National College of Ireland

Configuration Manual

MSc Research Project
Data Analytics

Rutuja Bhujbal
Student ID: x22123822

School of Computing
National College of Ireland

Supervisor: Prof. Teerath Kumar Menghwar

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Rutuja bhujbal
Student ID:	22123822
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Prof. Teerath Kumar Menghwar
Submission Due Date:	14th December 2023
Project Title:	Configuration Manual
Word Count:	751
Page Count:	11

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Rutuja Bhujbal
Date:	14th December 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	✓
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	✓
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	✓

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Rutuja Bhujbal
x22123822

1 Introduction

The code used to implement the project "Building a question-answering system to extract information from PDF files using BERT transformers" is described in detail in this file, along with the hardware and software requirements.

2 System configuration

Your second section. Change the header and label to something appropriate.

2.1 Hardware

- Processor: AMD Ryzen 5 5500U with Radeon Graphics 2.10 GHz
- 16.0 GB (15.3 GB usable) ,12.7 GB System RAM on Google Colab
- System type: 64-bit operating system, x64-based processor
- Hard Disk Storage: 100GB (Google Drive Storage)

2.2 Software

- Google Colab), Overleaf, Microsoft Excel, Google Scholar
- Browser Engine: Google Chrome/ Microsoft edge.
- Email: Gmail login to access Google Colab .

3 Project Development

3.1 Project management tool

- Google Drive with 100 GB of storage was used for creating a project environment. Domain-specific PDFs used for analysis were stored in separate folders.
- All notebooks were mounted on Google Drive and saved in the Colab Notebooks folder.

3.2 Installed libraries

Libraries required for implementing question-answering systems were installed at the beginning of each notebook. There are a total of six Colab notebooks. Some of the libraries installed were: transformers, Scikit Learn, PyPDF2, genism, pdf Plumber Devlin et al. (2018) Pearce et al. (2021)

```
!pip install pdfplumber          # Python library for extracting information from PDF
!pip install PyPDF2              #library for reading and manipulating PDF files in Python.
!pip install nltk                #library provides tools for tasks such as tokenization, stemming, lemmatization
!pip install -U gensim           # library for topic modeling and document similarity analysis
!pip install accelerate==0.20.3  #library for GPU acceleration in Python
!pip install transformers -U     #library for working with pre-trained models in transformer-based models

import pdfplumber
import re
import gensim
from gensim.parsing.preprocessing import remove_stopwords
from transformers import pipeline
import torch
from transformers import BertTokenizer
from transformers import BertForQuestionAnswering
from sklearn.metrics import f1_score
from sklearn.metrics import f1_score
import re
import pandas as pd
import nltk
```

Figure 1: Libraries installed in Financial Notebook

4 Design Flow

Perform design guidelines mentioned below such as 1) Data understanding 2) data pre-processing 3) Logic implementation for building question-answering system 4) Evaluation of system on Financial, Biomedical, and scientific PDF Dataset Alsentzer et al. (2019) Beltagy et al. (2019)

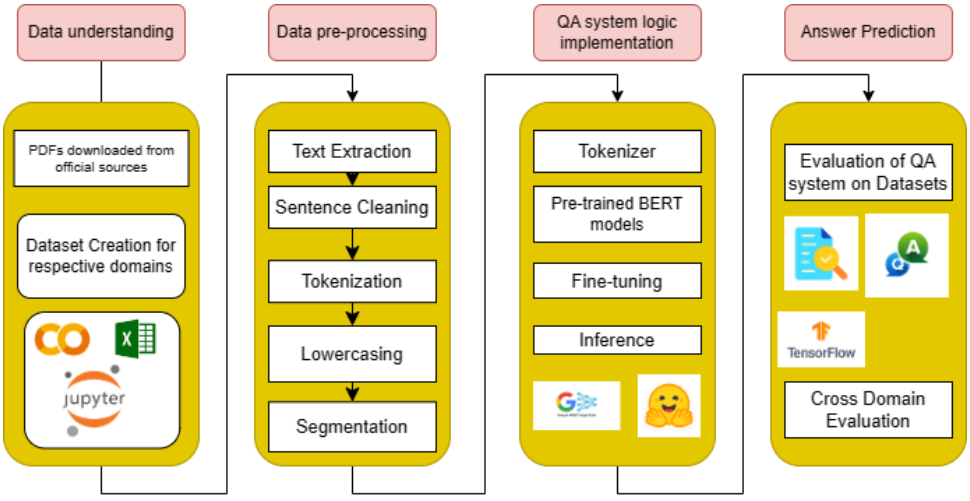


Figure 2: Design flow of building QA system on PDFs

5 Data Collection

PDFs for Financial and Biomedical domains were downloaded from the following sources respectively.

- Amazon's annual reports: Amazon.com Announces Third Quarter Results

- covid Research paper- Biomedical Research COVID-19 Impact Assessment
- For the scientific literature domain research papers used for the literature review of this research were analyzed. Papers were downloaded from Google Scholar.Google Scholar Search

6 Data Pre-processing

This step involves PDF text extraction, text cleaning, tokenization, and segmentation.

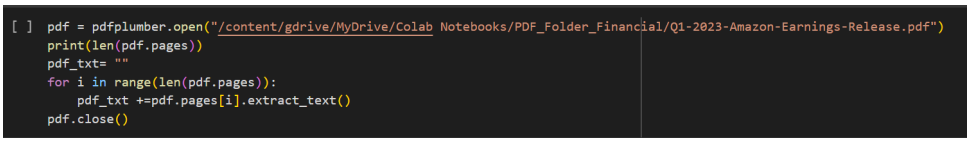


Figure 3: PDF text Extraction

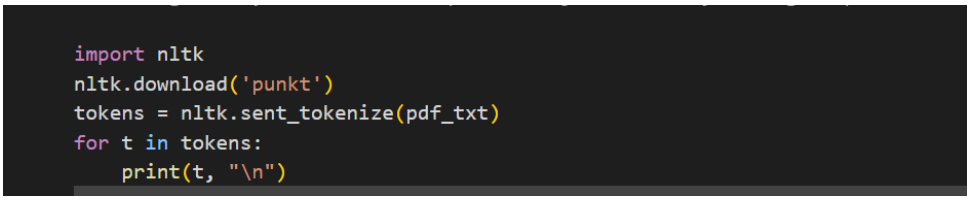


Figure 4: Sentence Tokenization

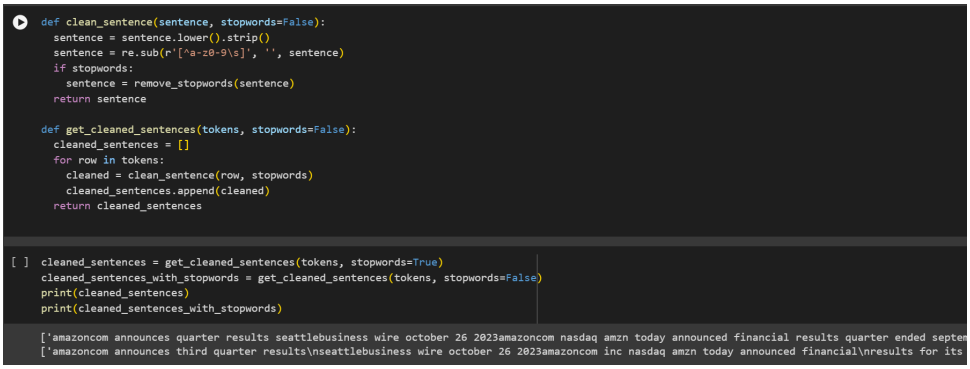


Figure 5: Sentence cleaning

7 Model Implementation Logic

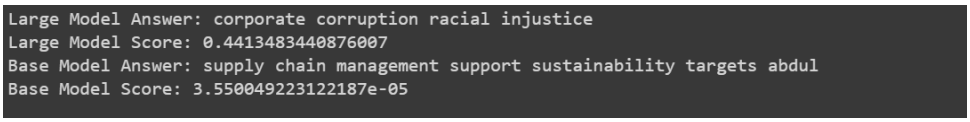
The stepwise model implementation is given below:

7.1 with pipeline library

Implementation of a QA system with a Pipeline library using a pre-trained BERT base and BERT large

```
# Experiment with different models
qa_pipeline_large = pipeline("question-answering", model="bert-large-uncased-whole-word-masking-finetuned-squad")
qa_pipeline_base = pipeline("question-answering", model="bert-base-uncased")

# Assign cleaned sentences to context
context = " ".join(cleaned_sentences)
question = "What is the topic of the document?"

# Get answers from different models
answer_large = qa_pipeline_large(question=question, context=context)
answer_base = qa_pipeline_base(question=question, context=context)

# Print answers and scores
print("Large Model Answer:", answer_large["answer"])
print("Large Model Score:", answer_large["score"])

print("Base Model Answer:", answer_base["answer"])
print("Base Model Score:", answer_base["score"])
```

Figure 6: Implementation of a QA system with a Pipeline



```
Large Model Answer: corporate corruption racial injustice
Large Model Score: 0.4413483440876007
Base Model Answer: supply chain management support sustainability targets abdul
Base Model Score: 3.550049223122187e-05
```

Figure 7: Outputs produced with pipeline library

7.2 with pre-trained BERT function

Implementation of a QA system with a Pipeline library using a pre-trained BERT base and BERT large

```
def answer_question(question, answer_text):
    input_ids = tokenizer.encode(question, answer_text, max_length=512, truncation=True)
    print('Query has {:,} tokens.\n'.format(len(input_ids)))
    sep_index = input_ids.index(tokenizer.sep_token_id)
    num_seg_a = sep_index + 1
    num_seg_b = len(input_ids) - num_seg_a
    segment_ids = [0]*num_seg_a + [1]*num_seg_b
    assert len(segment_ids) == len(input_ids)
    outputs = model(torch.tensor([input_ids]), token_type_ids=torch.tensor([segment_ids]))
    start_index = torch.argmax(outputs.start_logits)
    end_index = torch.argmax(outputs.end_logits)
    all_tokens = tokenizer.convert_ids_to_tokens(input_ids)
    score = float(torch.max(start_index))
    answer_start = torch.argmax(start_index)
    answer_end = torch.argmax(end_index)
    tokens = tokenizer.convert_ids_to_tokens(input_ids)
    answer = tokens[start_index]
    for i in range(start_index + 1, end_index + 1):

        if tokens[i][0:2] == ' ':
            answer += tokens[i][2:]
        else:
            answer += ' ' + tokens[i]
    return answer, score
```

Figure 8: Implementation of a QA system with a BERT large

7.3 With curated Financial dataset

The modified answer_question function is called with the curated dataset to evaluate the F1 score. Refer manual dataset creation step to generate the Financial dataset question-answer pairs.

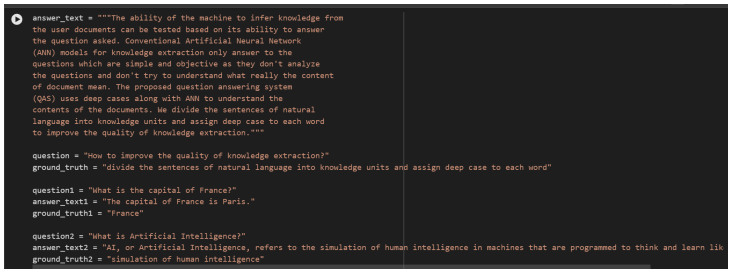


Figure 9: set of of question-answer pairs

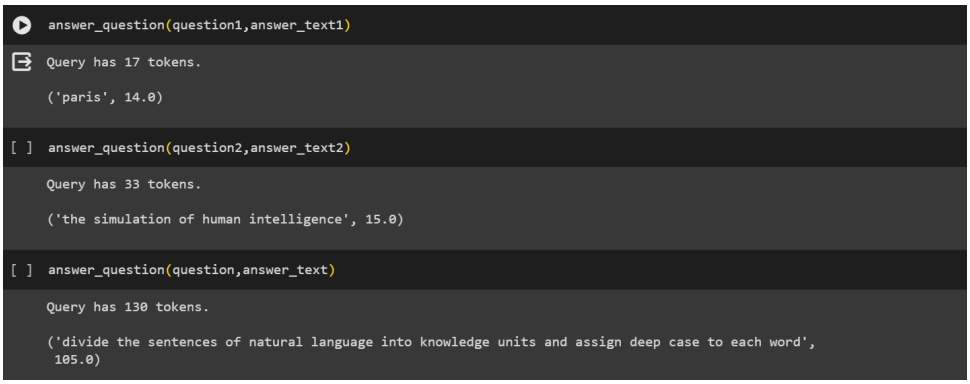


Figure 10: Outputs for set of question-answer pairs

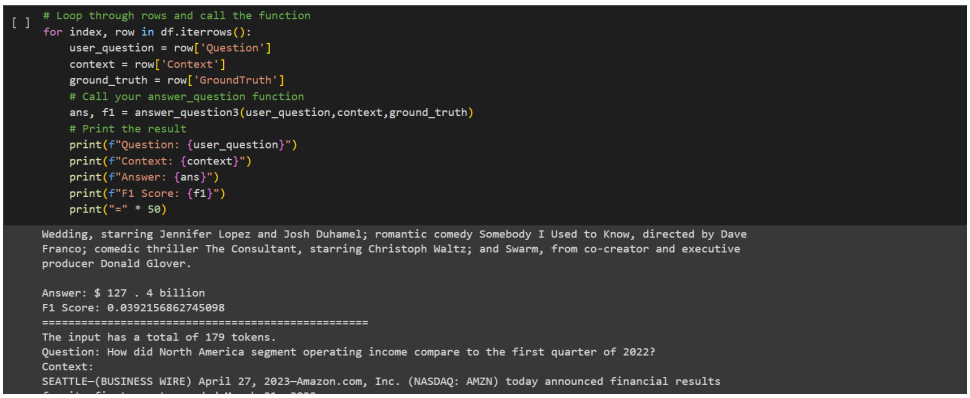
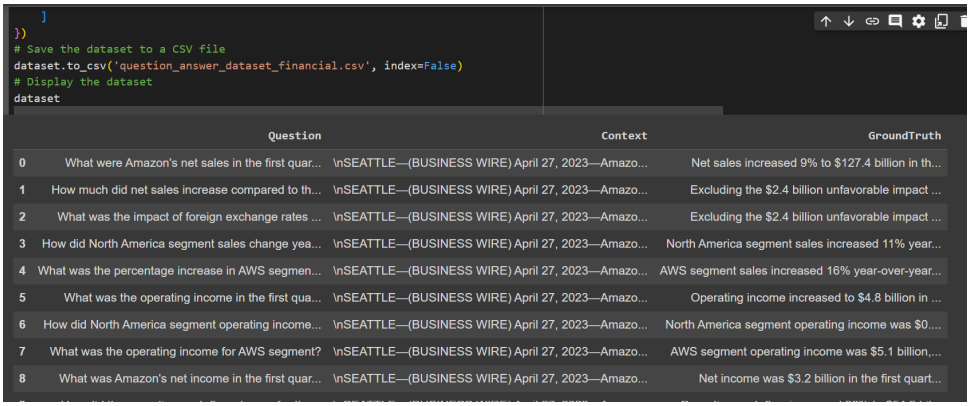


Figure 11: Implementation of QA with curated datset

8 Curated dataset creation

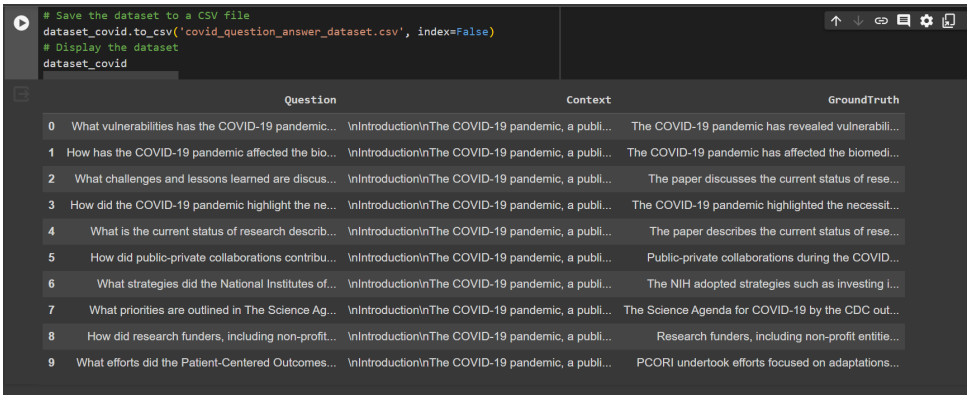
Datasets are created manually for each domain by running the dataset creation notebook.



```
    ]
})
# Save the dataset to a CSV file
dataset.to_csv('question_answer_dataset_financial.csv', index=False)
# Display the dataset
dataset
```

	Question	Context	GroundTruth
0	What were Amazon's net sales in the first quar...	\nSEATTLE—(BUSINESS WIRE) April 27, 2023—Amazo...	Net sales increased 9% to $127.4 billion in th...
1	How much did net sales increase compared to th...	\nSEATTLE—(BUSINESS WIRE) April 27, 2023—Amazo...	Excluding the $2.4 billion unfavorable impact ...
2	What was the impact of foreign exchange rates ...	\nSEATTLE—(BUSINESS WIRE) April 27, 2023—Amazo...	Excluding the $2.4 billion unfavorable impact ...
3	How did North America segment sales change yea...	\nSEATTLE—(BUSINESS WIRE) April 27, 2023—Amazo...	North America segment sales increased 11% year...
4	What was the percentage increase in AWS segmen...	\nSEATTLE—(BUSINESS WIRE) April 27, 2023—Amazo...	AWS segment sales increased 16% year-over-year...
5	What was the operating income in the first qua...	\nSEATTLE—(BUSINESS WIRE) April 27, 2023—Amazo...	Operating income increased to $4.8 billion in ...
6	How did North America segment operating income...	\nSEATTLE—(BUSINESS WIRE) April 27, 2023—Amazo...	North America segment operating income was $0....
7	What was the operating income for AWS segment?	\nSEATTLE—(BUSINESS WIRE) April 27, 2023—Amazo...	AWS segment operating income was $5.1 billion,...
8	What was Amazon's net income in the first quar...	\nSEATTLE—(BUSINESS WIRE) April 27, 2023—Amazo...	Net income was $3.2 billion in the first quart...

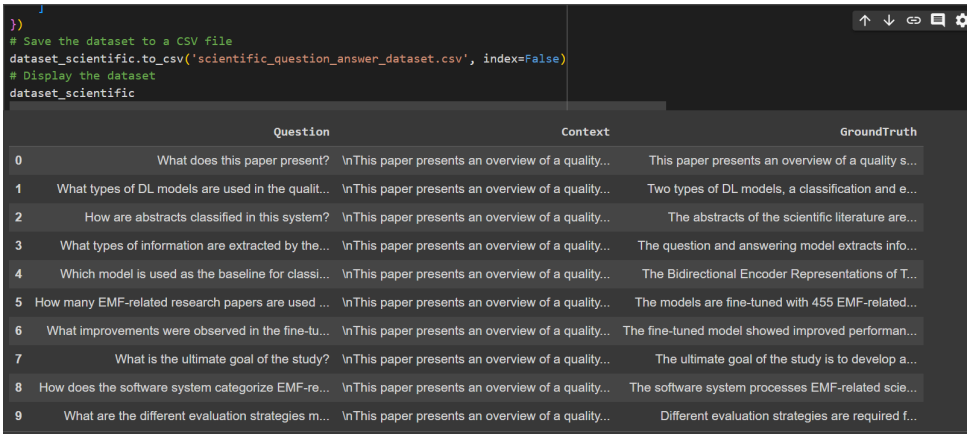
Figure 12: Financial dataset creation



```
# Save the dataset to a CSV file
dataset_covid.to_csv('covid_question_answer_dataset.csv', index=False)
# Display the dataset
dataset_covid
```

	Question	Context	GroundTruth
0	What vulnerabilities has the COVID-19 pandemic...	\nIntroduction\nThe COVID-19 pandemic, a publi...	The COVID-19 pandemic has revealed vulnerabili...
1	How has the COVID-19 pandemic affected the bio...	\nIntroduction\nThe COVID-19 pandemic, a publi...	The COVID-19 pandemic has affected the biomedi...
2	What challenges and lessons learned are discus...	\nIntroduction\nThe COVID-19 pandemic, a publi...	The paper discusses the current status of rese...
3	How did the COVID-19 pandemic highlight the ne...	\nIntroduction\nThe COVID-19 pandemic, a publi...	The COVID-19 pandemic highlighted the necessit...
4	What is the current status of research describ...	\nIntroduction\nThe COVID-19 pandemic, a publi...	The paper describes the current status of rese...
5	How did public-private collaborations contribu...	\nIntroduction\nThe COVID-19 pandemic, a publi...	Public-private collaborations during the COVID...
6	What strategies did the National Institutes of...	\nIntroduction\nThe COVID-19 pandemic, a publi...	The NIH adopted strategies such as investing i...
7	What priorities are outlined in The Science Ag...	\nIntroduction\nThe COVID-19 pandemic, a publi...	The Science Agenda for COVID-19 by the CDC out...
8	How did research funders, including non-profit...	\nIntroduction\nThe COVID-19 pandemic, a publi...	Research funders, including non-profit entitie...
9	What efforts did the Patient-Centered Outcomes...	\nIntroduction\nThe COVID-19 pandemic, a publi...	PCORI undertook efforts focused on adaptations...

Figure 13: Biological dataset creation



```
})
# Save the dataset to a CSV file
dataset_scientific.to_csv('scientific_question_answer_dataset.csv', index=False)
# Display the dataset
dataset_scientific
```

	Question	Context	GroundTruth
0	What does this paper present?	\nThis paper presents an overview of a quality...	This paper presents an overview of a quality s...
1	What types of DL models are used in the qualit...	\nThis paper presents an overview of a quality...	Two types of DL models, a classification and e...
2	How are abstracts classified in this system?	\nThis paper presents an overview of a quality...	The abstracts of the scientific literature are...
3	What types of information are extracted by the...	\nThis paper presents an overview of a quality...	The question and answering model extracts info...
4	Which model is used as the baseline for classi...	\nThis paper presents an overview of a quality...	The Bidirectional Encoder Representations of T...
5	How many EMF-related research papers are used ...	\nThis paper presents an overview of a quality...	The models are fine-tuned with 455 EMF-related...
6	What improvements were observed in the fine-tu...	\nThis paper presents an overview of a quality...	The fine-tuned model showed improved performan...
7	What is the ultimate goal of the study?	\nThis paper presents an overview of a quality...	The ultimate goal of the study is to develop a...
8	How does the software system categorize EMF-re...	\nThis paper presents an overview of a quality...	The software system processes EMF-related scie...
9	What are the different evaluation strategies m...	\nThis paper presents an overview of a quality...	Different evaluation strategies are required f...

Figure 14: Scientific dataset creation

9 Model Fine tuning

This step implements fine-tuning of DistillBERT on the SQAuD dataset where hyper-parameters are set and the model is trained with trainer class. This tuned model is evaluated with a Financial dataset

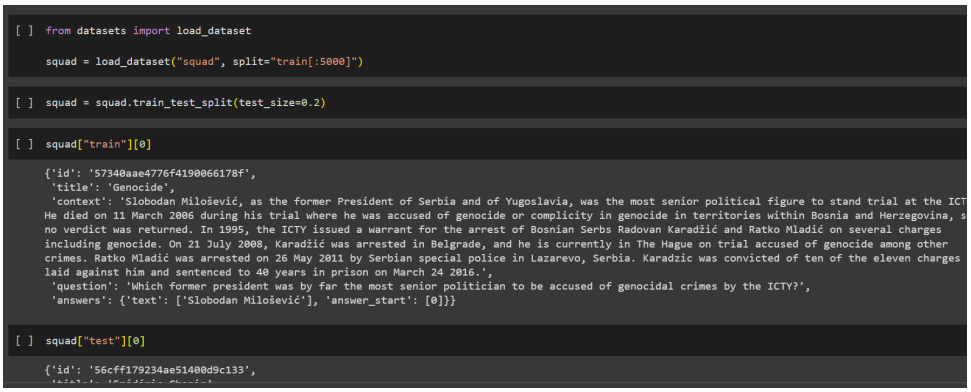


Figure 15: fine tuning with SQAud dataset

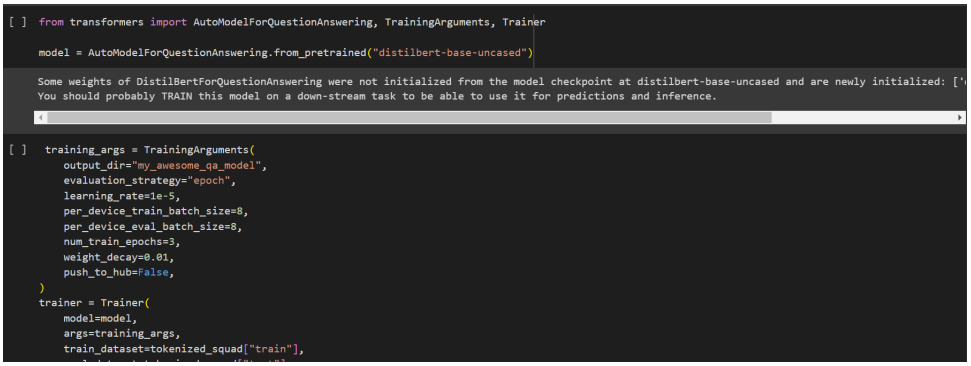


Figure 16: Hyperparameter tuning

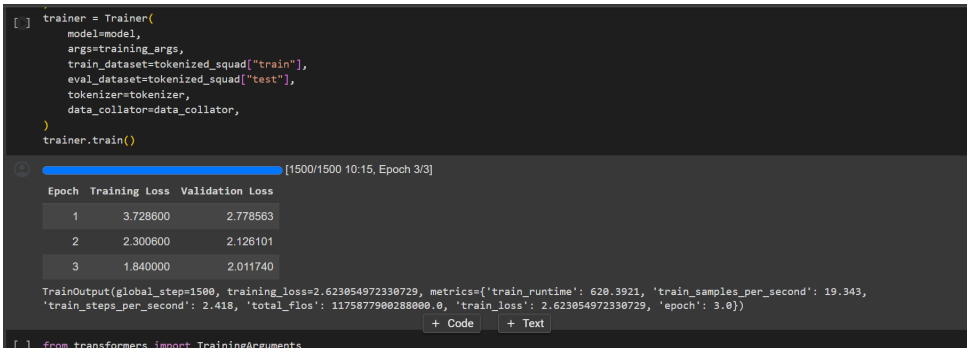


Figure 17: Fine-tuning of DistilBERT calling trainer class

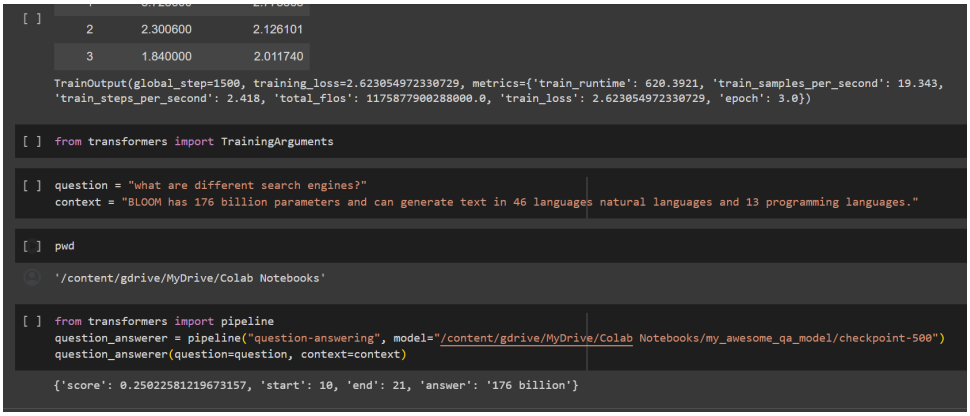


Figure 18: Evaluating financial dataset on fine-tuned model

10 QA system implementation for BioMedical dataset

The BioMedical dataset was implemented using pre-trained bioBERT

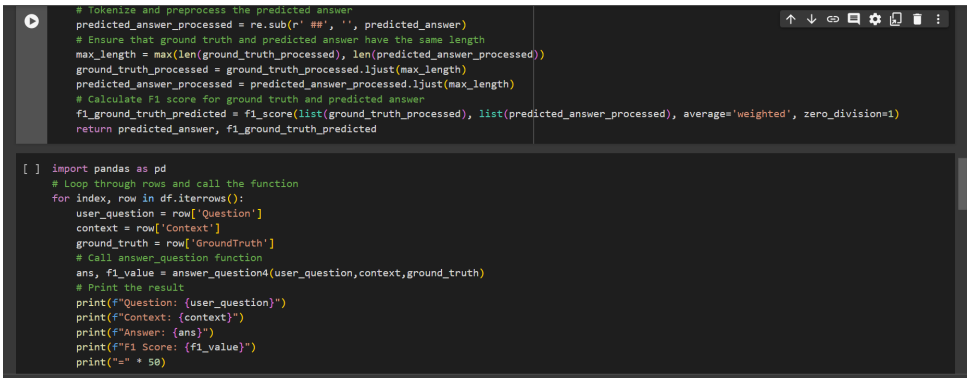


Figure 19: QA system with BioClinical BERT for biomedical dataset

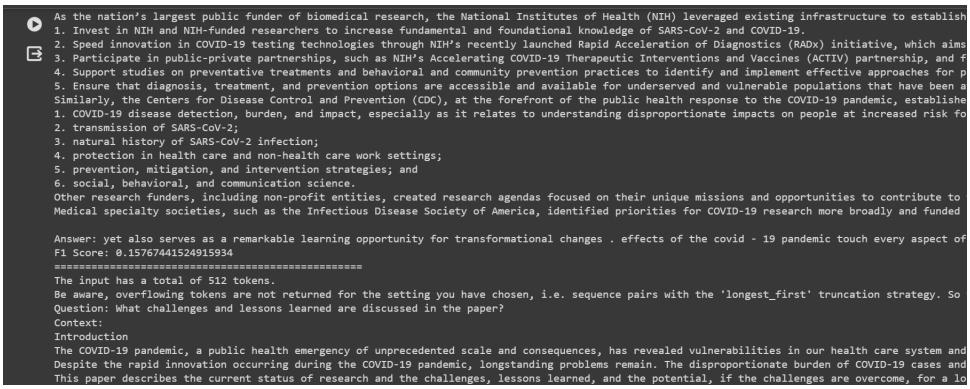


Figure 20: Predicted answers for Bioclinical BERT

11 QA system implementation for Scientific dataset

The Scientific dataset was implemented using pre-trained sciBERT

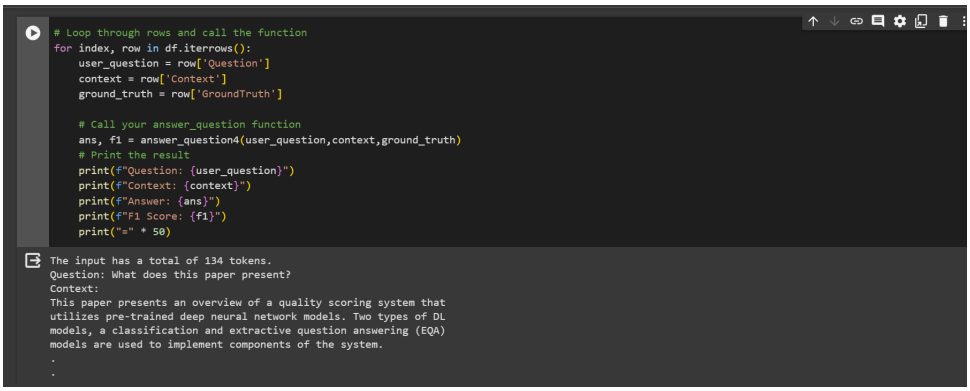


Figure 21: QA system with Scientific BERT for Scientific dataset

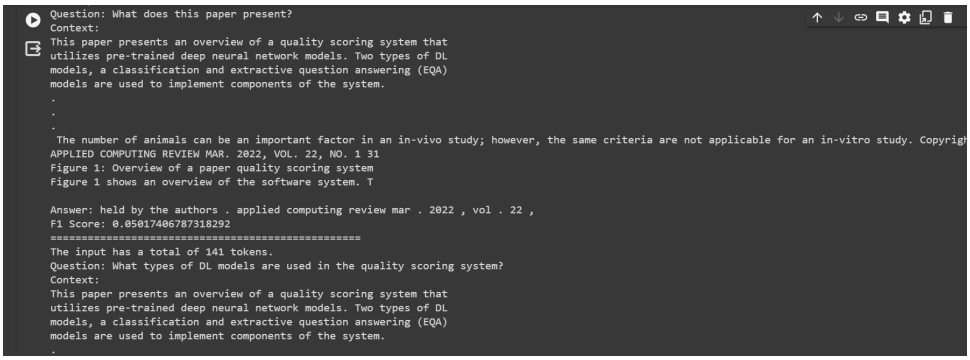


Figure 22: Predicted answers for Scientific BERT

12 Cross domain Evaluation

Biomedical and scientific datasets were evaluated on the BERT large model to check the QA system's generalizability.

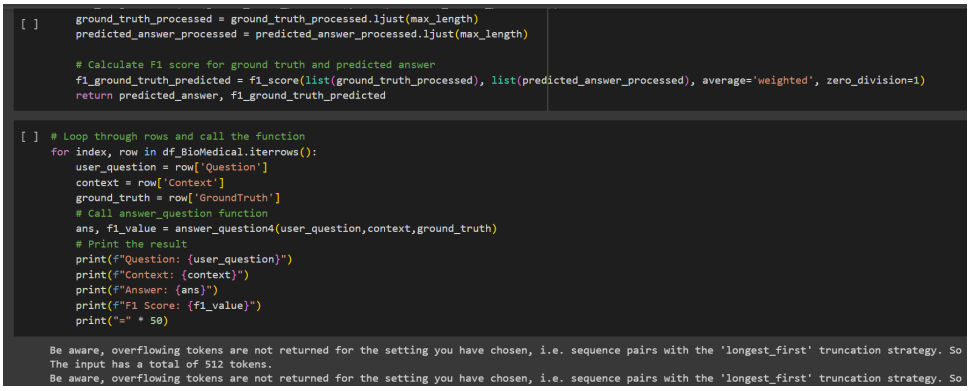


Figure 23: Cross-domain analysis of BiobERT on BERT large

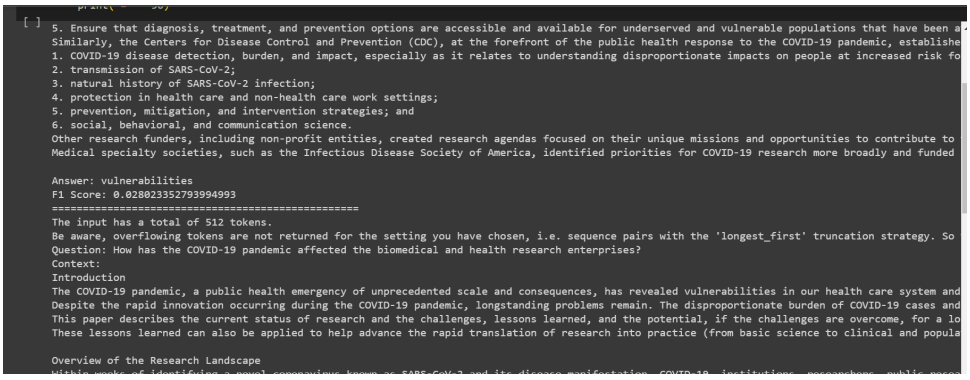


Figure 24: predicted answers for Cross-domain analysis of BiobERT on BERT large

Figure 25: Cross-domain analysis of sciBERT on BERT large

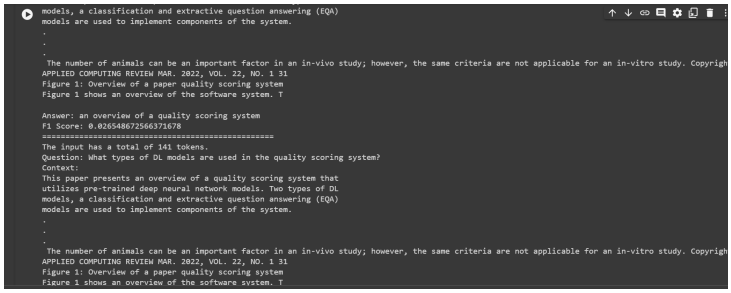


Figure 26: predicted answers for Cross-domain analysis of sciBERT on BERT large

References

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T. and McDermott, M. (2019). Publicly available clinical bert embeddings, *arXiv preprint arXiv:1904.03323* .

Beltagy, I., Lo, K. and Cohan, A. (2019). Scibert: A pretrained language model for scientific text, *arXiv preprint arXiv:1903.10676* .

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* .

Pearce, K., Zhan, T., Komanduri, A. and Zhan, J. (2021). A comparative study of transformer-based language models on extractive question answering, *arXiv preprint arXiv:2110.03142* .