium finish at Silverstone after coming fourth, still trails McLaren teammate Hamilton in second. Third-placed Webber stormed back into title contention after winning his third race of the season. The Australian leapfroged fellow Red Bull driver Sebastian Vettel, who is 24 points adrift of Hamilton in fourth. McLaren also lead Red Bull by 29 points at the top of the constructors' championship. Ferrari's Fernando Alonso stayed fifth overall but lost ground after earning no points, ending the race in 14th after being given a drive-through penalty for illegally overtaking Robert Kubica of Renault off the track. Nico Rosberg of Germany continues to outperform his Mercedes teammate Michael Schumacher, recording his third podium finish this season to replace Kubica in sixth....

| | |
|---|---|
| Q1 | What sport does Lewis Hamilton compete in? |
| A1 | Formula One |
| R1 | Formula One |
| | |
| Q2 | Did he compete in the British Grand Prix? |
| A2 | yes |
| R2 | British Grand Prix |
| | |
| Q3 | What did he place? |
| A3 | second |
| R3 | second |
| | |
| Q4 | Behind whom? |
| A4 | Mark Webber |
| R4 | Mark Webber |
| | |
| Q5 | How many races has he won this season? |
| A5 | 1 |
| R5 | one |
| | |
| Q6 | For what brand does he drive? |
| A6 | unknown |
| R6 | unknown |
| | |
| Q7 | Who is another driver for that brand? |
| A7 | Mark Webber |
| R7 | Mark Webber |
| | |
| Q8 | Which team is ahead of Red Bull? |
| A8 | McLaren |
| R8 | McLaren |

**Table 3.3:** An example of human conversation in COQA dataset

Indeed, to answer Question 2 (Did he complete in the British Grand Prix?), we must know whom we are talking about. And this information we can take from the first pair question-answer (Q1-A1).

### Naturalness of the answers

Also, human conversations consist of free-form answers. The answers can be either a proper sentence or just a one-word answer or "Yes" or "No". The presence of short answers in a conversation - is an essential part of the human discussion style. That is why this dataset is rich in short answers. More precisely, we will speak about it in the Characteristics session.

As we saw from the example of one conversation (Tab. 3.3), there are Rationales (R) for each question-answer pair. The concept of this dataset is to select a Rationale from the text span and only after the selection to adjust the Rationale to the free-form answer.

### Robust across different domains

The COQA dataset consists of 127k question-answering conversation pairs based on 8k text passages. The average conversation has 15 QA pairs. Each answer has a span-based rationale highlighted in the text passage. This dataset covers seven domains, while other datasets are concentrated on one specific field only. The domains are:

1. Wikipedia
2. Children's Stories
3. News
4. Middle and High School Exams
5. Literature
6. Science (out-of-domain)
7. Reddit (out-of-domain)

Five of them are in-domain, while Science and Reddit are out-of-domain. Let us specify from which source the text passage has been taken. The articles were selected from Wikipedia [4], children's stories from [26], CNN news from [27], Reddit articles from [28], science articles from [29], English exams from [30] and litterature from Gutenberg website [5]. In case of long articles, they have been truncated up to 200 words.

[4]https://en.wikipedia.org/wiki/Main_Page

[5]https://www.gutenberg.org/

### COQA dataset division

The whole dataset was divided into three sets:

- 100 text passages of each five in-domains are in the validation set
- 100 texts in the test set
- rest in the training set

The complete information on in-domains splitting we can find in Tab.3.4. The most considerable portion of data is on school exams and news. A little bit less is on literature and Wikipedia. The minor part is children's stories.

| **Domain** | **Number of texts** | **Number of QA pairs** | **Average passage length** | **Average number of QA pairs per passage** |
|---|---|---|---|---|
| Clildren's stories | 750 | 10,5k | 211 | 14 |
| Literature | 1815 | 25,5k | 284 | 15,6 |
| School Exams | 1911 | 28,6k | 306 | 15 |
| News | 1902 | 28,7k | 268 | 15,1 |
| Wikipedia | 1821 | 28k | 245 | 15,4 |

**Table 3.4:** Information about domains

Now let us see the distribution of these five domains in the training set (Fig.3.3) and validation set (Fig.3.4). For the training set, the distribution of text passages of the five topics are the same as in Table 3.4, while in the validation dataset, the portions of data are equal.

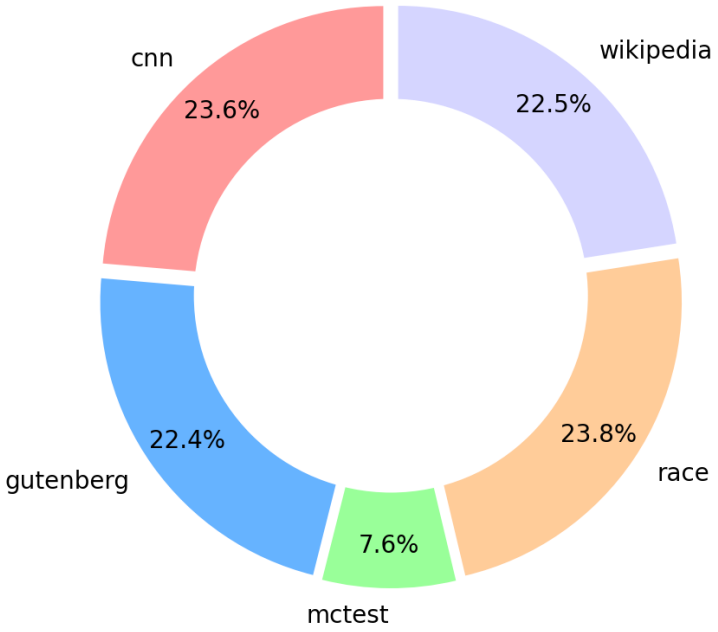


**Figure 3.3:** Distribution of five in-domains text passages in the training dataset

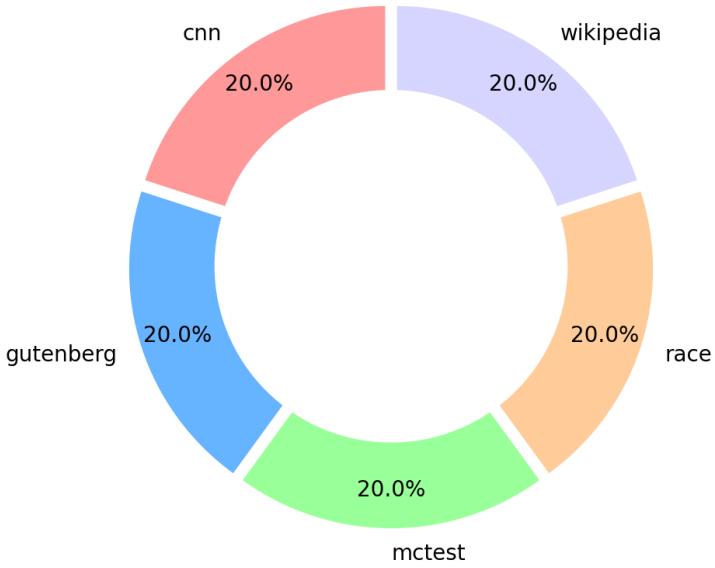


**Figure 3.4:** Distribution of five in-domains text passages in the validation dataset

## 3.3 COQA and SQUAD comparison

In the previous sections, SQuAD, SQuAD v.2 and COQA datasets were presented. It is time to deep into details, so let us compare two datasets: SQuAD v.2 with COQA.

### Question's types comparison

At first, we start with the analysis of question varieties of the datasets. COQA has a free-form of questions and answers, while SQuAD has not. This difference will lead to the variance in prefixes of questions. Let us have a look at Fig.3.5 and Fig.3.6. Half of the question starts with "WHAT" in SQuAD, while in COQA just the third part. COQA has a wider variety of questions than SQuAD. Some of them are present in COQA and missed in SQuAD (for instance: "Did").

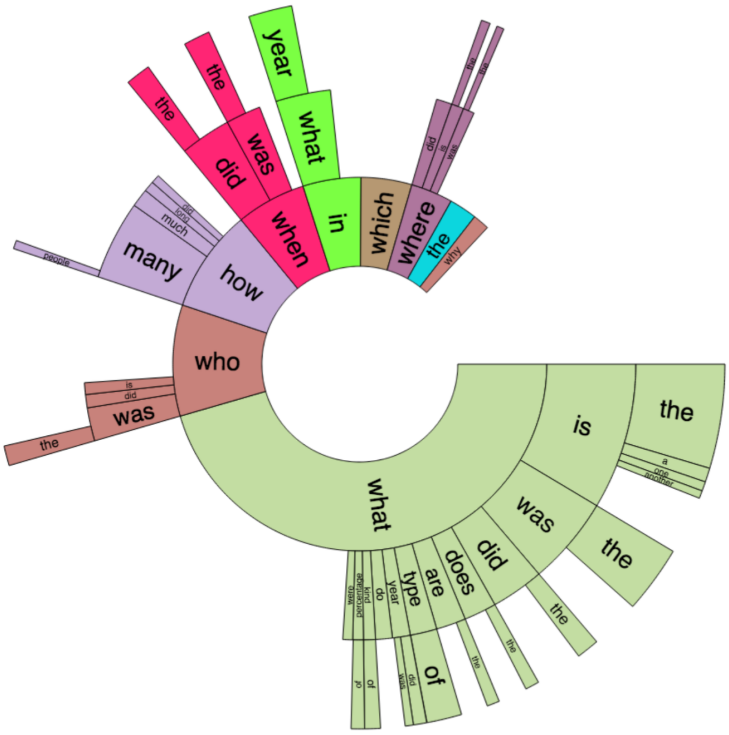


**Figure 3.5:** Question prefixes in SQuAD v.2

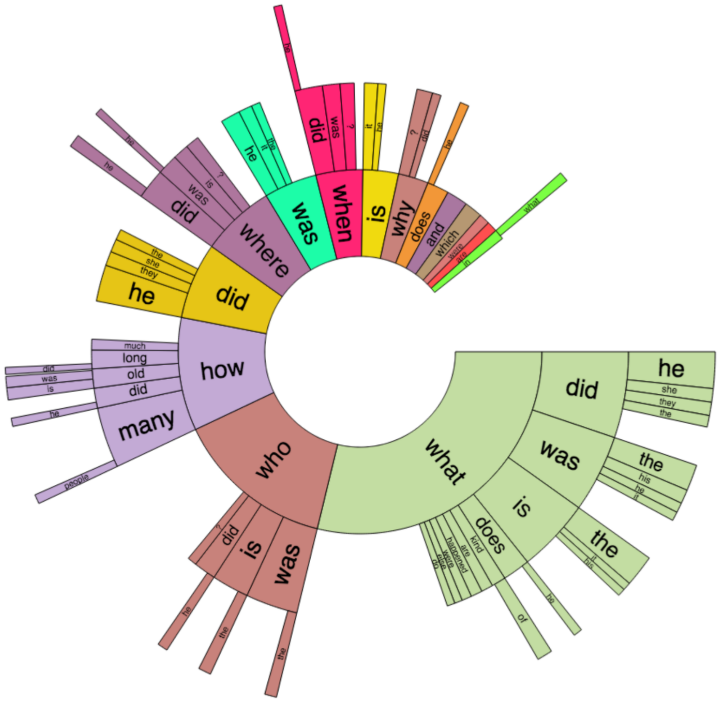


**Figure 3.6:** Question prefixes in COQA

### Lengths comparison

The average passage length in SQuAD is 117, while in COQA is 271. Nevertheless, the question length in COQA is twice less than in SQuAD: 5,5 and 10,1 respectively. The reason for that is free-form. The average length of answers in COQA (2,7) is less than in SQuAD (3,2) as well.

### Answers types comparison

Several name's entities and nouns phrases answers prevail in both datasets (Fig.3.7): 69% in SQuAD and 48.3% in COQA. However, the main difference is in "YES/NO" answers.The SQuAD does not have them at all, meantime in COQA, this is the fifth part of all answers. Both "Numbers" and "Date/Time" answers in SQuAD are twice bigger than in COQA.

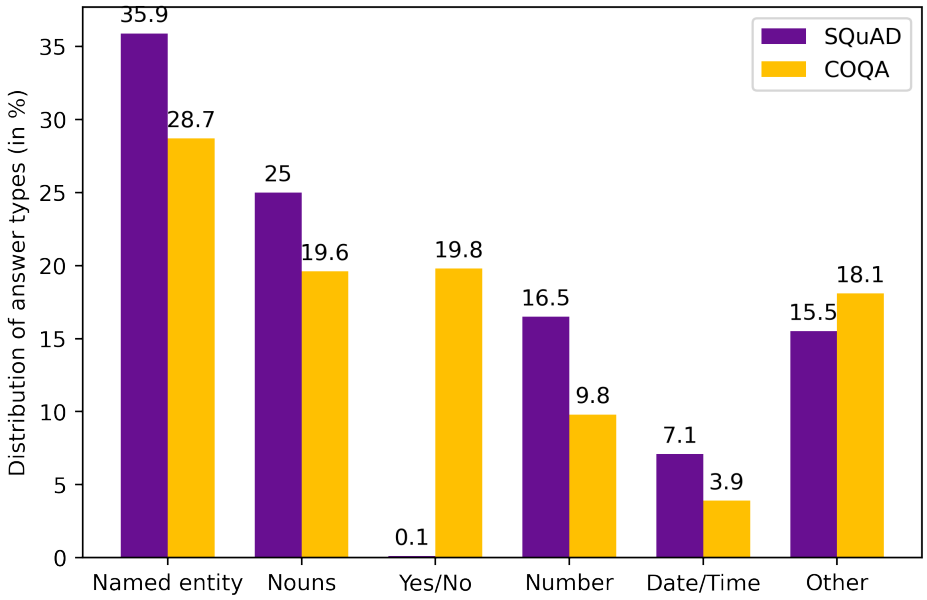


**Figure 3.7:** Distribution of answer types in SQuAD and COQA

## 3.4 BERT

### 3.4.1 What is BERT?

BERT (Bidirectional Encoder Representations from Transformers) is a popular machine learning language technique based on Transformers. Nowadays, the Transformer is the supreme architecture for natural language processing [31]. The transformer is a mechanism with attention which means that this deep learning model learns contextual relations between words. Weights of these words are dynamically calculated depending on the connection between the words. For example, when we want to read a text quickly, we use the skimming technique, in which we skip less important words and focus only on significant words in a text. Thus our visual system highlights the most informative words. Let us consider an example from Wikipedia [6]:

- Original text : "Wheatfield with Crows" is a July 1890 painting by Vincent van Gogh
- Skimming : "Wheatfield with Crows" is a July 1890 painting by Vincent van Gogh

[6]https://en.wikipedia.org/wiki/Wheatfield_with_Crows

In the Skimming row, the most important words are marked with red color. In these words, we pay our attention. Analogously to the human visual system, the mechanism with attention provides weights to the words. The essential words have the biggest weights [32].

Now let us speak about the main BERT difference from other language representation models, which can only read the text in one direction: either from left to right or from right to left, but can not read simultaneously a text passage in both directions simultaneously. Nevertheless, BERT can read in two directions at the same time. This capability gives BERT the possibility to pre-train BERT using two tasks [16] :

- Masked Language Model
- Next Sentence Prediction

Masked Language Model training aims to mask a word from the sentence and then predict the target word based on the masked word's context. Next Sentence Prediction training aims to predict the relationship between two sentences.

### 3.4.2 How does BERT work?

#### Architecture

The BERT architecture parameters are changeable. They depend on the model's id [7]. The main idea of working processes of all models are the same, but it is convenient to explain how BERT works on the specific model. In this thesis work, "bert-large-uncased-whole-word-masking-finetuned-squad" model was used. The parameters of the given model are [8]:

- Model is uncased (all capital letters are transformed into small letters)
- Whole Word Masking technique
- English language
- 24 layers
- 1024 hidden size
- 16 attention heads
- 336M total parameters
- Fine-tuned on SQUAD

#### Inputs preparing for BERT

The first thing we should do is prepare Question-Text pairs taken from a dataset for BERT. The structure of a BERT's input has to have this kind of form:

[CLS] QUESTION [SEP] TEXT PASSAGE [SEP]

, where CLS and SEP are important tokens, which are specially needed for the the BERT model. Every Question-Text pair starts with CLS (this token is necessitate for classification), then there are SEP tokens at the end of both question and text. These tokens are needed to separate the question from the text.

As we can see from Fig. 3.8,a BERT input consists of three embeddings:

- position embedding (the upper row)
- segment embedding (the central row)

[7] https://huggingface.co/transformers/pretrained_models.html

[8] https://huggingface.co/bert-large-uncased-whole-word-masking-finetuned-squad

- token embedding (the lower row)

Bert model understands the position of a given word, using position embedding. The segment embedding marks the question tokens with "A"s, while the tokens related to text - with "B"s. Token embedding is taken from the WordPiece token vocabulary.

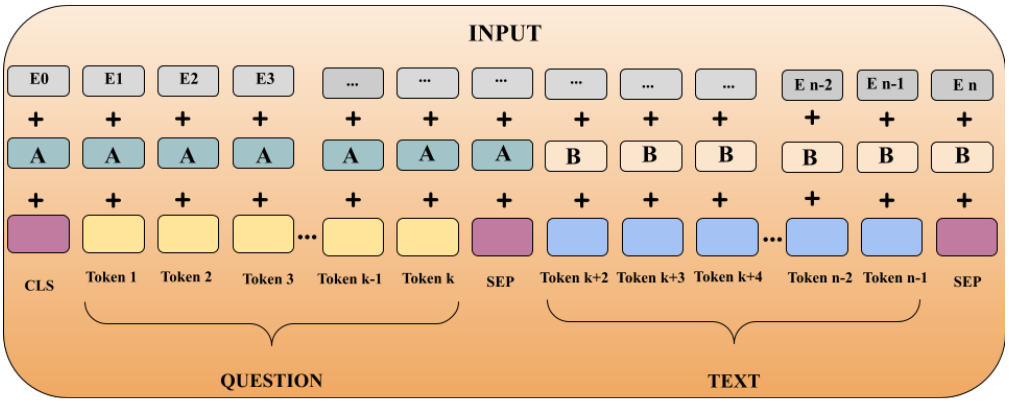


**Figure 3.8:** BERT input structure

The last phase of BERT implementation is feeding the model. The prediction will be given only if the most probable start token is before the most probable end token. Otherwise, the algorithm is not able to give us an answer.

## 3.5 Data mining

### 3.5.1 Tokenization

Tokenization is the first step in NLP [33]. The function of this method is to chop the sentences up into tokens. A token can be either a word, punctuation or even a part the word. For instance, let us have a look at the tokenization of COQA's question (Fig.3.9). As input, there is a question: "How old was the diary?". After tokenization, there is no plain text anymore. The text has been split into six chunks - tokens.

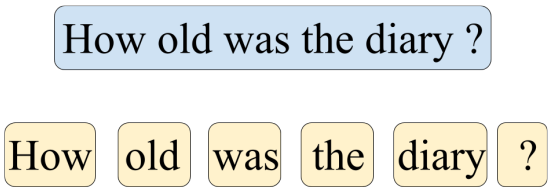


**Figure 3.9:** An example of the Tokenization on the question from COQA

This approach allows us to work with each token separately and use different techniques to achieve the best performance of the model. For example, if we want to measure the text similarity, we have to split sentences into tokens and then work with them using different strategies. Also, we need to keep in mind that machines do not understand lexical data. Algorithms need numerical data. For that reason, vector representation of the word is needed [34]. But in this thesis work we use another technique called TF-IDF (section 3.5.5).

### 3.5.2 Stop Words Removal

Stop Words Removing is a classic method to filter out a text from the less significant words. It helps to raise the accuracy of the predictions. Let us consider the example of two texts before stop words removing (Tab.3.5):

| | | |
|---|---|---|
| Before Stop Words Removal | **Text A** | The youngest brothers do play football on Mondays |
| | **Text B** | My brothers play football with friends on Mondays |
| After Stop Words Removing | **Text A** | brothers play football Mondays |
| | **Text B** | brothers play football friends Mondays |

**Table 3.5:** Two text examples before and after Stop Words Removing

We can see that the Stop Words : "do", "the", "my", "with", "on" have been eliminated. If we count the number of words in the intersection and then divide by the number of unique words presented in both texts, we achieve the accuracy score of text similarity. In case of the original texts (before the Stop Words Removing) the accuracy is equal 45%, while after Stop Words Removing the accuracy is higher and equals to 66%.

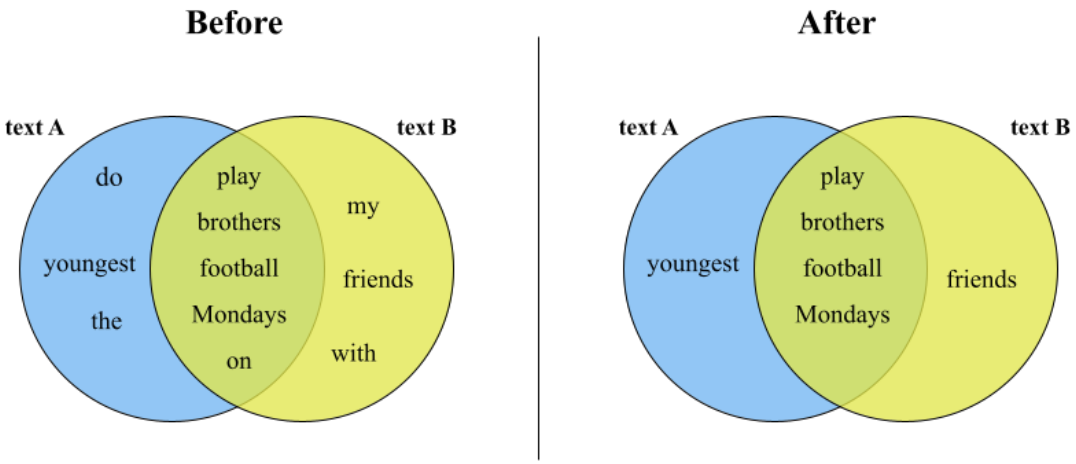


**Figure 3.10:** Euler circles before and after Stop Words Removing

### 3.5.3 Lemmatization

Lemmatization is another important approach in NLP. It is used to convert a word in his lemma (root word), reducing inflectional forms. In table 3.6 we can see different examples of how the lemmatization approach changes the words. These word transformations are significant in a text similarity problem. As in this thesis work we obtain the answer predictions, and the predictions can be slightly different from the given answer, thus lemmatization will be very useful.

| **Word** | **Lemma** |
|---|---|
| Was | Be |
| Working | Work |
| Cars | Car |
| Better | Good |

**Table 3.6:** Examples of lemmatization

### 3.5.4 Cosine Similarity

We are going to use the Cosine Similarity to understand how the two texts are similar. It is often used to measure document similarity in text analysis [35].

The Cosine Similarity is the measure of similarity between 2 vectors [9]. It equals to the ratio between a dot product of two given vectors and the product of their lengths.

$$\cos\theta = \frac{(\vec{a}, \vec{b})}{|\vec{a}||\vec{b}|} \tag{3.1}$$

As you can see from the formula, we find the cosine of the angle between two vectors. It defines how close the two vectors are directed to each other. The cosine ranges [10] from 0 to 1:

- $\cos\theta = 0$ means that two vectors are perpendicular to each other, thus they are the least similar
- $\cos\theta = 1$ means that two vectors have the same direction, thus they are the most similar

[9] https://en.wikipedia.org/wiki/Cosine_similarity

[10] By definition of the cosine, it ranges from -1 to 1, but in case of text similarity problem we work only with positive vectors, thus the cosine can not be the negative. Why we only use the positive non-zero vectors will be discussed in the next section "Tf-Idf"

If we want to find the similarity of the texts using the Cosine Similarity, we have to convert the text to the vector called a term-frequency vector. In the next section, we are going to define TF-IDF method.

### 3.5.5 TF - IDF Method

Term frequency-inverse document frequency (TF-IDF) is one of the most important methods in information retrieval and text mining [36]. The schema of TF-IDF is based on term-weight. The main idea of this method is to indicate how a word is important in a document or text in a collection of documents. This method is a combination of two concepts:

- Term frequency (TF) shows the frequency of a word in one specific document from the corpus of documents
- Inverse document frequency (IDF) shows the number of documents in which this word appears in the corpus

Suppose we have a document collection $D$ (the size of this corpus is $|D|$), and we want to calculate TF-IDF of a word $W$ in a document $d$, where $d$ belongs to the corpus $D$. The formula will have the form:

$$W_d = f_{W,d} * log(\frac{|D|}{f_{W,D}}) \tag{3.2}$$

, where $f_{W,d}$ is TF, and $\frac{|D|}{f_{W,D}}$ - IDF.

### 3.5.6 Text Similarity

There are a lot of Text Similarity techniques [37]. In this work the Text Similarity is based on two approaches :

1. TF method
2. Cosine Similarity valuation

Let us show how it works on the given example. There are two texts that we have seen it recently (Tab.3.7). The task is to evaluate their text similarity. At first, TF method will be implemented, the result of which is two vectors (Tab.3.8). The vector element is 1, if the word is present in the text, 0 if the word is absent in the text. After the Cosine Similarity has been evaluated (0.45).

| | |
|---|---|
| **Text A** | The youngest brothers do play football on Mondays |
| **Text B** | My brothers play football with friends on Mondays |

**Table 3.7:** Two texts example for text similarity evaluation

| **Word** | **Vector** | | **A*B** |
|---|---|---|---|
| | **A** | **B** | |
| The | 1 | 0 | 0 |
| youngest | 1 | 0 | 0 |
| brothers | 1 | 1 | 1 |
| do | 1 | 0 | 0 |
| play | 1 | 1 | 1 |
| football | 1 | 1 | 1 |
| on | 1 | 1 | 1 |
| Mondays | 1 | 1 | 1 |
| with | 0 | 1 | 0 |
| friends | 0 | 1 | 0 |
| my | 0 | 1 | 0 |
| | | | **COS SIM = 0.45** |

**Table 3.8:** Cosine Similarity of two vectors after TF implementation

# Chapter 4

# Implementation

## 4.1 Data selection

This work was based on two datasets:

1. training dataset [1] with 7199 text passages
2. validation dataset [2] with 500 test passages

As we mentioned before, there are 15 QA pairs per passage on average. Nevertheless, in this thesis work was taken only 1 QA pair for each passage. Thus the total amount of Question-Answering pairs is 7199 for the training dataset and 500 for validation dataset. The reason of this choice is to focus on the post-processing techniques rather than to predict the 15 answers for each passage. Thus, after downloading the data and structured it in a dataframe format (Tab.4.1), we can start to prepare the inputs for BERT implementation.

| id | Text passage | Question | Answer |
|---|---|---|---|
| 0 | Once upon a time, in a barn... | What color was Cotton? | white |
| 1 | Once there was a beautiful... | what was the name of the fish.. | Asta. |
| 2 | My doorbell rings. On the step | Who is at the door? | An elderly... |
| 3 | (CNN) – Dennis Farina, the ... | Is someone in showbiz? | Yes. |
| 4 | Kendra and Quinton travel to.. | Where do Quinton and Kendra.. | school |

**Table 4.1:** The first five samples of the training dataframe

[1] http://downloads.cs.stanford.edu/nlp/data/coqa/coqa-train-v1.0.json

[2] http://downloads.cs.stanford.edu/nlp/data/coqa/coqa-dev-v1.0.json

## 4.2 Bert Implementation

### 4.2.1 Prepare inputs for BERT

As we discussed in section 3.4.2, the input structure for BERT model has this order:

[CLS] QUESTION [SEP] TEXT PASSAGE [SEP]

let us consider a random example of Question-Text pair from the training COQA dataset:

- QUESTION : Who was in charge of FIFA?
- TEXT : "Cristiano Ronaldo provided ... 6 points behind Barca."

As we can see in Fig.4.1, there are two rows of tokens sequences. The first one is consisted of "A"s and "B"s. It is another necessary form for BERT, which is called "segment embedding". All tokens regarding the Question part are marked with "A", including CLS and SEP right after the question. In contrast, the text's tokens are marked with B. In the code these tokens marks are 0 and 1 instead of A and B. The next row represents the Question-Text pair divided into tokens with important tokens CLS and SEP. In this phase, the first step of preparing input for the BERT model is finished, and all the Question-Text pairs are ready to be encoded.

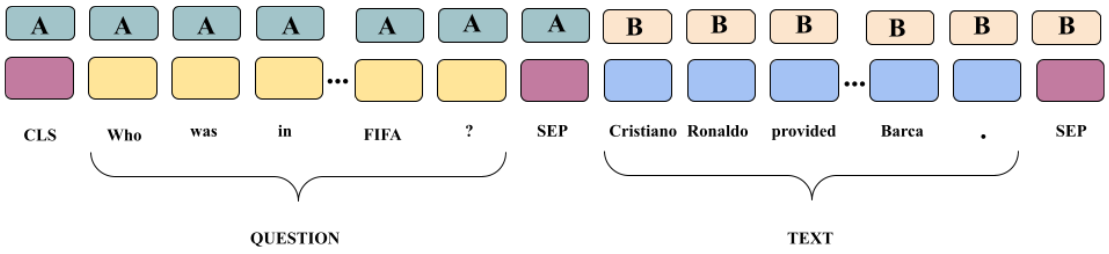


**Figure 4.1:** Example of an input of Question-Text pair for BERT implementation

### 4.2.2 Encoding the Question-Text pairs

As we can see from Tab4.2 BERT gives an ID to each token. All tokens have been lower-cased because of chosen uncased BERT model. Also, we can see the strange symbols: "##". This is the particularity of the BERT model, which is called "WordPiece" subword tokenization algorithm [3]. The main idea of this method is to divide rare tokens into pieces based on probability. This gives the model an advantage - reduced vocabulary for training performance. It is also possible that "##" symbols occur in predictions, but in this case, these symbols will be eliminated and subtokens will merge to the one token.

[3]https://huggingface.co/transformers/tokenizer_summary

| **TOKEN** | **TOKEN'S ID** |
|---|---|
| [CLS] | 101 |
| who | 2040 |
| was | 2001 |
| in | 1999 |
| charge | 3715 |
| of | 1997 |
| fifa | 5713 |
| ? | 1029 |
| [SEP] | 102 |
| cr | 13678 |
| ##ist | 2923 |
| ##iano | 15668 |
| ronald | 8923 |
| ##o | 2080 |
| provided | 3024 |
| ... | ... |
| ... | ... |
| ... | ... |
| 6 | 1020 |
| points | 2685 |
| behind | 2369 |
| bar | 3347 |
| ##ca | 3540 |
| . | 1012 |
| [SEP] | 102 |

**Table 4.2:** Token Id representation in BERT model

## 4.3 Post Processing Data Cleaning

After BERT implementation on the COQA dataset, we obtained the column "Prediction" (Fig.4.2). There are four columns : Text, Question, Answer and Prediction. To understand how good BERT performs on the given dataset we have to compare answers with prediction. Humans can distinguish easily if two phrases are similar, but for machines, this is not so clear. Let us consider some examples from (Fig.4.2). We can find an exact match only in row 9. Only in this case, the machine gives an output in which the answer and prediction match. Other answer-prediction pairs will not be recognized by machines as similar ones. There are many reasons why it is like that. For example :

- the presence of different words (row 15)
- the presence of additional information (rows 0, 1, 8)
- the difference in lower and upper case letters (rows 11, 18, 19)
- and so on..

All the issues related to comparing answers and predictions lead us to implement some data cleaning methods to understand how accurate the answers and the predictions match up.

| | text | question | answer | prediction |
|---|---|---|---|---|
| 0 | The Vatican Apostolic Library (), more commonl... | When was the Vat formally opened? | It was formally established in 1475 | 1475 |
| 1 | New York (CNN) -- More than 80 Michael Jackson... | Where was the Auction held? | Hard Rock Cafe | hard rock cafe in new york ' s times square |
| 2 | CHAPTER VII. THE DAUGHTER OF WITHERSTEEN \n\n"... | What did Venters call Lassiter? | gun-man | gun - man |
| 3 | (CNN) -- The longest-running holiday special s... | Who is Rudolph's father? | Donner | donner |
| 4 | CHAPTER XXIV. THE INTERRUPTED MASS \n\nThe mor... | Who arrived at the church? | the garrison first | monna valentina |
| 5 | Have you ever been to some big cities in the w... | Was Budapest always one city? | no | budapest became one city in 1872 |
| 6 | (CNN) -- A lawsuit filed by the family of Robe... | WHO IS FILING THE LAWSUIT? | The family of Robert Champion | the family of robert champion |
| 7 | Officials of the Chicago Transit Authority sai... | who recently had heart surgery? | Nicole Hobson | nicole hobson |
| 8 | Local businessmen are increasingly facing comp... | What is a valuable service? | brick and mortar stores | we spend 30 minutes to an hour with somebody a... |
| 9 | The University of Chicago (UChicago, Chicago, ... | When was teh University established? | 1890 | 1890 |
| 10 | An incandescent light bulb, incandescent lamp ... | What is the energy source for an incandescent ... | a wire filament | electric current through it , until it glows w... |
| 11 | Did you know that Albert Einstein could not sp... | What age did Einstein start talking? | Four | four |
| 12 | Traditionally considered the last part of the ... | What did Neolithic follow? | Holocene Epipaleolithic period | holocene epipaleolithic period |
| 13 | Poultry (/ˈpoʊltri/) are domesticated birds k... | When was poultry first domesticated? | several thousand years ago. | several thousand years ago |
| 14 | Mr. Laurence was not allowed to see Beth, and ... | Was Meg telling her Mom about Beth being sick? | Meg felt unhappy writing letters to her mother... | meg felt unhappy writing letters to her mother... |
| 15 | Once an Englishman named Jack Brown went to Ru... | What did Jack shoot? | A wolf | the first wolf |
| 16 | A Chinese actor's divorce from his wife, over ... | What is the name of the Chinese actor in the s... | Wang Baoqiang | wang baoqiang |
| 17 | Once upon a time in Greece, there lived a youn... | Did somebody did? | Yes | villagers would turn up on the streets to star... |
| 18 | Laura and Graham were having a party for their... | Who was throwing a party? | Laura and Graham | laura and graham |
| 19 | The Oscars ceremony at the 87th Academy Awards... | Who does Birdman star? | Michael Keaton | michael keaton |
| 20 | Czechoslovakia or Czecho-Slovakia (; Czech and... | Was Czechoslovakia ever apart of the Soviet bloc? | yes | from 1948 to 1990 |

**Figure 4.2:** First 20 examples of the training dataset just after the BERT implementation

This is a list of the methods which were implemented in this phase of the data cleaning:

- Lower case
- Punctuation and Space Elimination
- The Diacritics Elimination
- "Yes" "No" elimination
- Number to word conversion
- Lemmatization
- Stop Words Removing
- Answer presence in predictions
- Prediction presence in answers

### 4.3.1 Raw data

From now, the columns "Text" and "Question" will not be shown anymore in the following tables because we have to focus on the comparison between Answers and Predictions, while the information in the "Text" and "Question" columns is not relevant during the data cleaning faze. As we can see in table 4.3, there are three columns: Row, Answer and Prediction. The represented data is shown in a raw format, just after the BERT implementation. It is essential to mention that there is only one exact answer-prediction pair in row 9. All the other pairs are different. Some of them are slightly dissimilar, and some are entirely mismatched.

Let us perform the Cosine Similarity with the TF-IDF approach on the raw data (Tab.4.4) to have an idea of how similar the answer/prediction pairs are. **The average Cosine Similarity** over the whole dataset is equal to **0.2406**. As we mentioned before, only the 9th row has the Cosine Similarity equal to 1. If we tale a glance at the rows: 4,6,7,11,12,14,16,18,19,25,28,29, we can see that the Cosine Similarity equals to 0, while the texts are almost the same. The only difference is that some of the words begin with capital letters. Thus the first data cleaning step will be Lower Case implementation.

| ROW | ANSWER | PREDICTION |
|---|---|---|
| 0 | It was formally established in 1475 | 1475 |
| 1 | Hard Rock Cafe | hard rock cafe in.. |
| 2 | gun-man | gun - man |
| 3 | Donner | donner |
| 4 | the garrison first | monna valentina |
| 5 | no | budapest became one city in 1872 |
| 6 | The family of Robert Champion | the family of robert champion |
| 7 | Nicole Hobson | nicole hobson |
| 8 | brick and mortar stores | we spend 30 minutes to... |
| 9 | 1890 | 1890 |
| 10 | a wire filament | electric current through... |
| 11 | Four | four |
| 12 | Holocene Epipaleolithic period | holocene epipaleolithic period |
| 13 | several thousand years ago. | several thousand years ago |
| 14 | Meg felt unhappy writing ... | meg felt unhappy writing... |
| 15 | A wolf | the first wolf |
| 16 | Wang Baoqiang | wang baoqiang |
| 17 | Yes | villagers would turn... |
| 18 | Laura and Graham | laura and graham |
| 19 | Michael Keaton | michael keaton |
| 20 | yes | from 1948 to 1990 |
| 21 | yes | learning |
| 22 | tree trimmer | michigan tree trimmer |
| 23 | Yes | donny love |
| 24 | Dad | my dad |
| 25 | Wang Jiaming | wang jiaming |
| 26 | A giants | i ’ m a giant |
| 27 | 5th century | after the roman withdrawal .. |
| 28 | Chevrons | chevrons |
| 29 | Formula One | formula one |

**Table 4.3:** Comparison between Answers and Predictions for the first 30 examples after BERT implementation.

| ROW | ANSWER | PREDICTION | COS |
|---|---|---|---|
| 0 | It was formally established in 1475 | 1475 | 0.40 |
| 1 | Hard Rock Cafe | hard rock cafe in.. | 0 |
| 2 | gun-man | gun - man | 0 |
| 3 | Donner | donner | 0 |
| 4 | the garrison first | monna valentina | 0 |
| 5 | no | budapest became one city in 1872 | 0 |
| 6 | The family of Robert Champion | the family of robert champion | 0.4 |
| 7 | Nicole Hobson | nicole hobson | 0 |
| 8 | brick and mortar stores | we spend 30 minutes to... | 0.12 |
| 9 | 1890 | 1890 | 1 |
| 10 | a wire filament | electric current through... | 0.12 |
| 11 | Four | four | 0 |
| 12 | Holocene Epipaleolithic period | holocene epipaleolithic period | 0.33 |
| 13 | several thousand years ago. | several thousand years ago | 0.89 |
| 14 | Meg felt unhappy writing ... | meg felt unhappy writing... | 0.75 |
| 15 | A wolf | the first wolf | 0.40 |
| 16 | Wang Baoqiang | wang baoqiang | 0 |
| 17 | Yes | villagers would turn... | 0 |
| 18 | Laura and Graham | laura and graham | 0.33 |
| 19 | Michael Keaton | michael keaton | 0 |
| 20 | yes | from 1948 to 1990 | 0 |
| 21 | yes | learning | 0 |
| 22 | tree trimmer | michigan tree trimmer | 0.81 |
| 23 | Yes | donny love | 0 |
| 24 | Dad | my dad | 0 |
| 25 | Wang Jiaming | wang jiaming | 0 |
| 26 | A giants | i ’ m a giant | 0 |
| 27 | 5th century | after the roman withdrawal .. | 0.47 |
| 28 | Chevrons | chevrons | 0 |
| 29 | Formula One | formula one | 0 |

**Table 4.4:** First 30 raw answer-prediction pairs and their cosine similarity

### 4.3.2 Step 1 : Lower Case

After Lower Case performing, the **Average Cosine Similarity** was almost tripled and equaled to **0.6167**. Indeed, in all previous rows (4,6,7,11,12,14,16,18,19,25,28,29) the capital letters has been changed with the small letters (Table 4.5). Thus the Cosine Similarity of these pairs have been changed from 0 to 1 as well. Also, a new column "Label" was added to the table. This column is made for assigning the right predictions. If the label equals to 1, then the prediction is right from the semantic point of view.

**The rule for assigning the prediction to the label 1 is:**

- if the Cosine Similarity of the answer-prediction pair $> 0.9$, then we assume that the prediction is right and allocate the label = 1
- if the Cosine Similarity of the answer-prediction pair $< 0.9$, then we leave the label cell empty

After the marking of each answer-prediction pair, we can count the Accuracy as follows:

$$Accuracy = \frac{\sum(\text{rows where label=1})}{\text{number of all rows in dataset}} \tag{4.1}$$

**The Accuracy = 0.4232**, considering the "Lower case" method.

In the previous case of Raw Data, **The Accuracy = 0.1069**. As we can see the Accuracy has increased by 4 times.

| Row | Answer | Prediction | Cos | Label |
|---|---|---|---|---|
| 0 | it was formally established in 1475 | 1475 | 0.40 | |
| 1 | hard rock cafe | hard rock cafe in.. | 0.54 | |
| 2 | gun-man | gun - man | 0 | |
| 3 | donner | donner | 1 | 1 |
| 4 | the garrison first | monna valentina | 0 | |
| 5 | no | budapest became... | 0 | |
| 6 | the family of robert champion | the family of robert champion | 1 | 1 |
| 7 | nicole hobson | nicole hobson | 1 | 1 |
| 8 | brick and mortar stores | we spend 30 minutes .. | 0.12 | |
| 9 | 1890 | 1890 | 1 | 1 |
| 10 | a wire filament | electric current.. | 0.12 | |
| 11 | four | four | 1 | 1 |
| 12 | holocene epipaleolithic period | holocene epipaleolithic period | 1 | 1 |
| 13 | several thousand years ago. | several thousand years ago | 0.89 | |
| 14 | meg felt unhappy writing ... | meg felt unhappy writing ... | 0.89 | |
| 15 | a wolf | the first wolf | 0.40 | |
| 16 | wang baoqiang | wang baoqiang | 1 | 1 |
| 17 | yes | villagers would... | 0 | |
| 18 | laura and graham | laura and graham | 1 | 1 |
| 19 | michael keaton | michael keaton | 1 | 1 |
| 20 | yes | from 1948 to 1990 | 0 | |
| 21 | yes | learning | 0 | |
| 22 | tree trimmer | michigan tree trimmer | 0.81 | |
| 23 | yes | donny love | 0 | |
| 24 | dad | my dad | 0.70 | |
| 25 | wang jiaming | wang jiaming | 1 | 1 |
| 26 | a giants | i ' m a giant | 0.31 | |
| 27 | 5th century | after the roman .. | 0.47 | |
| 28 | chevrons | chevrons | 1 | 1 |
| 29 | formula one | formula one | 1 | 1 |

**Table 4.5:** First 30 answer-prediction pairs after Lower Case implementation.

### 4.3.3 Step 2 : Punctuation and Space Elimination

Sometimes the punctuation can be crucial for text similarity. Each punctuation mark is considered as an important element in the text. We can see in table 4.6, how important the presence of punctuation in the text is. There is no doubt that the answers and predictions are similar, but any punctuation element ruins the Cosine Similarity. In addition, the presence of spaces between words is just as important as punctuation marks. As we can see in row 2, the Cosine Similarity equals 0 due to unnecessary spaces in the prediction "gun-man".

| Row | Answer | Prediction | COS SIM | Label |
|---|---|---|---|---|
| 2 | gun-man | gun - man | 0 | |
| 37 | anna's parents | anna ' s parents | 0.57735 | |
| 39 | jeremy lin, | jeremy lin | 0.816497 | |
| 149 | noah, | noah | 0.707107 | |
| 153 | i'll have another | i ' ll have another | 0.67082 | |
| 171 | lieutenant gulston. | lieutenant gulston | 0.816497 | |
| 174 | robo sally. | robo sally | 0.816497 | |
| 188 | chapter xxv. | chapter xxv | 0.816497 | |
| 241 | mohawk guy, | mohawk guy | 0.816497 | |
| 244 | valentine's day | valentine ' s day | 0.57735 | |

**Table 4.6:** Cosine similarity before punctuation and space elimination

Let us see the same previous 10 examples after punctuation and space elimination (table 4.7).The results are improved because all of the answer-prediction pairs are exactly matched. **The Average Cosine Similarity** raised to **0.6557**, while **Accuracy** has increased up to **0.4948**

| Row | Answer | Prediction | COS SIM | Label |
|---|---|---|---|---|
| 2 | gun man | gun man | 1 | 1 |
| 37 | anna s parents | anna s parents | 1 | 1 |
| 39 | jeremy lin | jeremy lin | 1 | 1 |
| 149 | noah | noah | 1 | 1 |
| 153 | i ll have another | i ll have another | 1 | 1 |
| 171 | lieutenant gulston | lieutenant gulston | 1 | 1 |
| 174 | robo sally | robo sally | 1 | 1 |
| 188 | chapter xxv | chapter xxv | 1 | 1 |
| 241 | mohawk guy | mohawk guy | 1 | 1 |
| 244 | valentine s day | valentine s day | 1 | 1 |

**Table 4.7:** Cosine similarity after punctuation and space elimination

### 4.3.4 Step 3 : Presence of diacritics in the answers

In some cases, the Cosine Similarity can be less than 1 because of the letter with an apostrophe, dot, stress, or other symbols. In truth, all the predictions (Tab.4.8) are semantically correct, but there is no answer/prediction pair with Cosine Similarity equal to 1. Indeed, in rows: 91, 680, 1047, 1454 there is only one difference - the diacritics presence in the answer, and this distinction makes the Cosine Similarity less than 1. After the diacritics elimination, the Cosine Similarity has been changed to 1 as it was expected (Tab.4.9).

**The updated average Cosine Similarity = 0.6570, while Accuracy = 0.4968**

| ROW | ANSWER | PREDICTION | COS |
|---|---|---|---|
| 91 | napoléon bonaparte | napoleon bonaparte | 0.5 |
| 97 | córdoba al andalus | cordoba | 0 |
| 680 | café montmartre | cafe montmartre | 0.5 |
| 1047 | señor ramirez | senor ramirez | 0.5 |
| 1235 | grande île | the grande ile grand island | 0.35 |
| 1454 | kurt gödel | kurt godel | 0.5 |
| 1539 | erik möller | wikimedia foundation | 0 |
| 2044 | the río de la plata | rio de la plata | 0.75 |
| 2971 | gdańsk bay | gdansk bay | 0.5 |
| 3064 | alexis gonzález | omoa honduras cnn alexis gonzalez | 0.31 |
| 3338 | françois hollande | francois hollande | 0.5 |
| 3373 | ange félix patassé | ange felix patasse | 0.33 |
| 3605 | bärn | the city of bern | 0 |
| 4822 | mélanie joly | melanie joly | 0.5 |
| 5042 | álvaro de mendaña | alvaro de mendana | 0.33 |
| 5493 | the greek mīkros bios and logia | greek mikros small bios life and logia | 0.61 |
| 5594 | real madrid and atlético de madrid | real madrid and atletico de madrid | 0.75 |
| 5676 | frédéric auguste bartholdi | frederic auguste bartholdi | 0.66 |
| 5860 | anabella de león | anabella de leon | 0.66 |
| 6238 | sebastião de melo | sebastiao de melo | 0.66 |
| 6385 | juan de bermúdez got it first | juan de bermudez | 0.51 |
| 7039 | the greek κανών | greek κανων | 0.5 |

**Table 4.8:** The presence of diacritics in the answers

| ROW | ANSWER | PREDICTION | COS |
|---|---|---|---|
| 91 | napoleon bonaparte | napoleon bonaparte | 1 |
| 97 | cordoba al andalus | cordoba | 0.57 |
| 680 | cafe montmartre | cafe montmartre | 1 |
| 1047 | senor ramirez | senor ramirez | 1 |
| 1235 | grande ile | the grande ile grand island | 0.63 |
| 1454 | kurt godel | kurt godel | 1 |
| 1539 | erik moller | wikimedia foundation | 0 |
| 2044 | the rio de la plata | rio de la plata | 0.89 |
| 2971 | gdansk bay | gdansk bay | 1 |
| 3040 | productores de musica de espana | spanish albums chart ... | 0 |
| 3338 | francois hollande | francois hollande | 1 |

**Table 4.9:** Cosine similarity after diacritics elimination

### 4.3.5 Step 4 : Yes/No elimination

One of the differences between COQA and SQUAD datasets is the presence of "YES" and "NO" in COQA's answers. These types of answers occur in the "answer" column quite frequently (row 5, 17, 20, 21, 23 in Tab 4.4. However, in the "prediction" column, there are no "YES" or "NO". As we mentioned before, we used the pretrained BERT model, which was fine-tuned on the SQUAD dataset. It means that we can not expect "YES" / "NO" in the predictions because they were not presented in the SQUAD dataset. That is why we can remove the rows with "YES" "NO" answers from our data training and proceed with other techniques.

After removing these rows, the new shape of the data was reduced by 6,9 % from the original size of the data frame and equaled to 6520. The new Cosine similarity after removing the "YES"/"NO" rows is 0.7025. The Accuracy is 0.5309

### 4.3.6 Step 5 : Number to word

As we can see in the table(4.10) the answers and predictions can be represented either in a numerical format or in a word format. All Answer-Prediction pairs in this table have different formats. Even if the prediction is correct but differs from the given answer just with a format, the Cosine Similarity equals 0. This phenomenon has to be changed.

| Row | Answer | Prediction | COS SIM |
|---|---|---|---|
| 287 | eight | 8 | 0 |
| 852 | 19 | nineteen | 0 |
| 1304 | 11 | eleven | 0 |
| 1383 | 10 | nine | 0 |
| 1422 | 14 | fourteen | 0 |
| 1961 | 45 | 45th | 0 |
| 2545 | four | 4 | 0 |
| 5400 | 16 | sixteen | 0 |
| 5737 | 24 | nineteen | 0 |
| 6498 | 20 | twenty | 0 |

**Table 4.10:** Cosine Similarity before Number to word conversion

Let us convert numbers to words (Tab.4.11). In that case, the situation has improved, but a new problem is occurred - ordinal numbers. Indeed, comparing all rows except the row 1961 in two tables (4.10 and 4.11, we can see that the numbers have been represented. The Cosine Similarity for these cases is equal to 1. The pair in the Row 1961 contains the ordinal number. It means that this number has not been converted to a word because of ordinal suffixes. Hence, before converting the number to a word, we need to eliminate ordinal suffixes and implement the Number To Word method.

After eliminating the ordinal suffixes and converting the numbers to the words (Tab.4.12, we improved a bit **The Average Cosine Similarity 0.7069** and **the Accuracy 0.5352**.

| Row | Answer | Prediction | COS SIM |
|---|---|---|---|
| 287 | eight | eight | 1 |
| 852 | nineteen | nineteen | 1 |
| 1304 | eleven | eleven | 1 |
| 1383 | ten | nine | 0 |
| 1422 | fourteen | fourteen | 1 |
| 1961 | fourty five | 45th | 0 |
| 2545 | four | four | 1 |
| 5400 | sixteen | sixteen | 1 |
| 5737 | twenty four | nineteen | 0 |
| 6498 | twenty | twenty | 1 |

**Table 4.11:** Presence of ordinal numbers

| Row | Answer | Prediction | COS SIM |
|---|---|---|---|
| 287 | eight | eight | 1 |
| 852 | nineteen | nineteen | 1 |
| 1304 | eleven | eleven | 1 |
| 1383 | ten | nine | 0 |
| 1422 | fourteen | fourteen | 1 |
| 1961 | fourty five | fourty five | 1 |
| 2545 | four | four | 1 |
| 5400 | sixteen | sixteen | 1 |
| 5737 | twenty four | nineteen | 0 |
| 6498 | twenty | twenty | 1 |

**Table 4.12:** Cosine Similarity after converting all kind of number representations in words

### 4.3.7 Step 6 : Lemmatization

The lemmatization method was implemented in this way: First, we have to tokenize the answers and predictions. Then each token has to be lemmatized independently. After lemmatization, we have a list of the lemmatized words for every answer and prediction. Hence we need to build back the sentence from this list of words with the original order of the words. For instance:

- Prediction before lemmatization: "hard rock cafe in new york s **times** square"
- The list of tokens:[hard,rock,cafe,in,new,york,s,times,square]
- Lemmatization: [hard,rock,cafe,in,new,york,s,time,square]
- Finally lemmatized the prediction: "hard rock cafe in new york s **time** square"

As we can see, the order of the words in a Prediction phrase is the same as in the lemmatized result. The only difference is that the word "times" has been transformed into "time". There are other six examples in Table 4.13.

| **Answer or Prediction** | |
|---|---|
| **Before Lemmatization** | **After Lemmatization** |
| it was formally established in... | it wa formally established in... |
| in new york s times square | in new york s time square |
| brick and mortar stores | brick and mortar store |
| we spend thirty minutes | we spend thirty minute |
| villagers would turn up | villager would turn up |
| a giants | a giant |

**Table 4.13:** Comparison between phrases before and after lemmatization

At first, these changes seem to be insignificant, but even such little changes lead to an improvement in the efficiency of the algorithm: **The Average Cosine Similarity** after lemmatization is **0.7091**, while **Accuracy** is **0.5365**

### 4.3.8 Step 7 : Stop Words Removing

This approach affects the Cosine Similarity and the Accuracy because the number of words in Answers and Predictions decreases. Indeed, if we look at the Table 4.14 we can see how the Cosine Similarity has increased. All the articles and useless words were eliminated, and as the result the Cosine Similarity for each pair was roughly doubled. **The Average Cosine Similarity** slightly increased to the value **0.7295**, while **Accuracy** reached **0.5993**.

| Stop Words Removing | Row | Answer | Prediction | COS SIM |
|---|---|---|---|---|
| **Before** | 15 | a wolf | the first wold | 0.40 |
| | 26 | a giants | i m a giant | 0.35 |
| | 38 | into attra | attra | 0.7 |
| | 53 | the palace | the palace of qui si sane | 0.31 |
| **After** | 15 | wolf | first wolf | 0.70 |
| | 26 | giant | giant | 1 |
| | 48 | attra | attra | 1 |
| | 53 | palace | palace sane | 0.7 |

**Table 4.14:** Cosine Similarity before and after Stop Words Removing

It is important to mention that in this work was used only the list of English Stop Words. This parameter can be crucial in some cases. Let us consider the pair from the last section:

- Original pair:
  - Answer: gun-man
  - Prediction : gun - man
- After Stop English Words Removing and punctuation elimination:
  - Answer: gun man
  - Prediction : gun man

The result is exact match. But what if we use the whole corpus of the stop words? let us see what's happened:

- Original pair:
  - Answer: gun-man
  - Prediction : gun - man

- After Stop (all) Words Removing and punctuation elimination::
  - Answer: gun man
  - Prediction : gun

The result is not the same as before. The problem was in the Prediction word "man", which is considered as a stop word in some language from the whole NTLK corpus of the stop words. The result of using the whole corpus is less efficient, thus we have to use only the English stop words.

### 4.3.9 Step 8: Answer presence in prediction

Let us look at the first 20 rows of the dataset (Tab.4.15), where the given answers are present in the predictions. We can see the common feature in all these examples:

- a short free-form in answers (a peculiarity of COQA)
- long formal style in the prediction phrases

The presence of this difference is expected because the algorithm was not pretrained on the COQA dataset. However, it does not mean that the algorithm performs badly, not at all. The goal was to obtain the right prediction from a semantic point of view. Thus the answers and predictions can be slightly different. Indeed, let us consider the first row:

- The question was: "Where was the Auction held?"
- The answer after all data processing : "hard rock cafe"
- While the prediction : "hard rock cafe new york time square"

The obtained prediction is much more detailed than the given answer. Moreover, if we look at other examples, we can be convinced of it. In the 15th row, we get additional information about the wolf. In the 103rd row, we obtained an extra description of the student. In row 158, the name Jack was provided, and so on. In the 107th row, the prediction was too long to fit in the column (a part of the sentence was hidden).

Considering all the above, we set the new rule: if there is a complete presence of not empty answer in the prediction, then we consider this pair as right from a semantic point of view and assign it to label 1. **The Cosine Similarity** has not changed and equals to **0.7295**, but the **Accuracy** is risen up by 22% and equals to **0.7319**. The Cosine Similarity is the same because we did not modify anything in the predictions or answers. We have just assigned some pairs with label 1, which affects only the Accuracy.

| ROW | ANSWER | PREDICTION | COS |
|---|---|---|---|
| 1 | hard rock cafe | hard rock cafe new york time square | 0.65 |
| 15 | wolf | first wolf | 0.70 |
| 22 | tree trimmer | michigan tree trimmer | 0.81 |
| 27 | five century | roman withdrawal britain five century | 0.63 |
| 42 | english civil war | english civil war religious conflict | 0.65 |
| 46 | torpenhow | hiawatha torpenhow | 0.70 |
| 53 | palace | palace qui si sane | 0.5 |
| 61 | yan ran angel foundation | yan ran angel foundation harelipped.. | 0.81 |
| 64 | rape capital punishment | rape capital punishment big business | 0.70 |
| 68 | daily news | tuesday daily news | 0.81 |
| 98 | speculator | speculator neighboring council bluff .. | 0.44 |
| 99 | malian | malian island | 0.70 |
| 101 | build | build arbor around front door | 0.44 |
| 103 | student | fourteen year old junior student wuhan | 0.40 |
| 107 | hate crime | one count violating | 0.40 |
| 114 | copenhagen | copenhagen capital populous city | 0.44 |
| 122 | charles simony | charles simonyi | 0.5 |
| 135 | acob burn film center | jacob burn film center | 0.75 |
| 150 | venice | venice italy | 0.70 |
| 158 | broxton | jack broxton | 0.70 |

**Table 4.15:** The answer presence in prediction

### 4.3.10 Step 9 : Prediction presence in answer

Just before we analyzed the presence answer in the prediction, and it is logical to consider a reverse scenario : the presence of non-empty predictions in answers. In fact, if we see at examples in (Tab. 4.16) we can see that the situation is very similar with respect to the previous case. For instance, let us consider the last row:

- The question: "What does Jose Mourinho do for a living?"
- The answer after all data processing : "portuguese football coach"
- While the prediction : "football coach"

There is a full presence of the prediction in the answer. The difference between them is the adjective "portuguese", which is just the additional information. The luck or presence of the additional words as adjectives and adverbs should not affect the semantically correct prediction. Thus, the last rule in our post-processing pipeline is: if there is a complete presence of not empty prediction in the answer, we consider this pair as right from a semantic point of view and assign it to label 1. The percentage of the cases defined by the last rule equals to 5,8%. **The new Accuracy is equal to 0.7912**

| **ROW** | ANSWER | PREDICTION | COS |
|---|---|---|---|
| 0 | established one thousand four hund... | one thousand four hundred seventy five | 0.81 |
| 52 | meadow mouse | mouse | 0.70 |
| 67 | around five hundred | five hundred | 0.81 |
| 76 | thriving automobile manufacturin... | automobile manufacturin.. | 0.70 |
| 81 | revenge win | revenge | 0.70 |
| 82 | secret behind mona lisa smile | mona lisa smile | 0.77 |
| 87 | toy room | room | 0.70 |
| 108 | home north carolina state university | north carolina state university | 0.89 |
| 128 | sound distant gun | distant gun | 0.81 |
| 170 | strict mother | strict | 0.70 |
| 175 | two official radio | two | 0.33 |
| 189 | george edgar | george | 0.70 |
| 195 | alone church door | church door | 0.81 |
| 196 | engineering degree | engineering | 0.70 |
| 197 | twenty twenty five territory claim | twenty twenty five | 0.70 |
| 206 | weaker | weak | 0 |
| 232 | tthe school | school | 0.70 |
| 254 | american theoretical physicist | theoretical | 0.57 |
| 266 | entrance kaviri hut | kaviri hut | 0.81 |
| 272 | portuguese football coach | football coach | 0.81 |

**Table 4.16:** The Prediction presence in Answers

## 4.4 Results and Analysis

After implementing all Post Processing steps, we can see how the Cosine Similarity and Accuracy have been changed. Let us start with the Cosine Similarity. The visualization of the growth of Average Cosine Similarity is shown in Fig.4.3, while this graph's numerical information and description are in Tab.4.17. From the graph, we can see a spike after the first Post Processing Step - "Lower case". From the first step to the sixth, the Cosine Similarity has approximately linear growth. From the seventh to the ninth step, the value has not changed as it was discussed in "Step 8: Answer presence in prediction". Comparing the initial Cosine Similarity with the final, we can infer that it has increased three times.

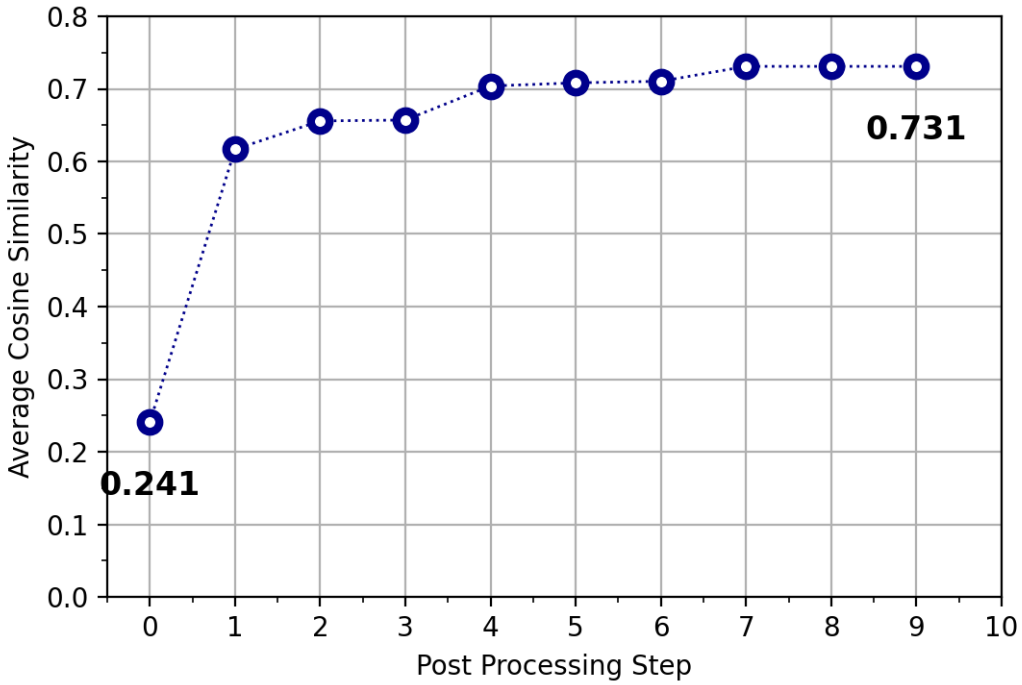


**Figure 4.3:** The growth of Average Cosine Similarity during all post processing steps in the training phase

The detailed information is contained in Tab.4.17, with 4 columns:

- The order of the Step
- The description of the Step
- Average Cosine Similarity
- Improvement (in percentages) with respect to the Cosine Similarity in the previous step

| STEP | DESCRIPTION | COS SIM | IMPROVEMENT |
|---|---|---|---|
| 0 | Raw data | 0.2406 | |
| 1 | Lower case | 0.6167 | +156% |
| 2 | Punctuation and space elimination | 0.6557 | +6,3% |
| 3 | Diacritics elimination | 0.6570 | +0.01% |
| 4 | Yes/No elimination | 0.7039 | +7,1% |
| 5 | Number to word | 0.7082 | +0,06% |
| 6 | Lemmatization | 0.7104 | +0,03% |
| 7 | Stop words removing | 0.7309 | +2,8% |
| 8 | Answer presence in prediction | | |
| 9 | Prediction presence in answer | | |

**Table 4.17:** Cosine Similarity

Now, let us consider the most important value, which shows the efficiency of the methods - Accuracy. Accuracy's growth information is given in numerical format (Tab.4.18) and visualization form (Fig.4.4). There is a huge difference between the first Accuracy 0.1069, and the last Accuracy 0.7947. The function of the graph(Fig.4.4) is similar to the previous function, especially in the beginning. Nevertheless, in the case of Accuracy, there is an increase after the sixth post-processing step.

After achieving the results, we can follow the validation phase, using the defined post-processes without monitoring each step.

| STEP | DESCRIPTION | ACCURACY | IMPROVEMENT |
|---|---|---|---|
| 0 | Raw data | 0.1069 | |
| 1 | Lower case | 0.4232 | +295% |
| 2 | Punctuation and space elimination | 0.4948 | +16,9% |
| 3 | Diacritics elimination | 0.4968 | +0.01% |
| 4 | Yes/No elimination | 0.5331 | +7,3% |
| 5 | Number to word | 0.5374 | +0,08% |
| 6 | Lemmatization | 0.5386 | +0,02% |
| 7 | Stop words removing | 0.6018 | +11,7% |
| 8 | Answer presence in prediction | 0.7351 | +22,1% |
| 9 | Prediction presence in answer | 0.7947 | +8,1% |

**Table 4.18:** The improvement of Accuracy during all post processing steps in training phase

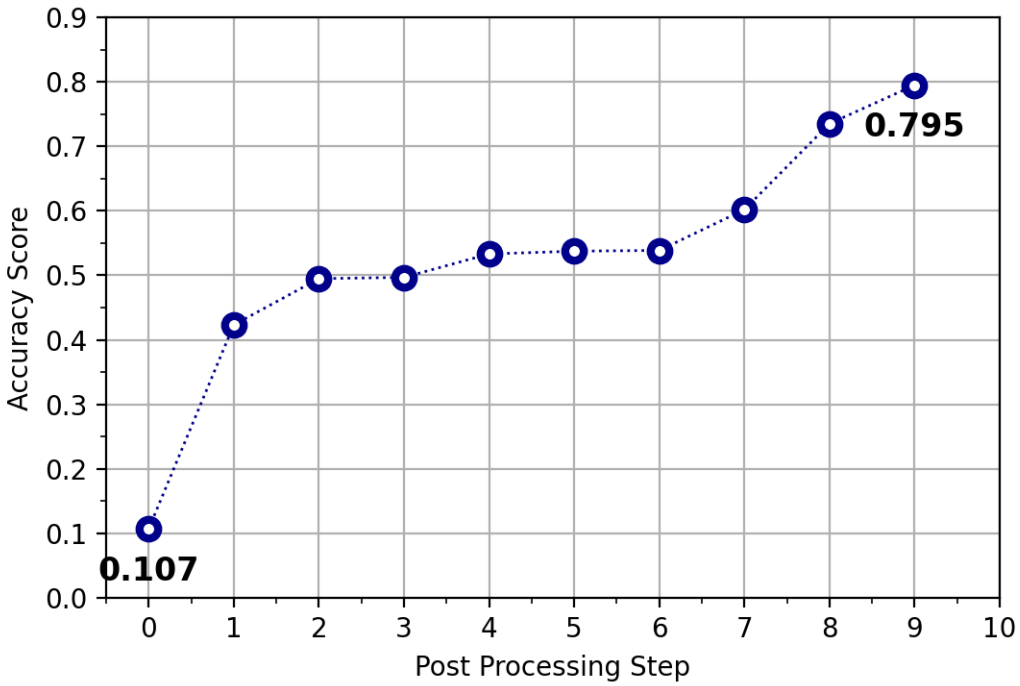


**Figure 4.4:** The growth of Accuracy during all post processing steps in the training phase

## 4.5 Validation

After defining all post-processing steps, it is time to implement them on the validation dataset. The Average Cosine Similarity reached 0,77 (Fig.4.5), which is 4% higher than the result on the training set. The Accuracy score in the validation phase is higher as well and equals 83,6%.

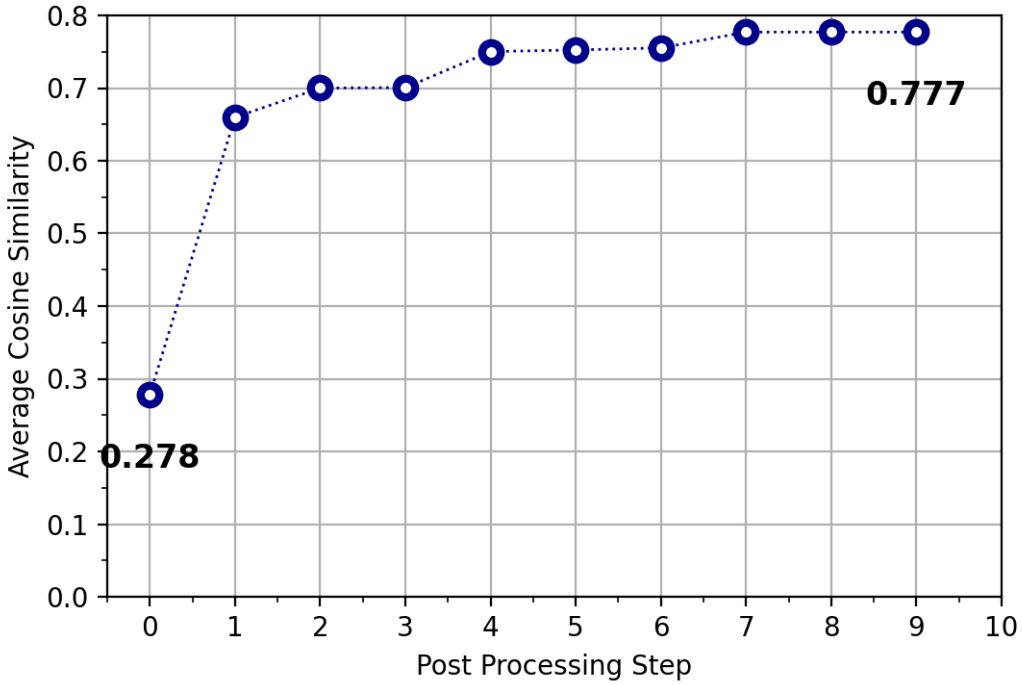


**Figure 4.5:** The growth of Average Cosine Similarity during all post processing steps in the validation phase

As the size of the validation dataset is 490 samples, it was possible to check and compare each answer-prediction pair manually. This will give us the possibility to count True Positives, True Negatives, False Positives, and False Negatives:

- True Positive = 382
- False Positive = 1
- True Negative = 69
- False Negative = 6

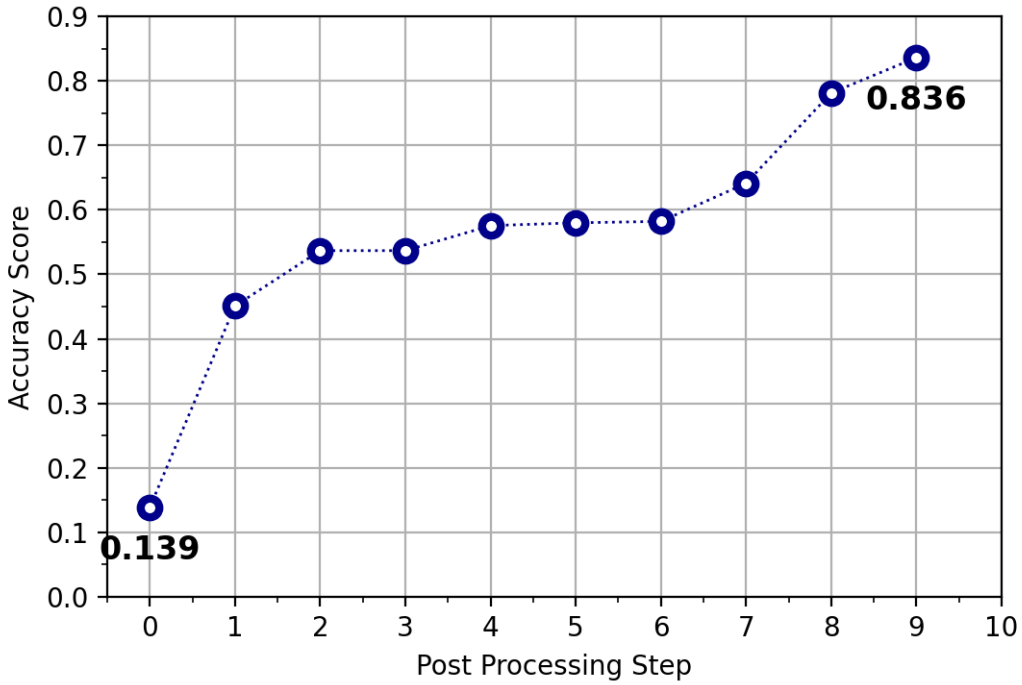


**Figure 4.6:** The growth of Accuracy during all post processing steps in the validation phase

Positive class is a case when prediction is correct, and Negative class means that prediction is wrong. From 383 Positive cases, only 1 sample is False, and all other pairs are True. We can see the example of the first 25 True Positives samples in Fig.B.1. let us consider this unique False Positive case:

**Text**: "We do a great deal too much, aunt," she said. "I am almost coming round to my father's opinion. You know, Mr.Maddison, he very seldom comes to London, and then only when he wants to pay a visit to his gunmaker, or to renew his hunting kit, or something of that sort. London life does not suit him at all."

**Question**: Who does city life not favor?

**Answer**: Helen's father

**Prediction**: my father

The answer and prediction have one common word, "father", but different adjectives. In this case, the most important information is consisted of the adjective, not in the noun. There is a reported speech in the text, which is why the algorithm model made a mistake. Now let us have a look at the 6 False Negatives samples (Tab.4.19).

| **row** | **Question** | **Answer (Original)** | **Prediction (Original)** | **COS SIM** | **Answer after post processing** | **Prediction after post processing** |
|---|---|---|---|---|---|---|
| 52 | which chapter are we reading? | chapter six | chapter vi | 0.5 | chapter six | chapter vi |
| 143 | Where was the banquet being held? | the old chateau | chateau d aumont | 0.5 | old chateau | chateau aumont |
| 155 | Who had malaria? | chelsea s star striker | didier drogba | 0 | chelsea star striker | didier drogba |
| 210 | what is it commonly known as | the welsh assembly | the national assembly for wales | 0.4 | welsh assembly | national assembly wale |
| 451 | What was Christopher's day job? | teaching | teacher | 0 | teaching | teacher |
| 465 | What was posted online? | an audio message purportedly recorded by osama... | arab spring audio message | 0.38 | audio message purportedly recorded | arab spring audio message |

**Table 4.19:** 6 False Negative samples

The prediction in row 52 has to be assigned as correct, but the roman number has not been converted. To fix this problem, we need to add a new rule in the Number to Word processing step: convert all the roman numbers to the Arabian numbers. In row 143, there is one common word, "chateau", which is significant, but due to different adjectives, the machine labeled this prediction wrong. The same problem is in Row 465 - many different and less significant adjectives. In row

155, both answer and prediction are right because Didier Drogba is Chelsea's star striker. Rows 210 and 451 is a clear example of missing the Stemming approach. The stemming approach converts a word to his base root with a truncation. In this thesis, Stemming was not implemented because it can as help as lead to mistakes. Indeed "Welsh" and "Wales" have the same root, "teaching" and "teacher" as well. Nevertheless, we can consider one True Negative example, when the absence of Stemming helps to assign the prediction as false 4.20. The question started with "Where". It is wrong to answer "canadian" because the question should start with "Who".

| **row** | **Question** | **Answer (Original)** | **Prediction (Original)** | **COS SIM** | **Answer after post processing** | **Prediction after post processing** |
|---|---|---|---|---|---|---|
| 187 | Where is Rick Hansen from? | canada | canadian | 0 | canada | canadian |

**Table 4.20:** Example of True Negative with Stemming absence

After this analysis the new parameters were calculated:

- Accuracy = 0.836
- Precision = 0.999
- Recall = 0.984

# Chapter 5

# Conclusion

In this chapter, we will summarize what was learned during this thesis work, which methods were investigated, which results were achieved, what was missing, and what can be done in the future for improvement.

### 5.0.1 Learned topics and techniques

During this thesis, four main topics were thoroughly researched:

1. 3 datasets (SQuAD, SQuAD v.2 and COQA)
2. BERT model implementation
3. Data mining techniques
4. Post processing methods

In sections 3.1 - 3.3 , each of the datasets were analyzed in detail. Their characteristics were discussed, the samples were shown, question and answers types were compared, and the differences between the datasets were found and highlighted. After that BERT model was described, the structure of the inputs was explained in 3.4. All the methods (section 3.5) and unnecessary definitions which were used during post-processing have been introduced and well explained :

- Tokenization
- Stop Words Removing
- Lemmatization
- Cosine Similarity
- TF-IDF method

- Text Similarity

Then experimental research was done. Already pretrained BERT model and finetuned on SQuAD was implemented on COQA dataset. Each post-processing step was described, and for each of them, the results were shown in a table format. Cosine Similarity and Accuracy scores were calculated as well. The list of post-processing steps:

- Lower case
- Punctuation and Space Elimination
- The Diacritics Elimination
- "Yes" "No" elimination
- Number to word conversion
- Lemmatization
- Stop Words Removal
- Answer presence in predictions
- Prediction presence in answers

### 5.0.2 Results

After defining the pipeline, the model was implemented on the validation dataset (section 4.5). Great results were obtained:

- The Average Cosine Similarity = 0,77
- The Accuracy = 0.836

As the validation data set size was less than 400, it was possible to evaluate every answer-prediction pair and calculate recall and precise.

- Recall = 0.984
- Precision = 0.999

The code was written in Python.

### 5.0.3 Future plans

The following research step could be implementing the same model on the same dataset using the first Question-Answer pair for each text passage and the whole number of turns for each passage. After that, it would be necessary to finetune the given BERT model on the COQA dataset and define new post-processing steps for evaluation. Finally, the final model could be defined and implemented for the student and professor's needs after comparing the results.

# Bibliography


[1] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. «Deep Learning Based Text Classification: A Comprehensive Review». In: vol. abs/2004.03705. 2020. arXiv: 2004.03705. URL: https://arxiv.org/abs/2004.03705 (cit. on p. 11).

[2] Kun Jing and Jungang Xu. «A Survey on Neural Network Language Models». In: vol. abs/1906.03591. 2019. arXiv: 1906.03591. URL: http://arxiv.org/abs/1906.03591 (cit. on p. 11).

[3] Felix Stahlberg. «Neural Machine Translation: A Review». In: vol. abs/1912.02047. 2019. arXiv: 1912.02047. URL: http://arxiv.org/abs/1912.02047 (cit. on p. 11).

[4] Chanwoo Kim, Dhananjaya Gowda, Dongsoo Lee, Jiyeon Kim, Ankur Kumar, Sungsoo Kim, Abhinav Garg, and Changwoo Han. «A review of on-device fully neural end-to-end automatic speech recognition algorithms». In: vol. abs/2012.07974. 2020. arXiv: 2012.07974. URL: https://arxiv.org/abs/2012.07974 (cit. on p. 11).

[5] Ahsaas Bajaj et al. «Long Document Summarization in a Low Resource Setting using Pretrained Language Models». In: vol. abs/2103.00751. 2021. arXiv: 2103.00751. URL: https://arxiv.org/abs/2103.00751 (cit. on p. 11).

[6] Mansi Gupta, Nitish Kulkarni, Raghuveer Chanda, Anirudha Rayasam, and Zachary C. Lipton. «AmazonQA: A Review-Based Question Answering Task». In: vol. abs/1908.04364. 2019. arXiv: 1908.04364. URL: http://arxiv.org/abs/1908.04364 (cit. on p. 11).

[7] Mark Yatskar. «A Qualitative Comparison of CoQA, SQuAD 2.0 and QuAC». In: vol. abs/1809.10735. 2018. arXiv: 1809.10735. URL: http://arxiv.org/abs/1809.10735 (cit. on p. 12).

[8] Ying Ju, Fubang Zhao, Shijie Chen, Bowen Zheng, Xuefeng Yang, and Yunfeng Liu. «Technical report on Conversational Question Answering». In: vol. abs/1909.10772. 2019. arXiv: 1909.10772. URL: http://arxiv.org/abs/1909.10772 (cit. on p. 12).

[9] Zhuosheng Zhang, Junjie Yang, and Hai Zhao. «Retrospective Reader for Machine Reading Comprehension». In: vol. abs/2001.09694. 2020. arXiv: 2001.09694. URL: https://arxiv.org/abs/2001.09694 (cit. on p. 12).

[10] Jianfeng Gao, Michel Galley, and Lihong Li. «Neural Approaches to Conversational AI». In: vol. abs/1809.08267. 2018. arXiv: 1809.08267. URL: http://arxiv.org/abs/1809.08267 (cit. on p. 13).

[11] Tomás Kociský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. «The NarrativeQA Reading Comprehension Challenge». In: vol. abs/1712.07040. 2017. arXiv: 1712.07040. URL: http://arxiv.org/abs/1712.07040 (cit. on p. 13).

[12] Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. «TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension». In: vol. abs/1705.03551. 2017. arXiv: 1705.03551. URL: http://arxiv.org/abs/1705.03551 (cit. on p. 13).

[13] Simon Suster and Walter Daelemans. «CliCR: A Dataset of Clinical Case Reports for Machine Reading Comprehension». In: vol. abs/1803.09720. 2018. arXiv: 1803.09720. URL: http://arxiv.org/abs/1803.09720 (cit. on p. 13).

[14] Debanjali Biswas, Mohnish Dubey, Md. Rashad Al Hasan Rony, and Jens Lehmann. «VANiLLa : Verbalized Answers in Natural Language at Large Scale». In: vol. abs/2105.11407. 2021. arXiv: 2105.11407. URL: https://arxiv.org/abs/2105.11407 (cit. on p. 13).

[15] Munazza Zaib, Wei Emma Zhang, Quan Z. Sheng, Adnan Mahmood, and Yang Zhang. «Conversational Question Answering: A Survey». In: vol. abs/2106.00874, 2021. arXiv: 2106.00874. URL: https://arxiv.org/abs/2106.00874 (cit. on p. 13).

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding». In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423 (cit. on pp. 15, 31).

[17] Yinhan Liu et al. «RoBERTa: A Robustly Optimized BERT Pretraining Approach». In: vol. abs/1907.11692. 2019. arXiv: 1907.11692. URL: http://arxiv.org/abs/1907.11692 (cit. on p. 15).

[18] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. «DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter». In: vol. abs/1910.01108. 2019. arXiv: 1910.01108. URL: http://arxiv.org/abs/1910.01108 (cit. on p. 15).

[19] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. «XLNet: Generalized Autoregressive Pretraining for Language Understanding». In: vol. abs/1906.08237. 2019. arXiv: 1906.08237. URL: http://arxiv.org/abs/1906.08237 (cit. on p. 15).

[20] Tom B. Brown et al. «Language Models are Few-Shot Learners». In: vol. abs/2005.14165¦ 2020. arXiv: 2005.14165. URL: https://arxiv.org/abs/2005.14165 (cit. on p. 15).

[21] Mark Yatskar. «A Qualitative Comparison of CoQA, SQuAD 2.0 and QuAC». In: vol. abs/1809.10735. 2018. arXiv: 1809.10735. URL: http://arxiv.org/abs/1809.10735 (cit. on p. 16).

[22] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. «SQuAD¦ 100, 000+ Questions for Machine Comprehension of Text». In: vol. abs/1606.05250¦ 2016. arXiv: 1606.05250. URL: http://arxiv.org/abs/1606.05250 (cit. on p. 17).

[23] Pranav Rajpurkar, Robin Jia, and Percy Liang. «Know What You Don't Know: Unanswerable Questions for SQuAD». In: vol. abs/1806.03822. 2018. arXiv: 1806.03822. URL: http://arxiv.org/abs/1806.03822 (cit. on p. 20).

[24] Haichao Zhu, Li Dong, Furu Wei, Wenhui Wang, Bing Qin, and Ting Liu. «Learning to Ask Unanswerable Questions for Machine Reading Comprehension». In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 4238–4248. DOI: 10.18653/v1/P19-1415. URL: https://aclanthology.org/P19-1415 (cit. on p. 20).

[25] Siva Reddy, Danqi Chen, and Christopher D. Manning. «CoQA: A Conversational Question Answering Challenge». In: vol. abs/1808.07042. 2018. arXiv: 1808.07042. URL: http://arxiv.org/abs/1808.07042 (cit. on p. 23).

[26] Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. «MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text». In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, Oct. 2013, pp. 193–203. URL: https://aclanthology.org/D13-1020 (cit. on p. 25).

[27] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. «Teaching Machines to Read and Comprehend». In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett. Vol. 28. Curran Associates, Inc., 2015. URL: https://proceedings.neurips.cc/paper/2015/file/afdec7005cc9f14302cd0474fd0f3c96-Paper.pdf (cit. on p. 25).

[28] Angela Fan, Mike Lewis, and Yann Dauphin. «Hierarchical Neural Story Generation». In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 889–898. DOI: 10.18653/v1/P18-1082. URL: https://aclanthology.org/P18-1082 (cit. on p. 25).

[29] Johannes Welbl, Nelson F. Liu, and Matt Gardner. «Crowdsourcing Multiple Choice Science Questions». In: vol. abs/1707.06209. 2017. arXiv: 1707.06209. URL: http://arxiv.org/abs/1707.06209 (cit. on p. 25).

[30] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. «RACE: Large-scale ReAding Comprehension Dataset From Examinations». In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 785–794. DOI: 10.18653/v1/D17-1082. URL: https://aclanthology.org/D17-1082 (cit. on p. 25).

[31] Thomas Wolf et al. «HuggingFace's Transformers: State-of-the-art Natural Language Processing». In: vol. abs/1910.03771. 2019. arXiv: 1910.03771. URL: http://arxiv.org/abs/1910.03771 (cit. on p. 30).

[32] Benyamin Ghojogh and Ali Ghodsi. «Attention Mechanism, Transformers, BERT, and GPT: Tutorial and Survey». In: Dec. 2020. DOI: 10.31219/osf.io/m6gcn (cit. on p. 31).

[33] Kyubyong Park, Joohong Lee, Seongbo Jang, and Dawoon Jung. «An Empirical Study of Tokenization Strategies for Various Korean NLP Tasks». In: vol. abs/2010.02534. 2020. arXiv: 2010.02534. URL: https://arxiv.org/abs/2010.02534 (cit. on p. 34).

[34] Tomás Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. «Distributed Representations of Words and Phrases and their Compositionality». In: vol. abs/1310.4546. 2013. arXiv: 1310.4546. URL: http://arxiv.org/abs/1310.4546 (cit. on p. 34).

[35] Jiawei Han, Micheline Kamber, and Jian Pei. «Data mining concepts and techniques, third edition». In: Morgan Kaufmann Publishers, 2012. ISBN: 0123814790 (cit. on p. 36).

[36] Amir Jalilifard, Vinicius Fernandes Caridá, Alex Mansano, and Rogers Cristo. «Semantic Sensitive TF-IDF to Determine Word Relevance in Documents». In: vol. abs/2001.09896. 2020. arXiv: 2001.09896. URL: https://arxiv.org/abs/2001.09896 (cit. on p. 37).

[37] Wael Gomaa and Aly Fahmy. «A Survey of Text Similarity Approaches». In: vol. 68. Apr. 2013. DOI: 10.5120/11638-7118 (cit. on p. 38).

# Appendix A

# Python libraries

There is a list of used Python libraries in this thesis work:

- Pandas
- Numpy
- Matplotlib
- NTLK
- Torch
- Transformers

# Appendix B

# Validation dataframe results

| | text | question | answer | prediction | Label |
|---|---|---|---|---|---|
| 0 | Once upon a time, in a barn near a farm house,... | What color was Cotton? | white | white | 1 |
| 1 | Once there was a beautiful fish named Asta. As... | what was the name of the fish | Asta. | asta | 1 |
| 2 | Kendra and Quinton travel to and from school e... | Where do Quinton and Kendra travel to and from... | school | school | 1 |
| 3 | Thunder was coming when Reginald Eppes woke up... | When did Reginald Eppes wake up? | Five in the morning | five in the morning | 1 |
| 4 | (CNN) -- FBI agents on Friday night searched t... | Whose house was searched? | Gary Giordano | gary giordano | 1 |
| 5 | Kabul, Afghanistan (CNN) -- The German news ou... | What news agency showed photos of American sol... | Der Spiegel | der spiegel | 1 |
| 6 | (CNN)A chiseled boxer's Instagram feed shows h... | Who are the two boxer featured in this article? | Floyd Mayweather and Manny Pacquiao | floyd mayweather and filipino manny pacquiao | 1 |
| 7 | Tommy was a little boy who lived by a big lake... | Where'd Tommy live? | by a big lake by the woods | by a big lake by the woods | 1 |
| 8 | CHAPTER XXII \n\nNorthward, along the leeward ... | What worked her way northward? | The _Ariel_ | the _ ariel _ | 1 |
| 9 | Brownie and Spotty were neighbor dogs who met ... | Who were the two canines who lived next door t... | Brownie and Spotty | brownie and spotty | 1 |
| 10 | (CNN) -- For Heather Neroy, it used to be a te... | What is the main character interested in? | arts-and-crafts | arts - and - crafts | 1 |
| 11 | Chapter XVIII \n\nThe Hound Restored \n\nOn th... | To whom did Archie pray? | Sir Earl | sir earl | 1 |
| 12 | Asuncion, Paraguay (CNN) -- Paraguay installed... | who made the announcement, the office of the p... | the armed forces | the armed forces | 1 |
| 13 | A women went shopping for a dress to wear to h... | what did the woman go shopping for ? | a dress | a dress | 1 |
| 14 | Chapter 17: The Battle Of Moncontor. \n\nWhen ... | Who was eating? | Philip | philip | 1 |
| 15 | Wiltshire ( or ) is a county in South West Eng... | What is Wiltshire characterised by? | its high downland and wide valleys | high downland and wide valleys | 1 |
| 16 | Brendan loves cats. He owns 8 cats. He has 7 g... | What does he care for? | cats | cats | 1 |
| 17 | Ryan and Adam love to play basketball. They li... | What sport do Ryan and Adam love to play? | basketball | basketball | 1 |
| 18 | CHAPTER XII \n\n"Throw your coat down anywhere... | Who did Wingate talk to? | Miss Baldwin | miss baldwin | 1 |
| 19 | A thorough understanding of adolescence in soc... | What connect childhood and adulthood? | adolescence | adolescence | 1 |
| 20 | CHAPTER XX. \n\nMoving against Captain Grady \... | Who is the colored man? | Jeff Jones | jeff jones | 1 |
| 21 | The Federal City of Bonn () is a city on the b... | What has a population of over 300,000 | The Federal City of Bonn | the federal city of bonn | 1 |
| 22 | CHAPTER II. ON A MOUNTAIN PATH \n\n"Armed men,... | who rose to his feet? | Aquila | aquila | 1 |
| 23 | Multimedia is content that uses a combination ... | What can be recorded and played? | Multimedia | multimedia | 1 |
| 24 | A team of British surgeons has carried out Gaz... | What is Adbelkader Hammad's job? | doctor | doctor | 1 |
| 25 | Discogs, short for discographies, is a website... | Who owns the websites servers? | Zink Media, Inc | zink media , inc . | 1 |

**Figure B.1:** First 25 True Positives samples