Question 4

The case study mentioned is about multilabel classification, as an input can be classified with more than 1 labels. The design of the AI application for the classification task can be illustration as below.

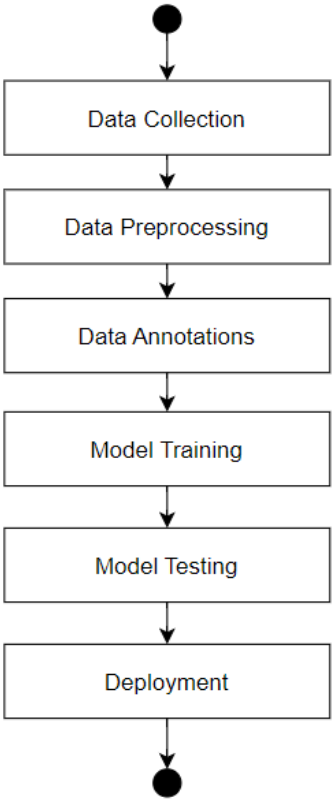

Figure 1: Classification Process

Data Collection

Firstly, data collection. The team can decide a few sectors to dive deep in, for example, telecommunication, financial service, and real estate. Then, the team can browse through some public online datasets such as Reuters, OHSUMED and 20NewsGroups or use web-scraping to collect long PDFs with topic labelling.

The raw data collected can be stored in cloud-based database or data lake.

Data Pre-processing

Several pre-processing techniques can be used to extract useful information from the long text of PDF documents.

- Apply lower casing for consistency
- Do stemming or lemmatization to analyse the word meaning
- Remove punctuations, stopwords, and HTML tagging

After the insignificant words have been removed, the words can be converted into the respective text embeddings. Truncation or paddings can be done to ensure each sentence have the same length.

Data Annotations

Since the multilabel classification needs to be carried out, the database table storing the PDFs is designed as shown as below with some important columns.

document_id	healthcare	telecommunication	financial_service	real_estate
001	1	0	1	0
002	0	1	0	1
003	0	0	1	0
004	0	1	0	1

The approach that used to treat this multilabel classification is using label powerset to transform the case into multiclass classification. Each unique label combination is considered as a different class. For example,

id	healthcare	telecommunication	financial_service	real_estate	final_label
001	1	0	1	0	1
002	0	1	0	1	2
003	0	0	1	0	3
004	0	1	0	1	2

- Class 1 is the combination of 1010,
- Class 2 is the combination of 0101,
- Class 3 is the combination of 0101, etc.

Model Training and Model Testing

The BERT model is fine-tuned to adapt to the case study of topic modelling. The text embeddings of the documents are being fed into the BERT classifier model and some parameters can be fine-tuned such as batch size, learning rate, and number of epochs. Training and testing accuracy shall be measured to check if the BERT classifier performs well or overfit the case study.

Confusion matrix can be used to visualize the model performance, by getting the figures of True Positive cases, True Negative cases, False Positive cases, and False Negative cases for the respective class.

Based on the confusion matrix, some other performance metrics can be derived such as precision, sensitivity, and f1-score.

- Precision represents the number of positive class predictions that belong to the positive class.
- Recall represents the number of positive class predictions out of all positive examples.
- F1-score balances both the concerns of precision and recall in one number.

Deployment

Once the model trained as reached a certain performance, it can be deployed into staging and production for further testing and improvements. The model weights can be stored, and an API request can be designed to allow user interface to use the features provided by this AI application for topic predictions.

Potential Challenges and Limitations

The potential challenges to develop this application are as below:

- Large amount of training data required to train the model
  - It is quite hard to find such an enormous number of long PDF documents online
  - The server storage should be huge and capable enough to support
- Pre-processing task which is time consuming
  - It takes 1-2 minutes to load PDF documents that is more than 100 pages in reality
  - Pre-processing tasks will extend the time taken longer
- Highly reliable and stable computing resources
  - The resources should be stable to support the data storage and data pre-processing to avoid unexpected system interrupt or crash, and data leakage.
- Overfitting might happen
  - The large number of classes might lead to overfitting that is undesirable

The limitations

- Performance of AI application
  - The performance might be slow due to the complexity and length of the PDF documents that make the system to load the input for a certain period.
- Quality assurance of the annotated data
  - The annotated data might not be always comprehensive and reliable, as the topic modelling is quite subjective to be decided among people