應用漸進學習於文本分類之資料標註加速器

Incremental Learning on Data Annotation Accelerator in Text Classification Tasks


指導教授:高宏宇


組員:楊智翔 林恩締

目錄 CONTENTS

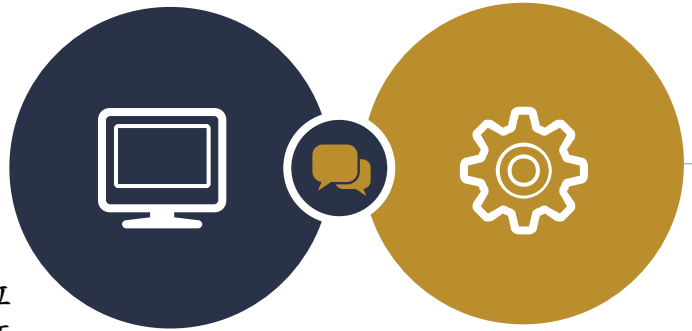
01 動機及目標

動機及目標

Motivation

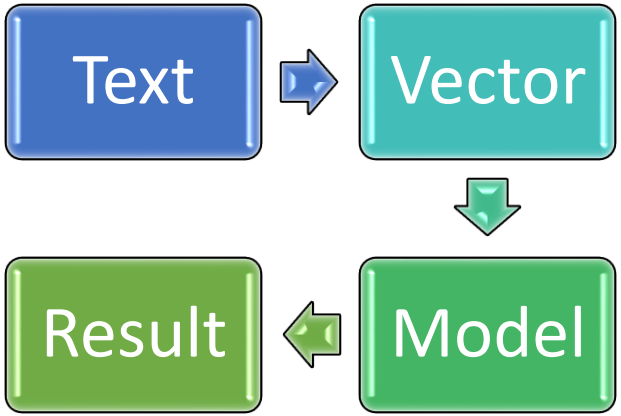
- 情境假設:

今天有一個未標記的資料，有位學生小明想把這些資料進行分類，因此手動開始標記資料，標記了一陣子之後發現這樣需要花費大量時間與精力又耗時，因此他想到他可以建立一個模型可以幫助標記，讓他不用標記完全部的資料就達到他的目標。



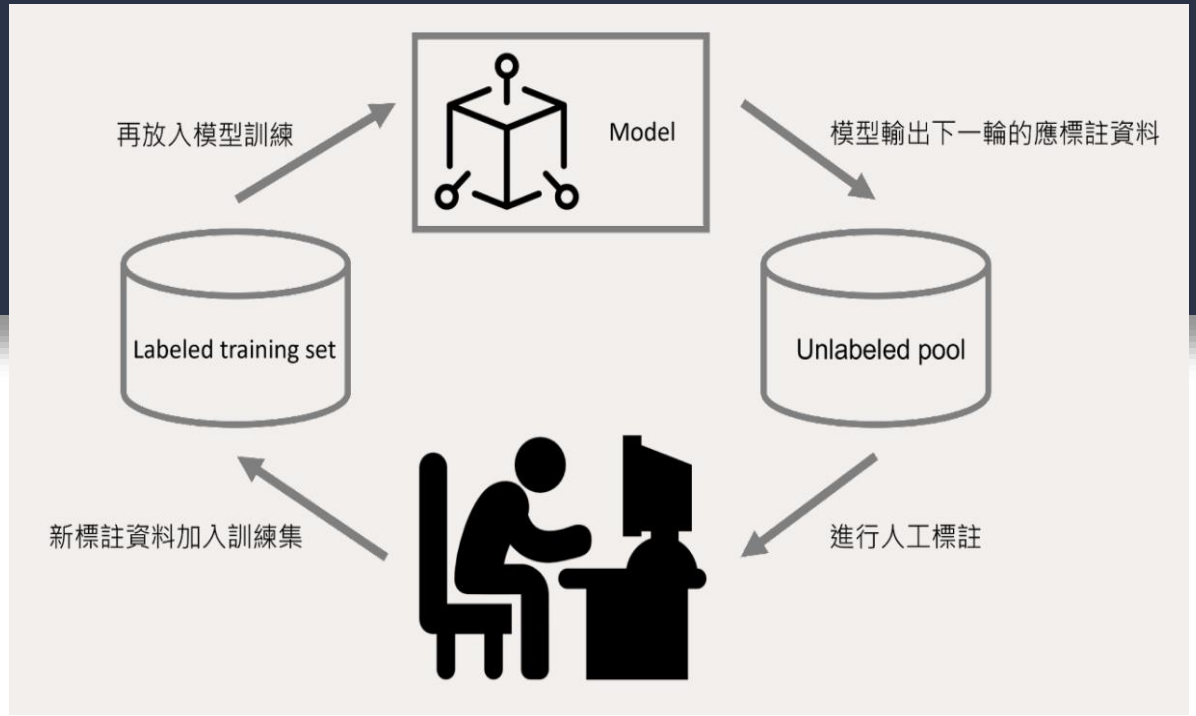
Goal

- 減少需要標記的資料



02 流程圖

流程圖


再放入模型訓練
Model
模型輸出下一輪的應標註資料
Labeled training set
Unlabeled pool
新標註資料加入訓練集
進行人工標註

03 資料一

資料一

1

資料來源: Kaggle Competition Natural Language Processing with Disaster Tweets

原比賽任務:

Predicts which Tweets are about real disasters and which one's aren't.

資料切割:

Labeled training set : 300筆

Unlabeled data pool : 6313筆

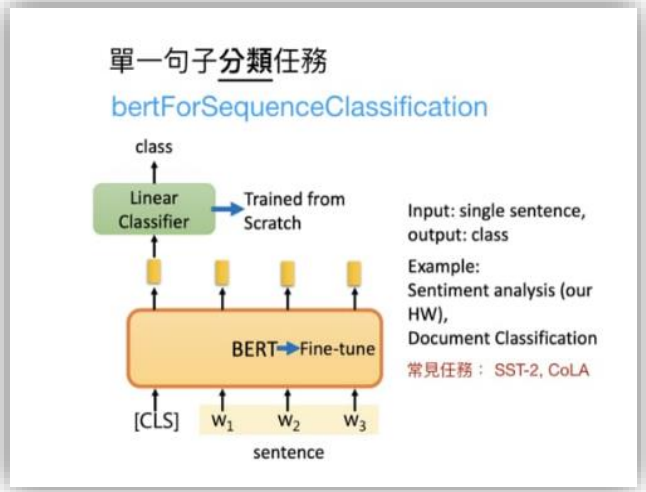
Test set : 1000筆

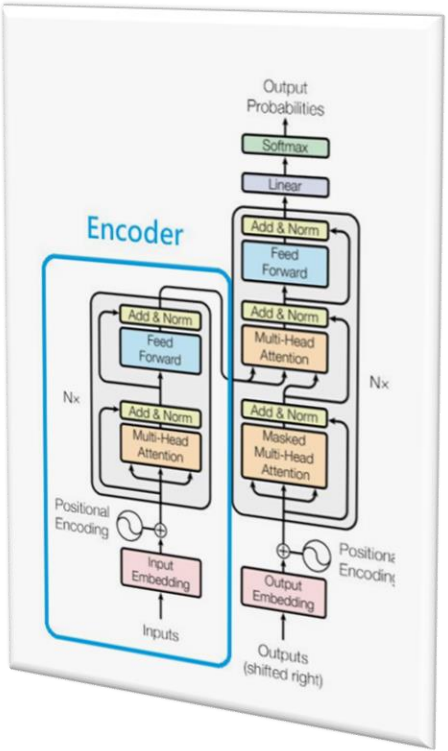
2

部分資料:(左邊為text,右邊為label)

People who say it cannot be done should not interrupt those who are doing it. ??? George I	0
'The first man gets the oyster the second man gets the shell.' Andrew Carnegie	0
Anyone need a P/U tonight? I play Hybrid Slayer ps4 EU. HMU @Cod8sandscrims @En	0
Experts in France begin examining airplane debris found on Reunion Island: French air acc	1
Strict liability in the context of an airplane accident: Pilot error is a common component of	1
@crobscarla your lifetime odds of dying from an airplane accident are 1 in 8015.	0
Experts in France begin examining airplane debris found on Reunion Island: French air acc	1
@AlexAllTimeLow awwww they're on an airplane accident and they're gonna die what a	1
family members of osama bin laden have died in an airplane accident how ironic ????? m	1
Man Goes into Airplane Engine Accident: http://t.co/TYJxrFd3St via @YouTube	1
Horrible Accident Man Died In Wings of Airplane (29-07-2015) http://t.co/i7kZtevb2v	1
A Cessna airplane accident in Ocampo Coahuila Mexico on July 29 2015 killed four men	1
#Horrible #Accident Man Died In Wings Airplane (29-07-2015) #WatchTheVideo http://t	1
Experts in France begin examining airplane debris found on Reunion Island http://t.co/LsM	1
Experts in France begin examining airplane debris found on Reunion Island: French air acc	1
#KCA #VoteJKT48ID mbataweel: #RIP #BINLADEN Family members who killed in ar	1
I almost sent my coworker nudes on accident thank god for airplane mode	0
@mickinyman @TheAtlantic That or they might be killed in an airplane accident in the ni	0
Experts in France begin examining airplane debris found on Reunion Island: French air acc	1

模型: bert (hugging face)





- Encoder of Transformer
- 大量文本以及兩個預訓練目標，事先訓練好一個可以套用到多個 NLP 任務的 BERT 模型，再以此為基礎 fine tune 多個下游任務。
- 預訓練模型: **bert-base-uncased**
- Fine-tuning階段 : **BertForSequenceClassification**
- Loss function: CrossEntropyLoss(), learning rate:1e-6

實作方法

01

先用300筆訓練資料去跑，跑50個epochs。

02

利用模型預測各分類的機率
選出要加入的資料。
策略一：用Least Confident去挑選
策略二：用Entropy去挑選
選完後再給模型去跑30個epochs。

03

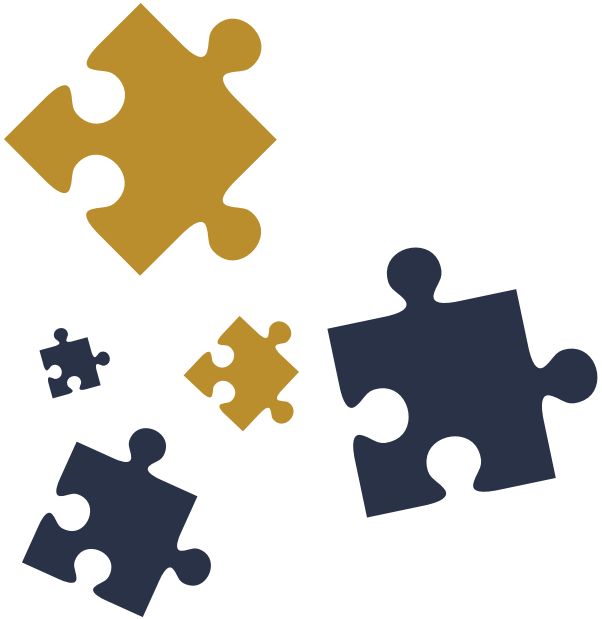
重複步驟二(每次改變取的數字區間)。

比較兩種策略

訓練集資料量 利用:Least Confident	測試集的 fl_score	訓練集資料量 利用:Entropy	測試集的 fl_score
300	0.7414	300	0.7414
2077	0.7870	2482	0.7935
2658	0.7832	3350	0.7866
3119	0.7931	4163	**0.7809**
3459	0.7935		
3814	0.7995		
4152	**0.8061**		

比較結果: 用Least Confident的方法比Entropy適合這個資料集

結果



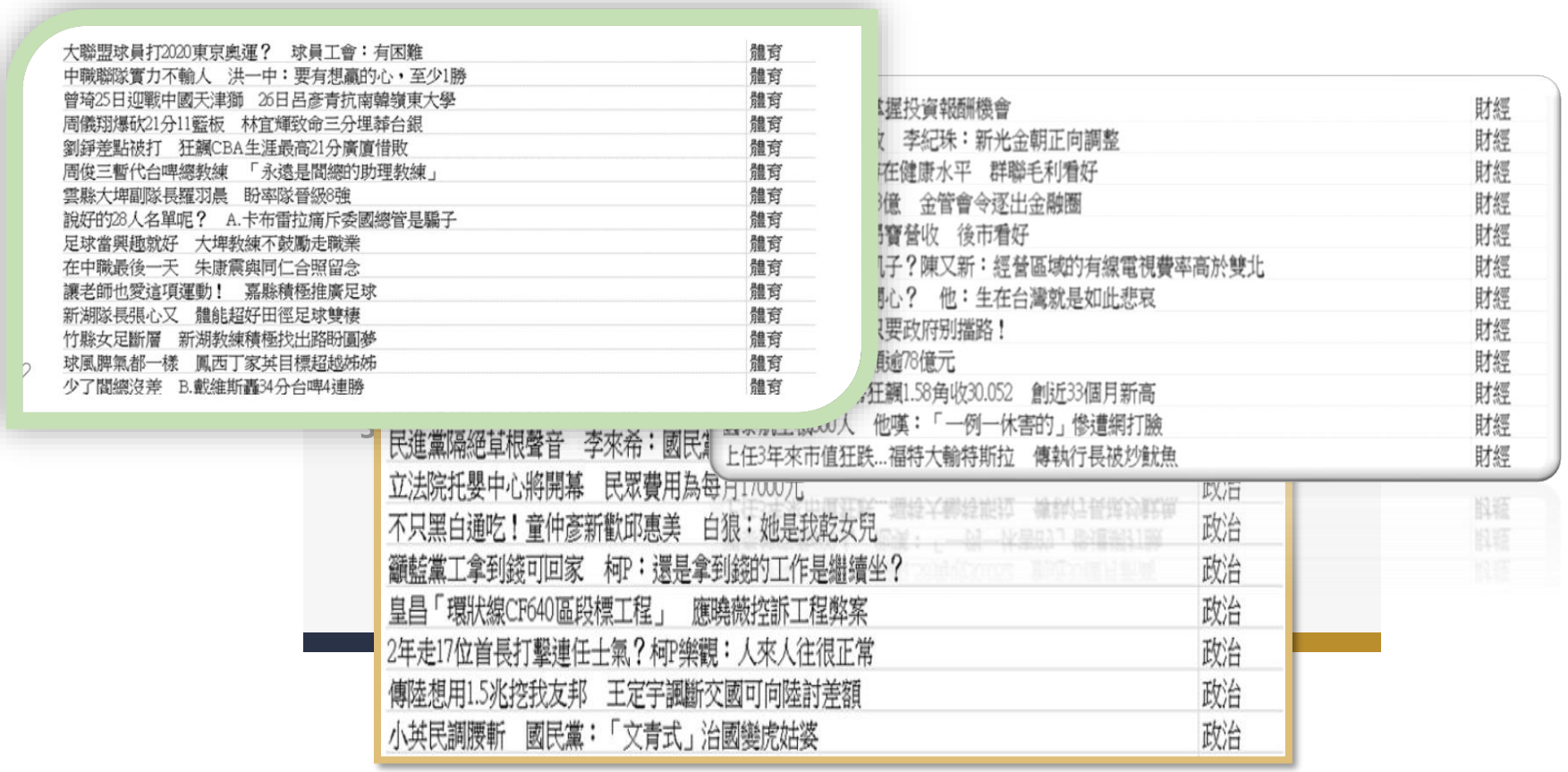
最終用4152筆訓練資料達到0.81的分數

對比丟入6613筆資料訓練可以達到0.81的分數

節省了37.2%標記量

04 資料二

資料二



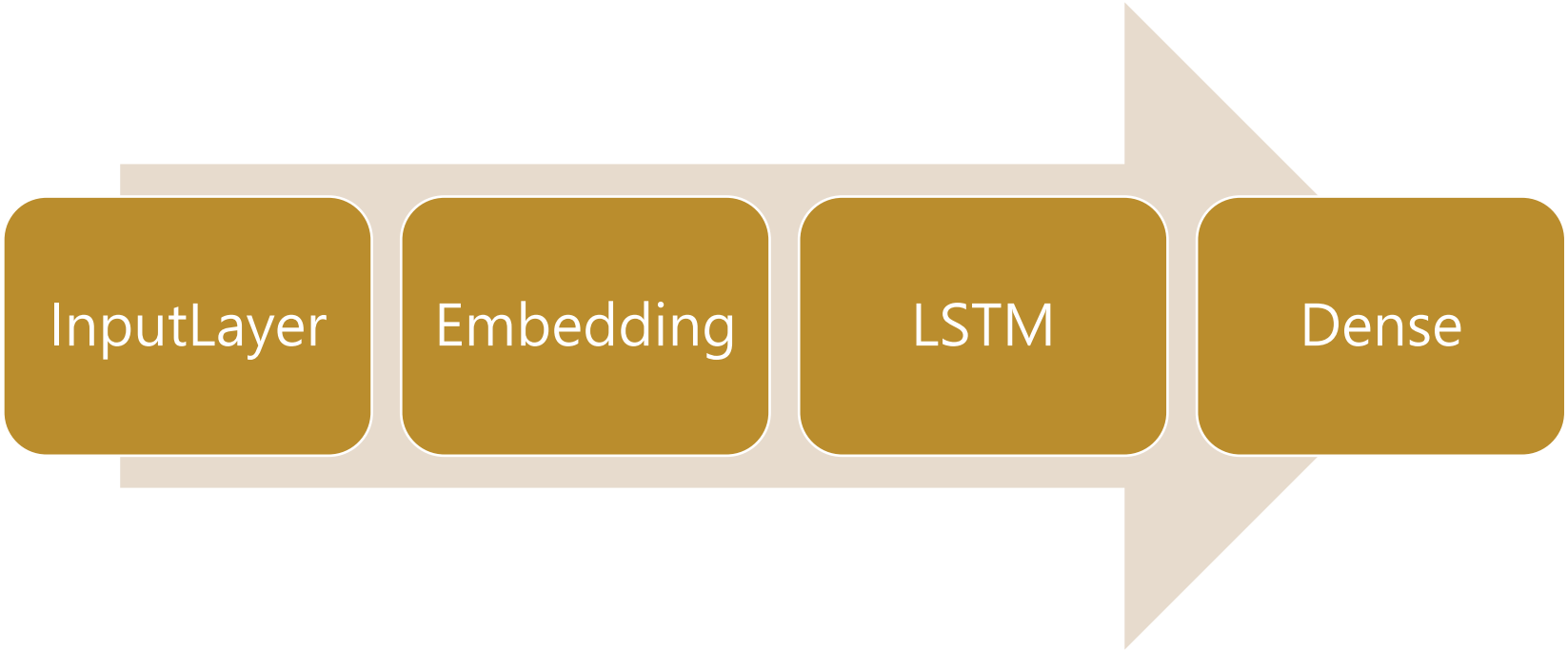
大聯盟球員打2020東京奧運？　球員工會：有困難	體育
中職聯隊實力不輸人　洪一中：要有想贏的心，至少1勝	體育
曾琦25日迎戰中國天津獅　26日呂彥青抗南韓嶺東大學	體育
周儀翔爆砍21分11籃板　林宜輝致命三分埋葬台銀	體育
劉錚差點被打　狂飆CBA生涯最高21分廣廈惜敗	體育
周俊三暫代台啤總教練　「永遠是閻總的助理教練」	體育
雲縣大埤副隊長羅羽晨　盼率隊晉級8強	體育
說好的28人名單呢？　A.卡布雷拉痛斥委國總管是騙子	體育
足球當興趣就好　大埤教練不鼓勵走職業	體育
在中職最後一天　朱康震與同仁合照留念	體育
讓老師也愛這項運動！　嘉縣積極推廣足球	體育
新湖隊長張心又　體能超好田徑足球雙棲	體育
竹縣女足斷層　新湖教練積極找出路盼圓夢	體育
球風脾氣都一樣　鳳西丁家英目標超越姊姊	體育
少了閻總沒差　B.戴維斯轟34分台啤4連勝	體育

握投資報酬機會	財經
李紀珠：新光金朝正向調整	財經
在健康水平　群聯毛利看好	財經
億　金管會令逐出金融圈	財經
寶營收　後市看好	財經
子？陳又新：經營區域的有線電視費率高於雙北	財經
心？　他：生在台灣就是如此悲哀	財經
要政府別擋路！	財經
逾78億元	財經
狂飆1.58角收30.052　創近33個月新高	財經
人　他嘆：「一例一休害的」慘遭網打臉	財經
上任3年來市值狂跌...福特大輸特斯拉　傳執行長被炒魷魚	財經

民進黨隔絕草根聲音　李來希：國民黨	
立法院托嬰中心將開幕　民眾費用為每月17000元	政治
不只黑白通吃！童仲彥新歡邱惠美　白狼：她是我乾女兒	政治
顧藍黨工拿到錢可回家　柯P：還是拿到錢的工作是繼續坐？	政治
皇昌「環狀線CF640區段標工程」　應曉薇控訴工程弊案	政治
2年走17位首長打擊連任士氣？柯P樂觀：人來人往很正常	政治
傳陸想用1.5兆挖我友邦　王定宇諷斷交國可向陸討差額	政治
小英民調腰斬　國民黨：「文青式」治國變虎姑婆	政治

05 模型一

模型一:Word embedding layer + LSTM layer (keras)



模型一 方法

01

前三次策略用margin sampling
選出最大機率減去第二大機率
分別小於1,小於3,小於5的

02

後面用entropy的值算出來介
於一個區間的，分別是
1.08~0.8和1.08~0.7

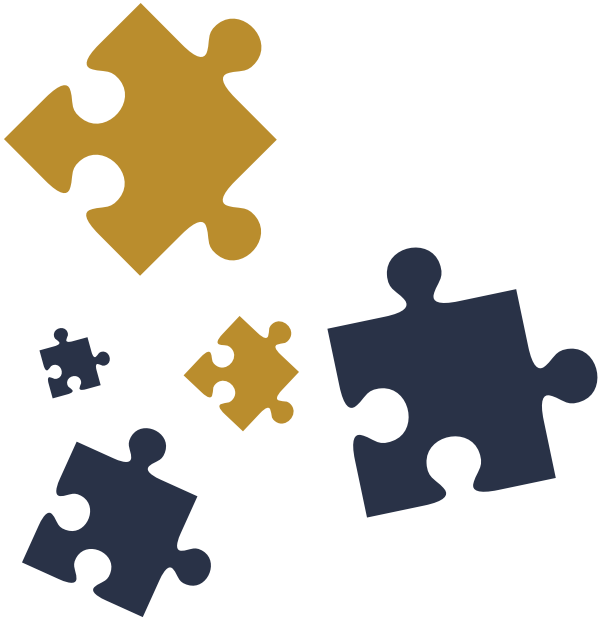
前半段	margin sampling
後半段	entropy

為啥設定小於1.08?
Ans:去除極端值

與隨機選取進行比較

丟進模型訓練的資料量	從模型輸出的結果去選取	Random選取同樣筆數
100	57.9%	57.9%
150	61.4%	62.8%
210	59.2%	65.6%
338	72.8%	69.4%
925	81.5%	79.3%
1139	83.0%	80.6%

模型一 結果



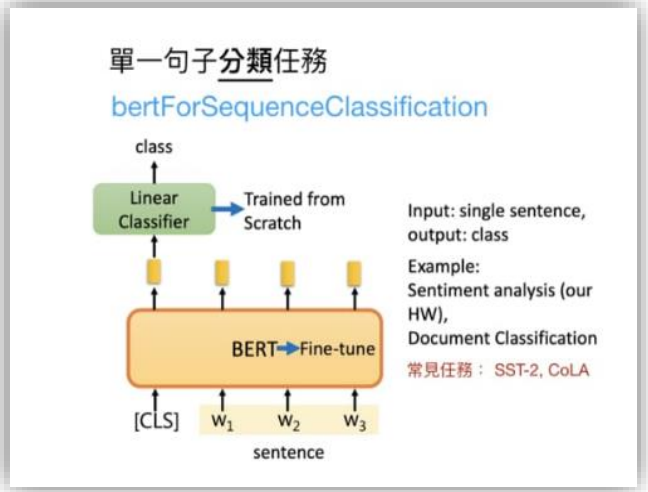
最終用1139筆訓練資料達到83.0%的準確率

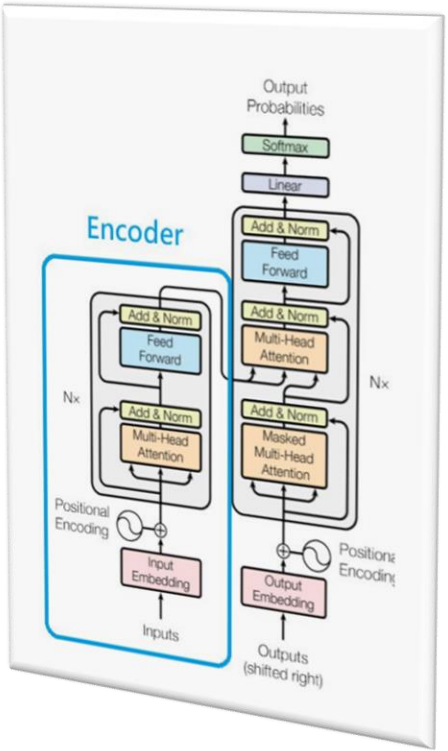
對比隨機丟入1600筆達到82.7%的準確率

節省了28.8%的標記量

06 模型二

模型二: bert (hugging face)





- Encoder of Transformer
- 大量文本以及兩個預訓練目標，事先訓練好一個可以套用到多個 NLP 任務的 BERT 模型，再以此為基礎 fine tune 多個下游任務。
- 預訓練模型: **bert-base-chinese**
- Fine-tuning階段 : **BertForSequenceClassification**
- Loss function: CrossEntropyLoss(), learning rate:1e-6

模型二 方法

01

先用100筆訓練資料去跑300個epochs。

02

利用模型預測各分類的機率，選出最大機率在某區間的資料加入training set，再給模型去跑80個epochs。

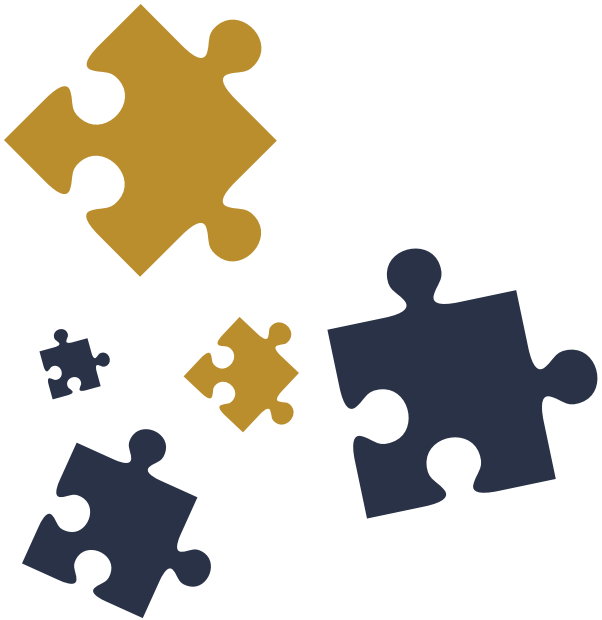
03

重複步驟二(每次改變取的機率區間)。

與隨機選取進行比較

丟入模型訓練的資料量	Select準確率	Random準確率	選取的機率區間
100	88.9%	88.9%	30%-75%
338	92.9%	91.8%	30%-86%
517	93.3%	92.3%	30%-90%
629	93.4%	92.1%	30%-93%
718	93.5%	93.0%	30%-95%
812	93.7%	93.2%	30%-97%
894	93.4%	93.1%	30%-98%
979	94%	93.3%	

模型二 結果



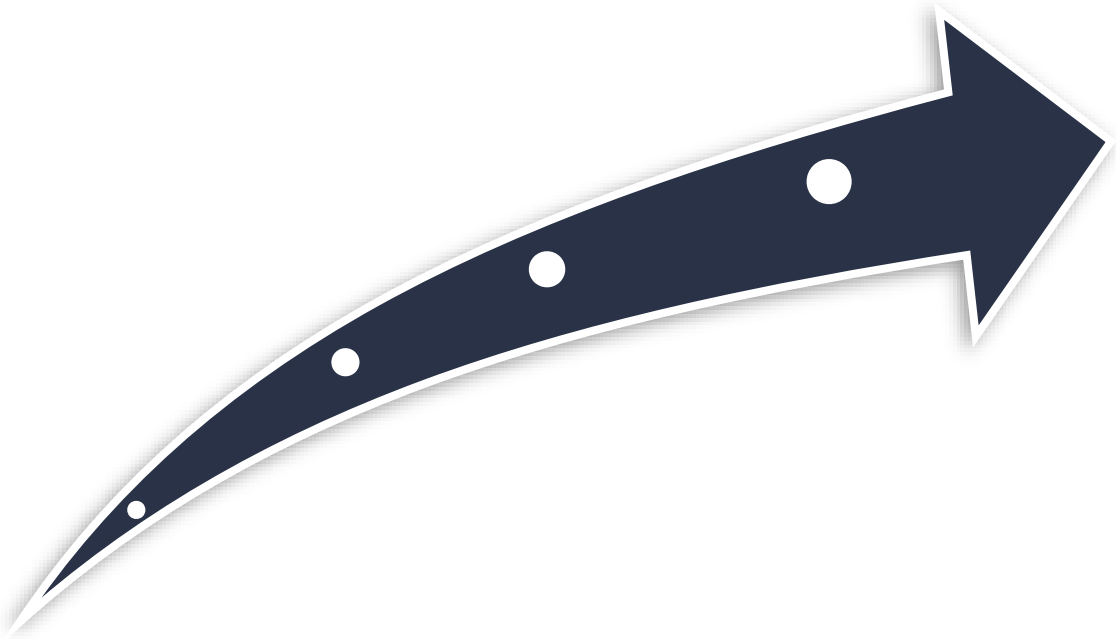
最終用979筆訓練資料達到94.0%的準確率

對比丟入2000筆資料訓練可以達到94.0%

節省了51%標記量

07 網頁呈現

網頁呈現



https://sites.google.com/view/nckuproject2022/%E9%A6%96%E9%A0%81

謝謝大家