Building a question-answering system to extract information from PDF files using BERT transformers

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 Bhujabl
Student ID:	x22123822
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Prof. Teerath Kumar Menghwar
Submission Due Date:	14/12/2023
Project Title:	Building a question-answering system to extract information from PDF files using BERT transformers
Word Count:	6767
Page Count:	20

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 a 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):	

Building a question-answering system to extract information from PDF files using BERT transformers

Rutuja Bhujbal
x22123822

Abstract

The comprehension of complex PDF such as research documents, clinical reports, and scientific manuals is a time-consuming task. Previous studies have demonstrated significant success in building question-answering systems to provide contextually relevant answers to user queries. However, addressing puzzling questions within a single end-to-end trained ML model remains a rigorous task. Such systems require a huge amount of labeled training data to train the base models for specific tasks. The creation of such datasets is still a challenge for complicated documents like the annual reports of big tech companies. This research paper addresses this challenge by focusing on the construction of a question-answering system tailored for PDF files, specifically targeting domains such as finance, biomedicine, and scientific literature. Curated data sets for the PDF from chosen domains were created manually for the evaluation. Pre-trained Bidirectional Encoder Representations from Transformers (BERT) Models from the Hugging Face Library were utilized for the chosen domains and evaluated with an F1 score. A score of 44% was achieved for the BERT Large.

Keywords: question answering, Bidirectional Encoder Representations from Transformers

1 Introduction

Particularly in question-answering systems, natural language processing (NLP) has advanced in recent years. These algorithms are integral to efficiently extracting relevant information from vast volumes of written content, thereby enabling people to obtain accurate and contextually appropriate responses to their inquiries. Even with end-to-end training, answering complex questions with a single machine learning model remains challenging [1]. The goal of this research paper is to develop a question-answering system specifically for PDF files related to scientific, financial, and biomedical literature. The technical analysis of these PDFs may be exhausting and require in-depth study of their complex text, which makes them hard to understand. On the other hand, a question-answering system can save users time and effort by rapidly retrieving the needed information. Prior research has demonstrated that machine learning models can be used to retrieve information from large documents, and these models can be evaluated using metrics like exact match and F1 score [1], [2]. Bidirectional Encoder Representations from Transformers, or BERT, for short, can perform better in deep learning tasks like quality control and summarization of text with minimal to no modifications after pre-training with just one additional output layer. Medical facilities might gain from accurate

QA systems by researching disease symptoms and treatment options. Pre-trained BERT models, like Bio-BERT and Sci-BERT, seem to perform more accurately than traditional models, in line with prior research. [Unsupervised Biomedical Question-Answering Pre-training]. This study will add to our learning of the effective usage of transformers in the fabrication of QA software for the selected PDFs. All things looked at, a QA system offers a more clever, effective, and user-friendly method of information extraction from huge PDFs, such as annual reports, articles, and medical documents.

Research Question: The problem mentioned in the above section motivates the following research question:

In the financial, biomedical, and scientific domains, how can BERT transformers be employed to effectively answer questions and retrieve knowledge from PDF files? This research intends to implement a QA system using pre-trained BERT models for the chosen domains and evaluate performance to find out the best-performing model. An overview of the question-answering system with domain-specific pre-trained BERT on documents is shown in Figure 1

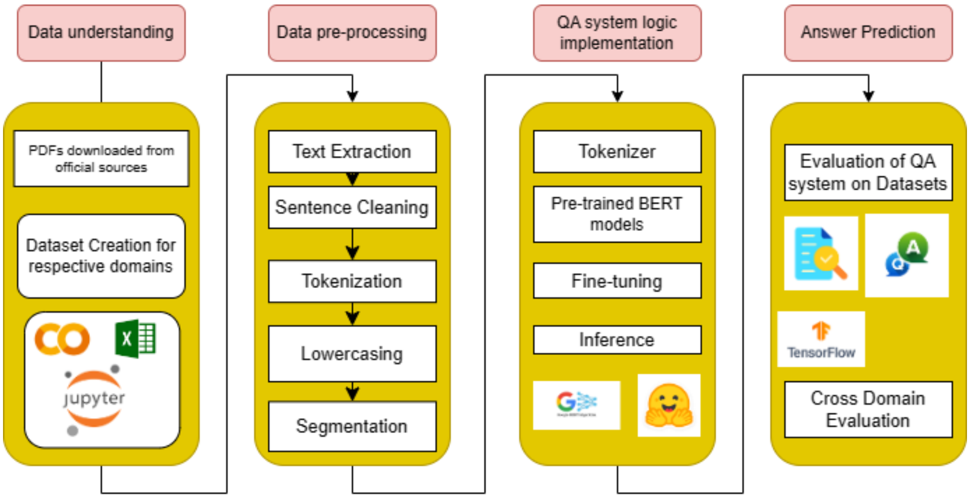

Figure 1: Question answering system with fine-tuned BERT

The paper's remaining sections are arranged as follows: The related work is discussed in Section 2, the research methodology is shown in Section 3, the design specification is explained in Section 4, the implementation aspects of the research are illustrated in Section 5, the evaluation results are examined in Section 6, and the conclusion and future work are highlighted in Section 7.

2 Related Work

BERT has demonstrated Advanced results in various NLP tasks due to its conceptual simplicity and empirical effectiveness. This research aims at building a BERT-based question-answering system specifically tailored for PDF documents, aiming to address the challenges associated with extracting nuanced information from this widely used format.PDF files, prevalent in academic and professional settings, pose challenges for effective information retrieval due to their diverse structures and complex formatting. Leveraging BERT's capabilities offers a promising solution to enhance comprehension and accessibility in PDF-based question answering. This literature review delves into existing research on integrating BERT models for question-answering in PDFs, aiming to identify gaps and opportunities. The goal is to advance intelligent systems for document comprehension by providing insights into the creation of a BERT-based QA system based on PDFs.

Researchers in this study Vaswani et al. (2017)introduced the Transformer architecture, a sequence transduction model based on attention mechanisms. The drawbacks of conventional recurrent neural networks in encoder-decoder architectures with multi-headed self-attention were overcome by this method. Transformer architectures outperform those built on recurrent networks in terms of speed. Transformer performed more effectively for language translation tasks than even previously reported ensemble models.

This work Pearce et al. (2021)investigates the performance of different pre-trained language models with the goal of determining if they can be fully generalized over a range of QA datasets. QA datasets vary in complexity, challenging models with various levels of reasoning. The study trains and fine-tunes pre-trained language models on a spectrum of datasets to identify models excelling in comprehensive generalization. The paper investigated whether enhanced bidirectionality improves QA model performance with BERT-BiLSTM architecture. Using the F1-score metric, the research identifies RoBERTa and BART as consistently outperforming others. BERT-BiLSTM also surpasses the baseline BERT model. The study sheds light on how QA models generalize and the impact of bi-directionality, contributing to robust systems for nuanced reasoning across domains. Future studies could investigate the wider effects of bidirectionality on language understanding and tailor pre-trained models for QA tasks. Covid-Twitter-BERT (CT-BERT) presented inMüller et al. (2023) was pre-trained on Covid-19-related Twitter messages and also utilized BERT large as a base model. The study Alsentzer et al. (2019) also supports the utilization of domain-specific models.

3. This Zayats et al. (2021) study's approach involves widening the BERT architecture to consider table inter-cell connections. A sizable table corpus taken from Wikipedia is used to retrain the parameters for these associations. Furthermore, by paying attention to relevant text representations in the surrounding article, a novel strategy improves table representations. By considering the contextual relationship between tables and text, the suggested method seeks to offer a more practical and efficient way to answer questions from documents. They laid the groundwork for a more comprehensive understanding of complex documents by integrating Text-Based and Table-Based Approaches.

Although BERT has shown unmatched ability to comprehend language, innovative approaches are needed when applying it to language generating problems. The research under review Chen et al. (2019) introduces the cascading masked language model(C-MLM) as a novel method that provides a mechanism to modify BERT for target generation tasks fine-tuning. Technique entails BERT's fine-tuning, acting as a "teacher" model for the goal generation tasks. Then, this improved BERT model serves as an extra supervisory source, augmenting traditional Sequence-to-Sequence (Seq2Seq) models, also called "students." This teacher-student approach improves the performance of Seq2Seq models in text production. The experiments show notable gains in performance over robust Transformer baselines in a variety of language generation tasks, such as summarized text and automatic translation.

Pre-training of Deep Bidirectional Transformers for Language Understanding According to this studyDevlin et al. (2018), BERT jointly trains on both left and right contexts across all layers to pretrain deep bidirectional representations from unlabeled text. So, by fine-tuning the previously trained BERT model with just one additional output layer, advanced models for a range of tasks, such as question answering and language inference, can be generated without necessitating significant modifications to the task-specific architecture. F1 score of 93 percent. showed empirical success for the question-answering task on SQuAD v1.1. and SQuAD v2.0 Test F1 to 83.1 percent PaperWadhwa et al. (2018) also compared the previous work done on the SQuAD dataset.

This study Kim et al. (2022)introduced an automated approach for extracting infrastructure damage information from textual data using BERT) and question answering (QA). The proposed method, trained on National Hurricane Center reports, demonstrates high accuracy in hurricane and earthquake scenarios, outperforming traditional methods. The method involves two steps: 1) Paragraph Retrieval using Sentence-BERT and 2) Information Extraction with a BERT model. The model was trained on 533 question-answer pairs from hurricane reports and tested on diverse datasets, achieving F1 scores of 90.5 percent and 83.6 percent for hurricanes and earthquakes. This research presented an innovative BERT-based QA approach for automated infrastructure damage retrieval, contributing to improved disaster management. Researchers were optimistic about generalizing the model to other disaster types.

This study Adhikari et al. (2019) pioneers the application of BERT to document classification, achieving state-of-the-art results across four datasets. Despite initial concerns, the proposed BERT-based model surpasses previous baselines, addressing computational expenses through knowledge distillation to smaller bidirectional LSTMs. This achieves BERT base parity with 30× fewer parameters on multiple datasets. Contributions include improved baselines for future document classification research, reflecting a change in basic assumptions in NLP towards pre-trained deep language representation models like BERT. The research highlights the feasibility of distilling BERT into simpler models for competitive accuracy with reduced computational cost.

BERT has demonstrated remarkable performance across various NLP tasks. This paperLiu (2019) introduced BERTSUM, a simplified BERT variant tailored for extractive summarizing. For extractive summarizing, despite recent neural models, further advancements have hit a wall. This research makes the case for using BERT to improve extractive summarizing performance because of its robust design and large pre-training dataset. The study investigates many BERT-based architectures for extractive summarization and finds that the best results are obtained on the job using a flat design with inter-sentence transformer layers. while the paper Yang et al. (2019) introduced a novel data augmentation technique, leveraging distant supervision for fine-tuning BERT in open-domain QA. challenges were noise and genre mismatch in distant supervision data, model sensitivity to diverse datasets, and hyperparameters.

This survey Mohammed and Ali (2021) analyzes various BERT types, including BioBERT for biomedical texts, Clinical BERT for clinical notes, SciBERT for scientific texts, Roberta as an enhanced version, and DistlBERT for smaller models. SCIBERT outperformed BERT-Base in scientific NLP tasks Beltagy et al. (2019), and its application to tasks like summarizing and answering questions is recommended for future research Mohammed and Ali (2021).

This paper Çelikten et al. (2021) tackles biomedical literature overload with a sequence labeling approach for keyword extraction, utilizing contextual embeddings from XLNET, BERT, BioBERT, SCIBERT, and RoBERTa. It avoids traditional methods, showcasing a 22 percent F1-score improvement. Similarly,Kommaraju et al. (2020) employs bioBERT and SciBERT in biomedical text. Another study Namazifar et al. (2021)finds ALBERT outperforming BERT with fewer parameters and faster training on natural language understanding benchmarks.

Conclusion: It is clear from the literature review that researchers have achieved high accuracy Zayats et al. (2021); Chen et al. (2019)for BERT-based models for various NLP tasksWadhwa et al. (2018), including question answering Pearce et al. (2021); Kim et al. (2022). Study Müller et al. (2023) used Roberta for PDFs containing tables and complex texts, while study Adhikari et al. (2019) illustrated the significance of domain-specific BERT models like BERT, BioBERT, and SCIBERT in the biomedical domain. The study Devlin et al. (2018)successfully achieved better results for document classification tasks using the distilling BERT model. The reviewed papers collectively highlight the versatility and robustness of BERT-based models across various NLP tasks. From language translation to document classification and biomedical question answering, BERT has proven to be a versatile and powerful tool. Some of the challenges were model sensitivity to different datasets, scalability concerns, potential overfitting, and difficulty in applying BERT to generative tasks, underscoring the ongoing complexities in refining these models.

3 Methodology

This section describes methods for implementing a PDF-based question-answering system using BERT base models. BERT is an excellent choice for question-answering (QA) on PDF documents for several reasons. Due to its extensive pre-training on a huge corpus of text material, BERT can acquire an in-depth contextual understanding of language. Understanding the context is essential to understanding the rich and varied content that may be found in PDF documents. Its bidirectional attention mechanism considers both the left and right context for each word in a documentDevlin et al. (2018). This approach is effective for capturing dependencies and relationships within the text, which is essential for accurate question answering. A pre-trained BERT model can be used as the base model for QA answering. Further Fine-tuning the pre-trained model with question-answer pairs specific to PDFs implements the QA system. BERT has a substantial number of parameters, and fine-tuning the pre-trained model with a small collection of question-answer pairs would result in overfitting. To avoid that, a fine-tuned BERT is used, which was trained on the SQAUD dataset.

3.1 Data Collection

Links to download the PDFs are given in the configuration manual.

3.1.1 Financial Domain

For the financial domain, annual reports from Amazon were collected as representative documents. The PDFs were obtained from official sources, ensuring the authenticity and relevance of the financial data.

3.1.2 Biomedical Domain

In the biomedical domain, research papers related to COVID-19 and Diabetes were selected for analysis. The dataset includes papers from reputable journals and conferences, ensuring a diverse and comprehensive coverage of biomedical information. PDFs were downloaded from Google Scholar.

3.1.3 Scientific Domain

Scientific literature documents were sourced from various research papers related to question-answering systems. These papers were selected to represent the breadth of scientific literature and were obtained from Google Scholar.

3.2 Question-Answer Pairs Dataset Creation

3.2.1 Financial Domain

For the financial domain, a dataset of question-answer pairs was manually curated. Questions were formulated to cover various aspects of financial reports, and the corresponding answers were extracted from relevant sections of the annual reports. The dataset includes the 'question,' 'context,' and 'ground truth' columns, where 'context' represents the document chunk and 'ground truth' provides the correct answer.

3.2.2 Biomedical Domain

In the biomedical domain, a similar approach was taken to create question-answer pairs related to COVID-19 research papers. Questions were designed to capture key biomedical information, and answers were extracted from the respective document chunks. The dataset structure includes 'question,' 'context,' and 'ground truth' columns.

3.2.3 Scientific Literature Domain

The creation of question-answer pairs for the scientific literature domain followed a similar methodology. Questions were tailored to cover diverse scientific topics, and answers were extracted from relevant chunks of scientific papers. The dataset structure includes 'question,' 'context,' and 'ground truth' columns.

3.3 Pre-processing

When employing BERT or other transformer-based models for question answering, preprocessing is essential. Preprocessing ensures that the input text is appropriate for BERT models, that are made to handle text in a specific way. The following justifies the requirement for preprocessing:

3.3.1 Text Extraction

For the financial domain, a dataset of question-answer pairs was manually curated. Questions were formulated to cover various aspects of financial reports, and the corresponding answers were extracted from relevant sections of the annual reports. The dataset includes the 'question,' 'context,' and 'ground truth' columns, where 'context' represents the document chunk and 'ground truth' provides the correct answer.

3.3.2 Lowercasing and Stripping

To maintain consistency and reduce redundancy, all text was converted to lowercase. Leading and trailing whitespaces were removed to enhance the uniformity of the data. BERT was trained on a large amount of lowercase text. For the optimal performance of the BERT, lowercasing of the text was a necessary pre-processing step.

3.3.3 Sentence Cleaning

Prior to model training and evaluation, it is crucial to preprocess the raw text data to enhance the quality and relevance of information. This involves cleaning sentences to ensure uniformity and remove noise. The following functions were employed for sentence cleaning:

- clean_sentence():This function is designed to clean individual sentences. Converts the sentence to lowercase. Removes special characters using regular expressions. Optionally removes stopwords, leveraging the gensim library's remove_stopwords function.

- get_cleaned_sentences():This function applies the clean_sentence function to a list of sentences. The optional parameter remove_stopwords_flag controls whether stopwords are removed from the sentences.

3.3.4 Tokenization

BERT employs a particular tokenization technique that divides text into smaller pieces known as tokens. The input text is split into words or subwords, and an embedding vector is given to each token. The PDF text is tokenized using the nltkṡent_tokenize method, which tokenizes the text into sentences.

3.3.5 Chunking Strategy

To overcome the token limit of BERT (512 tokens), the PDFs were pre-processed by breaking them into smaller chunks. Each chunk was then tokenized using the appropriate BERT-based model for the respective domain (BERT base for financial, SciBERT for scientific literature, and BioBERT for biomedical).

3.4 Model Implementation

3.4.1 Model Selection

A critical first step in implementing a QA system is choosing suitable pre-trained models. In this section, details of the models chosen for each domain and the rationale behind these selections are given and it was hugely inspired by the literature review conducted [9,10]. BERT base and BERT large models were utilized for the financial domain. For the scientific literature domain, SciBERT, a BERT model pre-trained on scientific text, was employed. In the biomedical domain, BioBERT, pre-trained on biomedical literature, was used. For the financial domain, two variants of BERT models were utilized: BERT Base Model: A base BERT model was employed to capture general financial information and nuances. BERT Large Model: A larger version of BERT was utilized to grasp more complex financial patterns and relationships within the text.

$$\boldsymbol{F_1 = \frac{2}{\mathbf{recall}^{-1} + \mathbf{precision}^{-1}} = 2 \cdot \frac{\mathbf{precision} \cdot \mathbf{recall}}{\mathbf{precision} + \mathbf{recall}}}$$

Figure 2: Calculation of F1 score

- Financial Domain For the financial domain, two variants of BERT models were utilized: BERT Base Model: A base BERT model was employed to capture general financial information and nuances. BERT Large Model: A larger version of BERT was utilized to grasp more complex financial patterns and relationships within the text.

- Biomedical Domain In the biomedical domain, a specialized BERT model pre-trained on biomedical literature, known as BioBERT, was chosen. BioBERT is appropriate for the study of COVID-19 research publications since it is designed to comprehend the distinct terminologies and ideas found in biomedical texts.

- Scientific Literature Domain For the scientific literature domain, we utilized SciBERT, a BERT model pre-trained on a diverse range of scientific texts. SciBERT is designed to capture the intricacies of scientific language, making it suitable for extracting information from research papers and scientific literature.

3.4.2 Model Fine-tuning

To adjust a pre-trained BERT model to a specific task or domain, fine-tuning entails training the model on a domain-specific dataset. With hundreds of millions to more than 300 million parameters, BERT is an extensive neural network architecture. Thus, overfitting would occur if a BERT model were trained from scratch on a small dataset. A refined, pre-trained BERT model that was trained on a sizable dataset is preferable. Using data from the Stanford Question Answering Dataset (SQuAD), the BERT model has been improved.

3.5 Question Answering Setup

The task of question answering was framed as identifying relevant information within the chunks. Domain-specific BERT models for QA were used for question-answer pairs created for each dataset.

3.6 Evaluation

3.6.1 Metrics

The models' performance was assessed using the F1 score, an accepted measure for question answering. Initially, a confidence score was also used to check how confident the is model in predicting answers. F1 score is a popular and extensively used measurement in quality assurance for classification problems. In cases where we value recall and precision equally, it is appropriate. The foundation of the F1 score is the number of words that are shared between the prediction and the truth: recall is the ratio of shared words to the total number of words in the ground truth, and precision is the ratio of shared words to the total number of words in the prediction.

3.6.2 Cross-Domain Evaluation

To assess the models' generalizability, cross-domain evaluation was performed by testing fine-tuned BERT large models on datasets from other domains. This helps understand the adaptability and transferability of the models across diverse types of documents. When the performance of fine-tuned BERT was tested on the Biomedical and Scientific domains following insights were driven: 1. For the biomedical domain, both BERT large and Bio-clinical BERT gave the partial answers for some question-answer pairs and no answers for a few pairs. 2.For the scientific domain, both BERT large and SciBERT models predicted answers partially correct.

3.6.3 Comparative Analysis

Comparisons were made between the performance of the BERT large model within each domain. Additionally, insights were drawn from the cross-domain evaluation to identify potential areas for improvement.

4 Design Specification

In this section, the foundational elements underpinning the implementation of the BERT-based QA system, catering specifically to the distinct characteristics of the financial, biomedical, and scientific domains.

4.1 Techniques

BERT-based QA system integrates several key techniques to address the unique challenges posed by diverse domains: Domain-Specific Fine-Tuned BERT: For each of the domains, a domain-specific fine-tuned BERT model was employed.

- Domain-Specific Fine-Tuned BERT: For each of the domains, a domain-specific fine-tuned BERT model was employed.
- Transfer Learning: Transfer learning in BERT (Bidirectional Encoder Representations from Transformers) involves leveraging pre-trained models on large corpora and fine-tuning them for specific downstream tasks. Google's BERT algorithm has demonstrated impressive results across a range of natural language processing (NLP) applications. The key idea behind transfer learning in BERT is to utilize the pre-trained knowledge encoded in the model's parameters and adapt it to a particular task or domain with limited labelled data [19].

4.2 Architecture

4.2.1

Multi-Head Attention Mechanism: A multi-head attention mechanism in the architecture [20] enables the model to focus on different parts of the input text at the same time. This is especially beneficial for capturing complex relationships and context within diverse domain-specific documents.

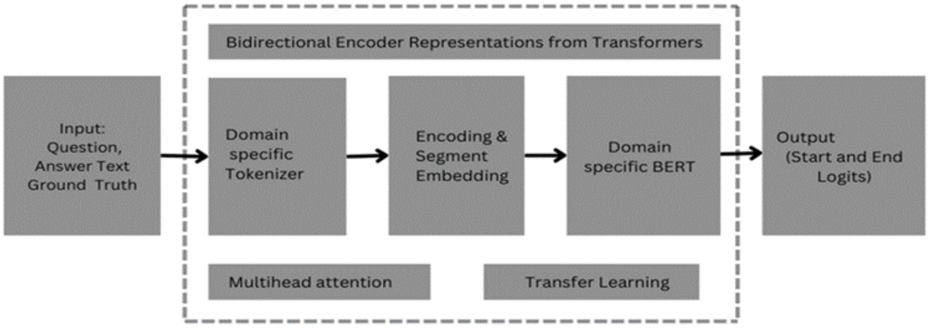
Bidirectional Encoder Representations from Transformers
Input: Question, Answer Text Ground Truth
Domain specific Tokenizer
Encoding & Segment Embedding
Domain specific BERT
Output (Start and End Logits)
Multihead attention
Transfer Learning

Figure 3: Question answering system using fine-tuned BERT

4.2.2

Domain-Specific Embeddings: We utilize domain-specific embeddings to augment the pre-trained BERT embeddings. These embeddings are tailored to the vocabulary and context prevalent in the financial, biomedical, and scientific domains.

4.3 Framework

Implementation is built on the PyTorch framework, providing a robust and flexible platform for deep learning.

- PyTorch Transformers Library: PyTorch Transformers library was used, which facilitates seamless integration with pre-trained BERT models. This library offers a comprehensive set of tools for tokenization, model configuration, and training.

4.4 New Algorithm Description

4.4.1 Algorithm Functionality

BERT-based QA system for financial, biomedical, and scientific domains introduces the following functionalities: Document Chunking Strategy: Given the potentially lengthy and complex nature of documents in these domains, our system employs a document chunking strategy to handle large texts efficiently, ensuring that relevant context is preserved.

4.4.2 Algorithm Requirements

To implement and deploy QA system successfully, certain requirements must be met: Preprocessing Modules: Custom preprocessing modules are designed to handle data cleaning, tokenization, and embedding generation. Hardware Acceleration: For optimal performance, the system benefits from hardware acceleration, such as GPUs provided by Google Collab, to expedite training and inference.

4.5 Tools and Languages

The implementation leveraged the following tools and languages: Programming Language: Python's extensive libraries and versatility in the fields of data science and machine learning led to its selection as the main programming language. Machine Learning Frameworks: Machine learning models were implemented and trained with the help of the scikit-learn, TensorFlow, PyTorch, and Hugging Face Transformers libraries.

5 Implementation

In the final stage of the implementation, the input sequences are prepared for processing by the model, adhering to the model's specific input requirements and constraints.

5.1 Domain-Specific Implementation

Each domain requires a tailored approach to model implementation due to the distinct characteristics of the data. Below, we provide an overview of the strategies employed in each domain.

5.1.1 Financial Domain

The financial domain involves the analysis of annual reports from Amazon. The fine-tuned BERT base and BERT large models were used on a curated dataset of financial question-answer pairs. The stages of implementation of the QA system are as follows:

- With pipeline library: When models were implemented with pipeline library, confidence scores for both BERT base and BERT large models were extremely low even though the answers were correct.

- With tokenization and segmentation: Then models were implemented with different approach where pre-processed input question and answer text were tokenized using the pretrained tokenizer. The tokenized input is then segmented into question-and-answer segments. A pretrained model was trained with the tokenized and segmented input to estimate the beginning and ending positions of the response within the input text. Post-processing is used to handle any spaces at the start of the answer tokens after model inference.The last answer is reconstructed by concatenating these tokens.

- With chunking strategy: Chunking strategy refers to the process of breaking down a large document, such as a PDF, into smaller chunks or segments to be processed by a model. Due to limitation of 512 tokens Bert models were not efficient for long documents as it will only consider first 512 tokens. To overcome that limitation input text was stripped into chunks of 512 tokens and then fed to the model in a loop. While chunking can be effective in handling lengthy documents, it comes with certain limitations:

  Context Discontinuity: Breaking a document into chunks may result in the loss of contextual information that spans across different chunks. BERT models use context to interpret words, so if a question's pertinent context is divided into two chunks, the model's performance might be impacted.

Answer Span Across Chunks: Sometimes a question's answer can be found in more than one section.If the model processes each chunk independently, it might miss the context necessary to identify the correct answer span that extends beyond a single chunk. Incoherent Context: The chunks processed in isolation might not provide coherent context, leading to potential misunderstandings by the model. Since BERT is meant to record contextual relationships between words, breaking up the text into smaller sections might cause this continuity to be broken.

Increased Complexity: Chunking introduces additional complexity into the preprocessing and post-processing stages. Managing the boundaries of chunks and ensuring a seamless flow of information between them requires careful handling.

- With a Curated Dataset of question-answer pairs

  A dataset of ten examples was created manually from PDFs containing question, context, and ground truth columns. This dataset in CSV format was then read as a data frame and fed to the model to calculate the F1 score. This strategy overcomes the following limitations of the chunking method: Context Preservation: The curated dataset contains question-answer pairs carefully crafted to ensure that the context necessary for answering the questions is preserved. In contrast, chunking large documents may introduce discontinuities in context, potentially affecting the model's performance.

  Reduced Complexity: Utilizing a curated dataset might simplify the training process compared to managing the complexities introduced by chunking. Dealing with context boundaries, overlaps, and potential information loss associated with chunking can be challenging.

  A curated dataset of question-answer pairs has additional advantages as follows: Training Data Quality: If a curated dataset is well-constructed and diverse, it provides a clean and controlled environment for training the model. The model learns from specific examples that are explicitly designed for the task, which can be beneficial in terms of generalization to similar scenarios.

  Task Relevance: If a task is well-represented in the curated dataset, and the questions and answers cover a diverse range of scenarios, a model may perform better compared to a model trained on chunks of documents. This is particularly true if the curated dataset is domain-specific or tailored to the types of documents.

  Reduced Complexity: Utilizing a curated dataset might simplify the training process compared to managing the complexities introduced by chunking. Dealing with context boundaries, overlaps, and potential information loss associated with chunking can be challenging.

  Evaluation: Curated datasets often come with predefined evaluation metrics and benchmarks, like ground truth, making it easier to assess the model's performance and compare it against other models in the field.

  Efficiency: Training on a curated dataset may be computationally more efficient than training on large, chunked documents, especially if the documents are extensive.

- Fine-tuning of DistilBERT on SQAuD dataset:

To fine-tune the pretrained BERT model, Trainer class from the PyTorch library was utilized. A small subset of SQAuD dataset was loaded from Datasets library and was split into train test datasets using the train_test_split method. Then DistilBERT ,a distilled version of BERT was loaded to process question and answer. Then dataset was preprocessed to truncate the context and map the answer tokens to the context. Map function from Dataset library was used to apply preprocessing to the entire dataset. A batch of examples were created using Data Collator. Next step was to define hyperparameters in training arguments such as learning rate, number of epochs and weight decay. After that, the trainer was given training arguments that included the model, dataset, tokenizer, and data collator. Train function was called to finetune the model. This fine-tuned model was saved and used for inference for financial dataset.

5.1.2 Biomedical Domain

In the biomedical domain, fine-tuned BioBERT was applied to COVID-19 research papers. Curated Dataset for biomedical PDFs was tested on the model to predict the answers. A dataset containing question, context, and ground truth columns was used to calculate the F1 score.

5.1.3 Scientific Literature Domain

For the scientific literature domain, SciBERT was employed to analyze scientific research papers. A curated scientific dataset was used to calculate the F1 score.

6 Evaluation

This section presents an in-depth evaluation of the findings from the experimental research conducted in each domain. The analysis focuses on the most relevant findings that contribute to addressing the research question.

6.1 Financial Domain

- Case study 1: Implementation of QA system with QA pipeline library. In this case study, the implementation of a Question Answering (QA) system using a dedicated QA pipeline library is explored. The Hugging Face Transformers library was used to perform question-answering tasks using two different models: "bertlargeuncasedwholewordmaskingfinetunedsquad" and "bertbaseuncased." Preprocessed text from Amazon's annual report was fed to the QA pipeline as context and model were evaluated with a confidence score that measures model confidence. Comparison of the question-answering performance of two different BERT models on a specific question and context helps evaluate how the choice of model can impact the quality of answers provided by the question-answering system. The large model's answer was more relevant and contextually appropriate for the given question about the document's topic. The higher score of 0.44 indicates a higher confidence level compared to the base model, which provided a less relevant answer with a significantly lower score.

Model	Score
BERT base	3.44E-05
BERT large	0.441348344

Table 1: Confidence score of BERT base and BERT large

- Case study 2: Implementation of QA system with Tokenization and segment embeddings text and questions were tokenized using the tokenizer's encoding method. The [SEP] token index separated the question and answer segments. Segment IDs were created, assigning 0s to segment A (question) and 1s to segment B (answer). The tokenized input and segment IDs were passed to the model to obtain outputs. The start and end indices of the predicted answer were determined, and the answer span was constructed by concatenating the corresponding tokens. The score was calculated as the maximum value of the start index, but it does not provide a direct measure of the model's confidence or certainty in the predicted answer. When a small text from PDF was tested, the model predicted the answer correctly (Refer Table No. 2). Further function was modified to incorporate the F1 score, to measure the model's predictions. The limitation was that BERT could only consider 512 tokens. So, this method was not useful for longer documents.

Question	Context	Predicted answer
How much was the net sales in the year 2022?	Net sales increased 13% to $143.1 billion in the third quarter, compared with $127.1 billion in the third quarter of 2022.	"$ 127.1 billion"

Table 2: Predicted answer for question pair with Tokenization strategy

- Case study 3: Implementation of QA with chunking strategy. Further input sequences were divided into chunks of 510 and special tokens [CLS] and [SEP] were added to separate the question and answer. Zero-padding was done to ensure consistent sizes. For each chunk, the answer_question function was called with the question, and the chunk's tokens were converted back to a string as the answer text. Table 3 shows the predicted answer by this strategy.

Question	**Context**	**Predicted answer**
How AWS helped Amazon to grow in the year 2022?	PDF text from Amazon's quarterly report	Segment sales increased 12% year-over-year to $23.1 billion. Operating income increased to $11.2 billion in the third quarter, compared with $2.5 billion in the third quarter of 2022. North America segment operating income was $4.3 billion, compared with an operating loss of $0.4 billion in the third quarter of 2022. International segment operating loss was $0.1 billion, compared with an operating loss of $2.5 billion in the third quarter of 2022. AWS segment operating income was $7.0 billion, compared with an operating income of $5.4 billion in the third quarter 2022.

Table 3: Predicted answer for question pair with chunking strategy

- Case study 4: Implementation of QA with curated dataset. A curated dataset from the PDF text was created manually to test the model's performance for multiple questions. This approach overcomes the limitation of the chunking strategy which could cause a loss of context and it was more efficient in evaluating long text as well.

Question	Context	Ground Truth	Predicted answer	F1 score
What were Amazon's net sales in the first quarter of 2023?	PDF Text	Net sales increased 9% to $127.4 billion in the first quarter, compared with $116.4 billion in the first quarter 2022.	$127.4 billion	0.039
How much did net sales increase compared to the first quarter of 2022?	PDF Text	Excluding the $2.4 billion unfavorable impact from year-over-year changes in foreign exchange rates throughout the quarter, net sales increased 11% compared with the first quarter of 2022.	9%	0.0
How did North America segment sales change year-over-year?	PDF Text	North America segment sales increased 11% year-over-year to $76.9 billion.	Foreign exchange rates	0.199
What was the operating income for AWS segment?	PDF Text	AWS segment operating income was $5.1 billion, compared with operating income of $6.5 billion in the first quarter of 2022.	$5.1 billion	0.0
How did the operating cash flow change for the trailing twelve months?	PDF Text	Operating cash flow increased 38% to $54.3 billion for the trailing twelve months, compared with $39.3 billion for the trailing twelve months ended March 31, 2022.	Net sales increased 9%	0.074

Table 4: Predicted answers and F1 scores for curated financial dataset

- Case study 5: Implementation of QA with fine-tuned DistilBERT. The next step was to see if fine-tuning the BERT model improves the score. The distilling version of BERT was loaded and finetuned with hyperparameters learning_rate=1e-5, num_train_epochs=3, per_device_train_batch_size=8, per_device_eval_batch_size=8. When the fine-tuned model was inference for simple QA pairs, the confidence score was 0.250225812,(Table no 5). The question-answer pairs of the financial dataset were inference to evaluate the performance. The results are given in Table No 6.

Question	Context	Predicted answer	Score
What are different search engines?	BLOOM has 176 billion parameters and can generate text in 46 natural languages and 13 programming languages.	176 billion	0.250225812

Table 5: Result for small question-answer pair on fine-tuned BERT

Question	Context	Predicted answer	Score
What were Amazon's net sales in the first quarter of 2023?	PDF text	$127.4 billion	0.07002584
How much did net sales increase compared to the first quarter of 2022?	PDF text	$127.4 billion	0.070099174
What was the impact of foreign exchange rates on net sales?	PDF text	$116.4 billion	0.070099174
How did North America segment sales change year-over-year?	PDF text	$116.4 billion	0.080088
How did the operating cash flow change for the trailing twelve months?	PDF text	$127.4 billion	0.03641737

Table 6: Results for Financial dataset on fine-tuned BERT

6.2 Bio-medical Domain

- Case study 6: Implementation of QA using pre-trained Bio-BERT with curated dataset. PDF text from the Covid research paper was fed to the Bio-Clinical BERT on pre-trained on biomedical and clinical text. A curated dataset of 10 question-answer pairs was tested on the model and evaluated with an F1 score. The results are given in Table no 7.

- Case study 7: Implementation of QA using pre-trained BERT large with curated dataset. For cross-domain evaluation, the Bio-medical dataset was then tested on a fine-tuned BERT large model to assess the model's generalizability to other domains and to investigate which model performs the best for bio-medical documents.

Question Answer pairs	Bio-ClinicalBERT F1-Score	Bio-ClinicalBERT Average F1	BERT Large F1-Score	BERT Large Average F1 score
1	0.035960107	0.033	0.028023353	0.043
2	0.037663124		0.083908594	
3	0.030660377		0.038654608	
4	0.034029851		0.040029685	
5	0.032238806		0.122826651	
6	0.026143791		0.022364851	
7	0.031790556		0.022641509	
8	0.038139441		0.025694735	
9	0.023529412		0.024219489	
10	0.040635086		0.025724567	

Table 7: Comparison of F1 scores of BioBERT and BERT Large for biomedical dataset

6.3 Scientific Domain

- Case study 8: Implementation of QA using pre-trained Sci-BERT with curated dataset. Pre-trained Sci-BERT trained on a large corpus of scientific literature, including scholarly articles, research papers, and other documents from the biomedical and life sciences domains. The model was tested for F1 score.

- Case study 9: Implementation of QA using pre-trained BERT large with curated dataset. To test the pre-trained BERT model's generalizability, a scientific dataset was also evaluated on the BERT large model. Results are given in Table no 8.

Question Answer pairs	SciBERT F1-Score	SciBERT Average F1	BERT Large uncased F1-Score	BERT Large uncased Average F1 score
1	0.050174068	0.053	0.026548673	0.053
2	0.029270798		0.110177404	
3	0.091635101		0.058608186	
4	0.070933213		0.045015747	
5	0.044015319		0.05486294	
6	0.062091503		0.029445595	
7	0.038116426		0.031267685	
8	0.036177633		0.071584253	
9	0.054175516		0.034844972	
10	0.060310427		0.068245615	

Table 8: Comparison of F1 scores of SciBERT and BERT Large for scientific literature dataset

7 Discussion

Developing a QA system for PDFs is a challenging task since PDFs may contain complex text, tables, images, or complex layouts. PDFs related to the financial, biomedical, and scientific sectors are even more complex and time-consuming to comprehend. To utilize pre-trained BERT models for the respective domains, experiments were carried out to implement a QA system for the chosen PDFs. From case study 1, BERT Large gives a 44% score (see Table No. 1) for the pre-processed financial PDF text. The research was successful in implementing a QA system for the smaller texts. As we can see from Case Study 2, the model predicts the answer correctly (see Table 2). The chunking strategy implemented in Case Study 3 successfully overcomes the limitation of 512 tokens in the BERT model (see Table 3). The limitation of case study 3 was overcome in case study 4 with a curated dataset that preserves the context (see Table 4). Case Study 5 implemented fine-tuning of pre-trained DistilBERT. The model's confidence score was slightly higher for the simple question-answer pairs (see Table 5) than for the complex texts (see Table 6). However, as the research progressed to design an end-to-end QA system on longer PDFs, the following limitations were found during the experiments:

a) Chunking text into 512 tokens is only useful for small PDFs. Amazon's annual reports used for analysis are 16 pages long and the prototype developed here lacks the implementation for longer PDFs. Additionally, this could result in context loss and incoherence in answer generation while processing multiple chunks.

b) Low F1 scores for the curated dataset for respective domains as shown in Table 9 suggest that the proposed research needs optimization and should consider fine-tuning the curated dataset.

c) A dataset was created for each domain using only a few pages of the PDFs. Creating datasets manually for fine-tuning and evaluation is challenging.

d)When cross-domain evaluation was conducted in case studies 6,7,8 and 9 did not show much difference. As described in the literature review, previous studies show domain-specific BERT models have achieved significant results for respective domains.

Domain	BERT model used	Average F1 score
Financial	BERT large uncased	0.03913
Biomedical	Bio-ClinicalBERT	0.033
Scientific	SciBERT	0.053

Table 9: Average F1 scores for the respective domains.

8 Conclusion and Future Work

This study intended to utilize the pre-trained BERT models for implementing a QA system on PDFs from various domains. Several strategies were used to implement a QA system for financial, scientific, and bio-medical domains. The proposed research successfully implemented a question-answering pipeline with a pre-trained BERT base and BERT large models. For longer documents, chunking the long text into chunks of 512 and extracting answers from the chunks was implemented successfully. Datasets were created manually for evaluation for the chosen domains with question, context, and ground truth columns. These datasets were tested on different BERT models like BioClinical BERT, SciBERT, BERT large, and DistilBERT. This research poses few limitations such as lower confidence score of BERT models even after fine-tuning with hyperparameters. The creation of correct datasets manually from PDFs was also challenging and needs to be addressed for better evaluation of the models. This research holds the potential to utilize personalized chatbots for various fields like education, medicine, and finance. This research can be extended in the future for the improvisation of the model's confidence score and the creation of question-answer pairs from complex PDFs.

References

Adhikari, A., Ram, A., Tang, R. and Lin, J. (2019). Docbert: Bert for document classification, *arXiv preprint arXiv:1904.08398* .

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* .

Çelikten, A., Uğur, A. and Bulut, H. (2021). Keyword extraction from biomedical documents using deep contextualized embeddings, *2021 International Conference on Innovations in Intelligent Systems and Applications (INISTA)*, IEEE, pp. 1–5.

Chen, Y.-C., Gan, Z., Cheng, Y., Liu, J. and Liu, J. (2019). Distilling knowledge learned in bert for text generation, *arXiv preprint arXiv:1911.03829* .

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* .

Kim, Y., Bang, S., Sohn, J. and Kim, H. (2022). Question answering method for infrastructure damage information retrieval from textual data using bidirectional encoder representations from transformers, *Automation in construction* **134**: 104061.

Kommaraju, V., Gunasekaran, K., Li, K., Bansal, T., McCallum, A., Williams, I. and Istrate, A.-M. (2020). Unsupervised pre-training for biomedical question answering, *arXiv preprint arXiv:2009.12952* .

Liu, Y. (2019). Fine-tune bert for extractive summarization, *arXiv preprint arXiv:1903.10318* .

Mohammed, A. H. and Ali, A. H. (2021). Survey of bert (bidirectional encoder representation transformer) types, *Journal of Physics: Conference Series*, Vol. 1963, IOP Publishing, p. 012173.

Müller, M., Salathé, M. and Kummervold, P. E. (2023). Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter, *Frontiers in Artificial Intelligence* **6**: 1023281.

Namazifar, M., Papangelis, A., Tur, G. and Hakkani-Tür, D. (2021). Language model is all you need: Natural language understanding as question answering, *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 7803–7807.

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* .

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. (2017). Attention is all you need, *Advances in neural information processing systems* **30**.

Wadhwa, S., Chandu, K. R. and Nyberg, E. (2018). Comparative analysis of neural qa models on squad, *arXiv preprint arXiv:1806.06972* .

Yang, W., Xie, Y., Tan, L., Xiong, K., Li, M. and Lin, J. (2019). Data augmentation for bert fine-tuning in open-domain question answering, *arXiv preprint arXiv:1904.06652* .

Zayats, V., Toutanova, K. and Ostendorf, M. (2021). Representations for question answering from documents with tables and text, *arXiv preprint arXiv:2101.10573* .