- AIdea 醫病訊息決策與對話語料分析競賽 -

2021 春季賽：醫病決策預判與問答

組名：NTUNLP_衝衝衝鴨

組員：R09946001 陳知遙 、 R09946006 何青儒、 R09946021 黃瀞瑩

I. 任務一 ： 決策預判與風險評估

1. **目的：**從診間的對話脈絡，判斷該求診民眾是否需要再進一步評估，例如感染疾病的風險高低、是否需要再回診或檢查等。

2. **資料集**

資料集	檔案名稱	數量
訓練資料集	Train_risk_classification.csv	346
開發資料集	Develop_risk_classification.csv	100
測試資料集	Test_risk_classification.csv	5314

3. **預處理**

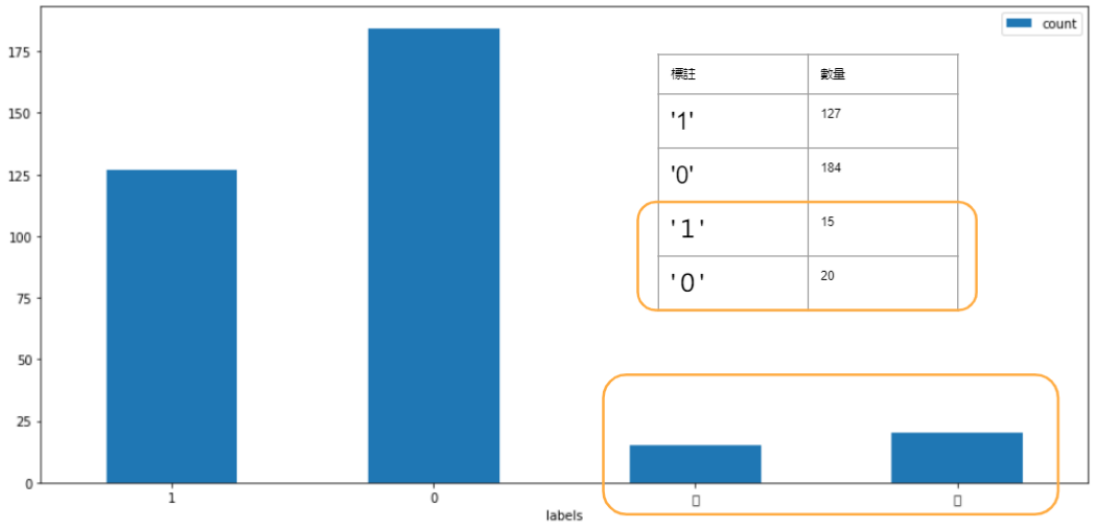
   a. **統一標籤：**將未統一的標籤轉換為統一格式。原始資料集中的標籤因人工或其他因素而導致了錯誤的標柱，如圖一所示半形與全形混用的情況，這會導致資料集中的類別數量不正確，因此我們在預處理的階段中將類別統一為「0」和「1」兩類。

4. **方法**

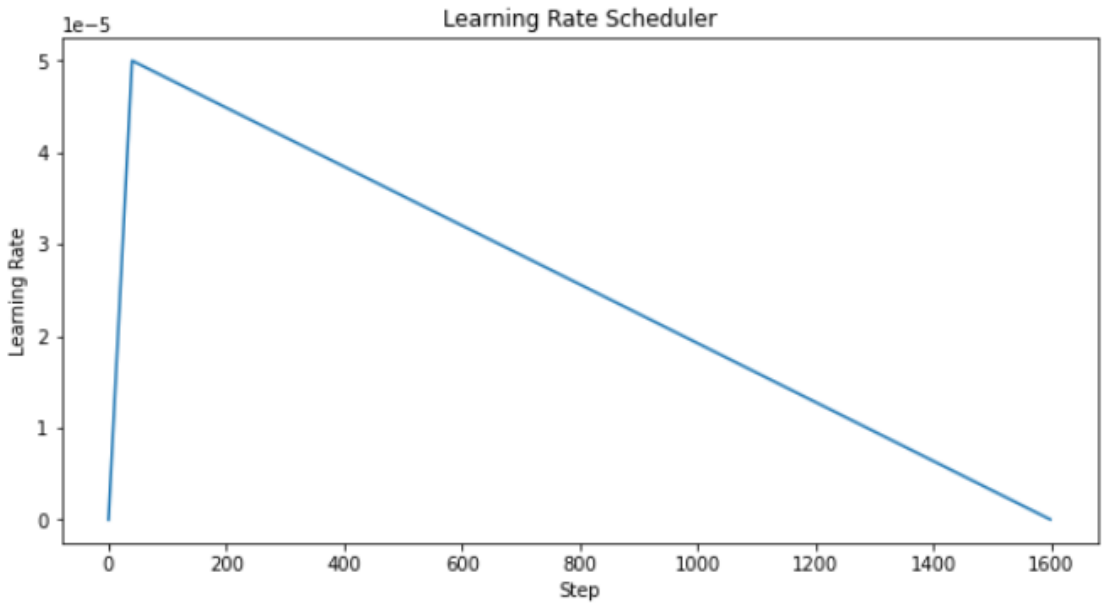
   a. **文本字句刪減：**對於此項任務，我們其實很難定義特定的指標來針對對話內容做篩選，而最簡單且直接的方式就是篩掉重複性極高的詞彙，以及其他較短的句子。例如將「蛤？」、「嗯。」、「對。」、

「個管師」、「民眾」、「醫師」等詞彙刪除，僅專注在對話的實際內容。

b. **模型訓練技巧：**Learning Rate 是在模型訓練過程中非常重要的參數之一，合適的 Learning Rate 可以幫助模型快速收斂。這裡使用 lr_scheduler.LambdaLR 訓練技巧，給予各個訓練階段的模型不同的 Learning Rate，如圖二所示，使得模型快速收斂。



▲ 圖一：黃色框框的全形將被視為錯誤標記，會將它們處理回半形。



▲ 圖二：Learning Rate 在使用 LambdaLR 訓練方式下的變化。

5. 模型與參數

a. 模型： BertForSequenceClassification

b. 參數

Pretrained Model	Bert-base-chinese
Optimizer	Adam + Linear Schedule Warmup (warmup steps = 40)
Learning Rate	5e-5
Betas	(0.9, 0.98)
Eps	1e-9
Epochs	80
Batch Size	16
Max Grad Norm	1.0
Input Lengths	400

6. 在資料集上的表現

資料集	表現 (AUROC)
開發資料集	0.7100
測試資料集	0.6827

II. 任務二 ： 醫病問答

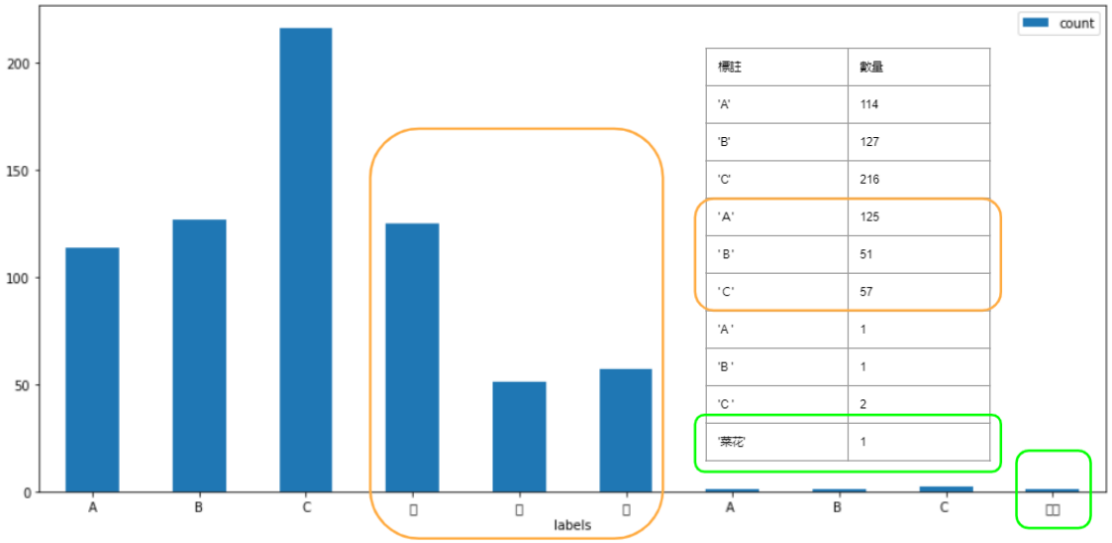
1. **目的：**從門診對話資料中擷取出與該次看診狀況有關的問題，並從複數的選項中選出正確的答案。

2. **資料集**

資料集	檔案名稱	總筆數
訓練資料集	Train_qa_ans.json	694
開發資料集	Develop_QA.json	192
測試資料集	Test_QA.json	12921

3. **預處理：**

   a. **觀察標籤形態：** 將未統一的標籤轉換為統一格式，並將含錯誤標籤的資料刪除，如圖三所示。這同樣是為了將雜訊去除並將同樣的類別正確歸類，使得模型能夠更好的學習並達到我們期望的效果。



▲ 圖三：黃色框框中的全形標記將被視為錯誤標記，會將它們處理回半形。
至於綠色框框中的標記將被視為雜訊刪除。

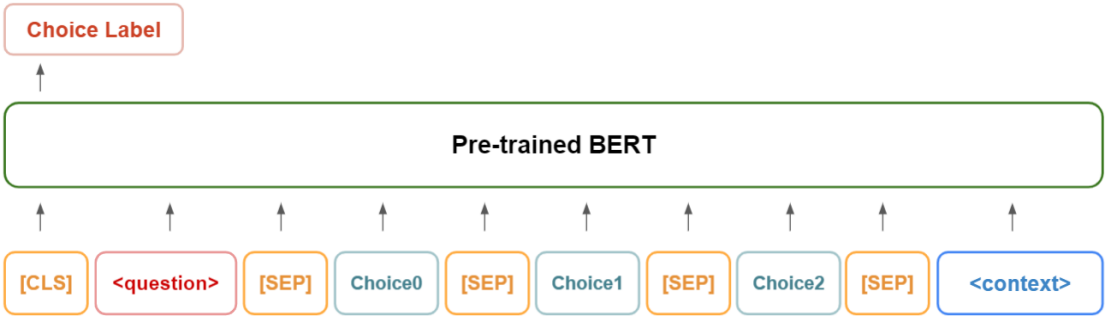
4. **方法**

a. **萃取重要語句：**由於對話性資料集的天性是有許多與主題或問答並不相關的閒聊，這將會導致 (1) 無法將所有資訊輸入給 BERT 模型，以及 (2) 不重要的訊息佔據輸入太多的篇幅兩個問題。
因此我們嘗試了以下方法：

- **方法 1：**我們將對話內容以句號、問號、驚嘆號等標點符號作為斷句點，並將問題文字與斷句後的對話內容分別以 Sentence-BERT[1] 的方式來進行編碼，再計算問題與對話的餘弦相似度（Cosine Similarity）。
  對於每個樣本，我們只保留問題與對話斷句後前 10 相似 (k=10) 的句子，從而大大的減少了每個樣本的 token 數，達到去蕪存菁的效果。

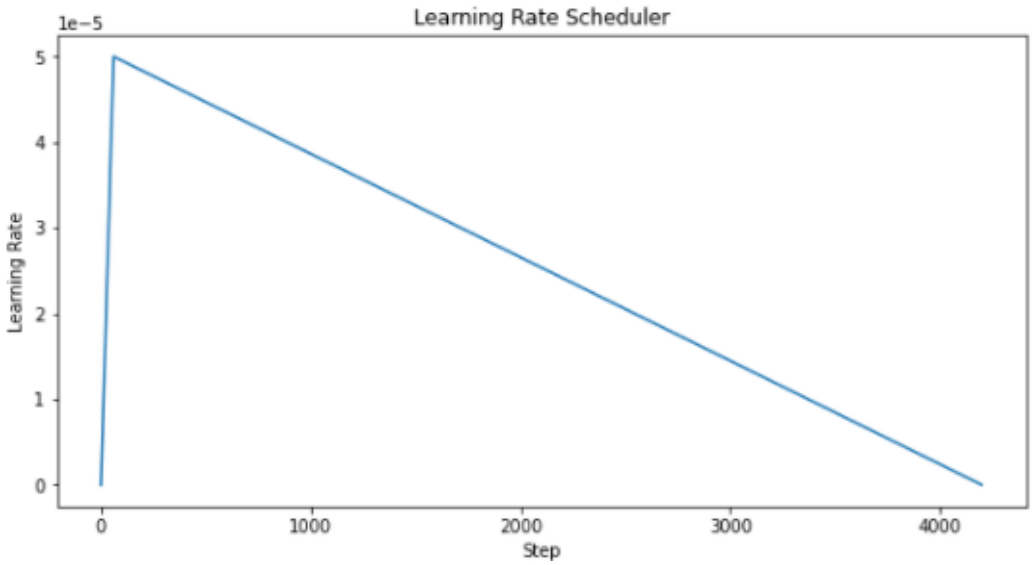
- **方法 2：**我們先將對話內容以句號來做斷句，觀察每一句斷句後的對話內容，若內容裡頭所包含的單詞與問題單詞重疊，且重疊單詞數大於五，我們就會選擇該句、以及它往後的兩句對話內容，作為與問題有關且含資訊的對話文本。
  而最終的結果則顯示，透過這個方法會有比較好的表現，所以可以解釋為，問題的答案會出現在與問題有關（重疊單詞）的句子附近。

b. **文本輸入格式：**主要以作答方式習慣設定的輸入格式，一般人首先會先理解問題，觀察問題所給出的選項，再從文章裡頭找出相對應的解答，故以這樣的概念將 [SEP] 標註穿插於輸入文本中，有效的間隔開問題、選項以及文章。



▲ 圖四：將 [SEP] 穿插在斷句後的詞 token 之中，然後再一起送入模型裡面。

c. **模型訓練技巧：**一樣使用 lr_scheduler.LambdaLR 訓練技巧。



▲ 圖五：Learning Rate 在使用 LambdaLR 訓練方式下的變化。

5. 模型與參數

a. **模型：** BertForSequenceClassification

b. **參數：**

Pretrained Model	Bert-base-chinese
Optimizer	Adam + **Linear Schedule Warmup (warmup steps = 60)**
Learning Rate	5e-5
Betas	(0.9, 0.98)
Eps	1e-9
Epochs	100
Batch Size	16
Max Grad Norm	1.0
Input lengths	256

6. 在資料集上的表現

資料集	表現 (Accuracy_score)
開發資料集	0.5237
測試資料集	0.5089

III. 困難與解決方法

1. 訓練資料稀少：由於醫病資料的收集不易，在本次競賽中能取得的資料相當有限，只能透過適當的預處理和 pre-trained BERT 模型具備的強大能力來盡力克服。
2. 對話資料的雜訊多：在醫病問答資料集中可以以肉眼明顯看到有許多偏日常性的寒暄對話，這些語句和問題本身是無關的。若未經過適當的前處理，不僅會讓 GPU 更容易被塞滿，更可能使得真正重要的資訊沒辦法被當作 input 讓 BERT 學習。
   而我們針對這個問題的解決方式是想辦法透過 Sentence-BERT 取得 sentence-level 的 embeddings，如此便可以計算問題與對話中的哪些句子是比較相關的，並保留前 k 相關（在我們的實作中取 k=10）的句子即可。

IV. 組員分工

組員 / 分工	討論	資料處理	程式部分	報告內容	整理排版
陳知遙	v	v	v	v	
何青儒	v	v		v	v
黃瀞瑩	v	v	v	v	