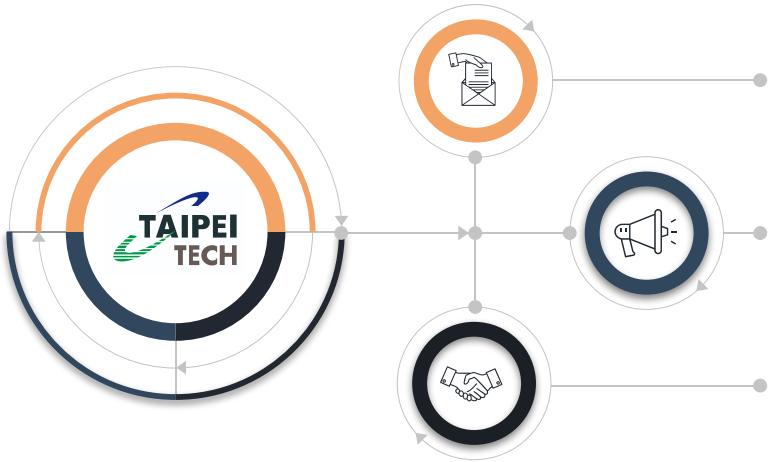
# LLM - Detect AI Generated Text

# Table of Contents



# 1 技術選題

# The origin of the topic "LLM - Detect AI Generated Text"




- **題目需求**: 偵測 AI 產生文字的論文。
- **解決方案**: 運用 NLP 技術之模型，進行偵測。
- **參考方案**:

Plan A. LLM Detect: Text Cluster [中文] ⇦ **主要選擇**

Plan B. LLM Detect AI Generated (Bert)

Plan C. Detect AI Generated Text Using BLSTM & DistilBERT

- **修改方案**:

Plan D. LLM Detect AI Generated Text (Optimization)

【LLM - Detect AI Generated Text】https://www.kaggle.com/competitions/llm-detect-ai-generated-text

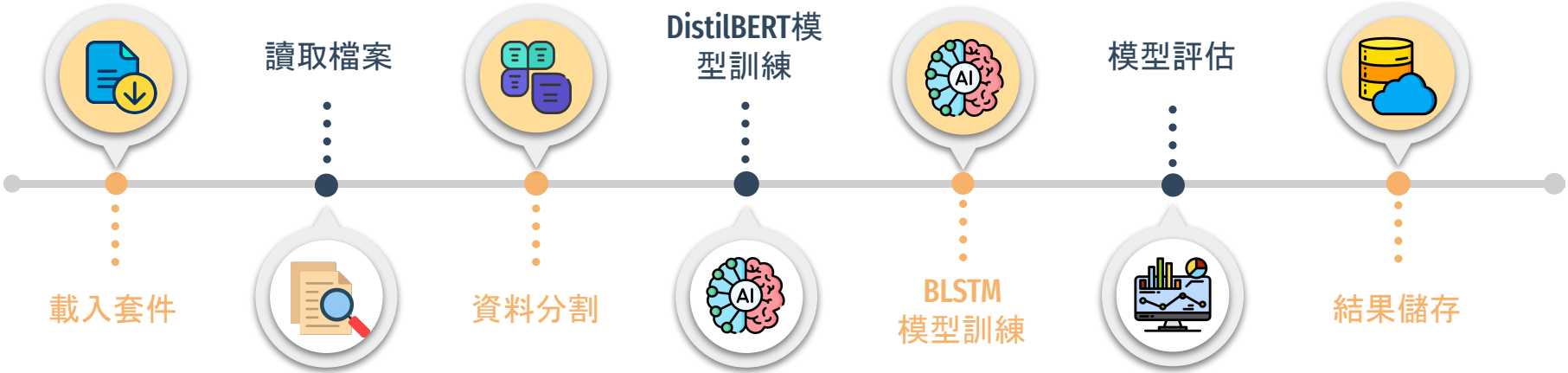
# A LLM Detect: Text Cluster [中文]


【LLM Detect: Text Cluster [中文] 】https://www.kaggle.com/code/finlay/llm-detect-text-cluster

# B LLM Detect AI Generated (Bert)

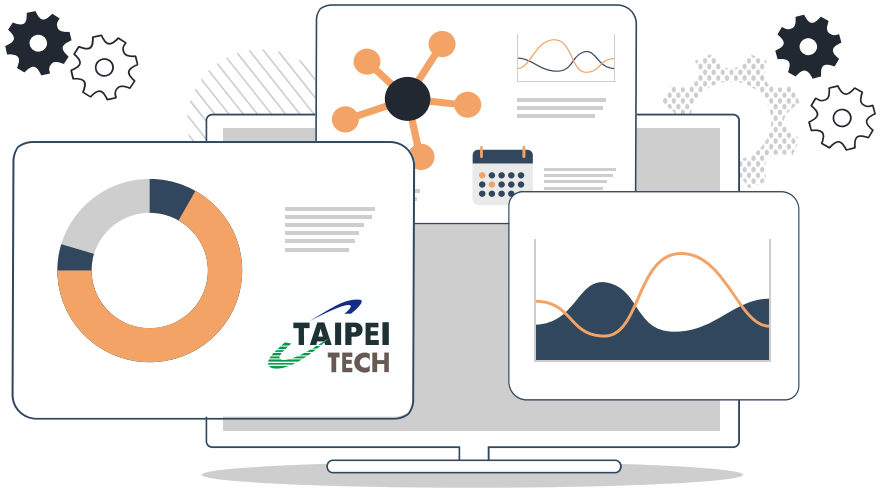

【LLM Detect AI Generated (Bert)】https://www.kaggle.com/code/shahbodsobhkhiz/detect-ai-generated-text-using-blstm-distilbert

# C Detect AI Generated Text Using BLSTM & DistilBERT

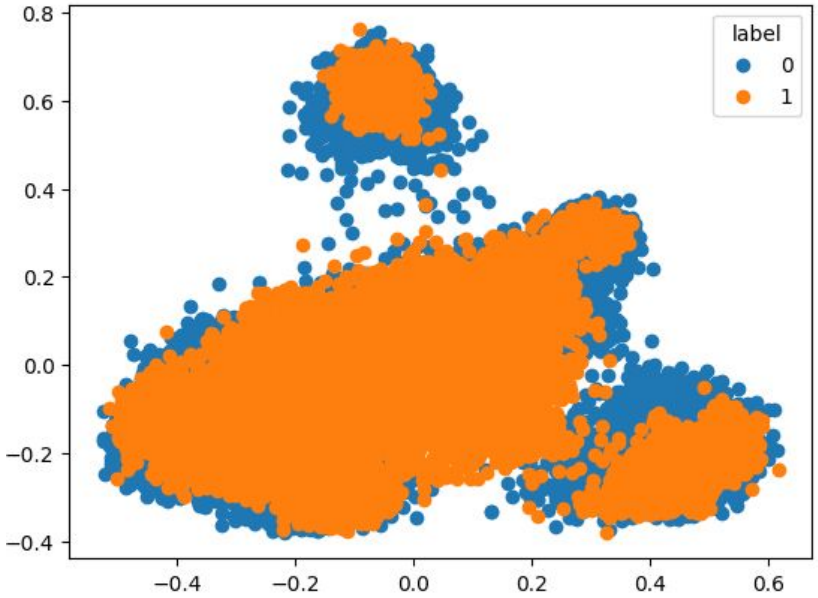



【Detect AI Generated Text Using BLSTM & Distilbert】https://www.kaggle.com/code/sunshine888888/llm-detect-ai-generated-bert

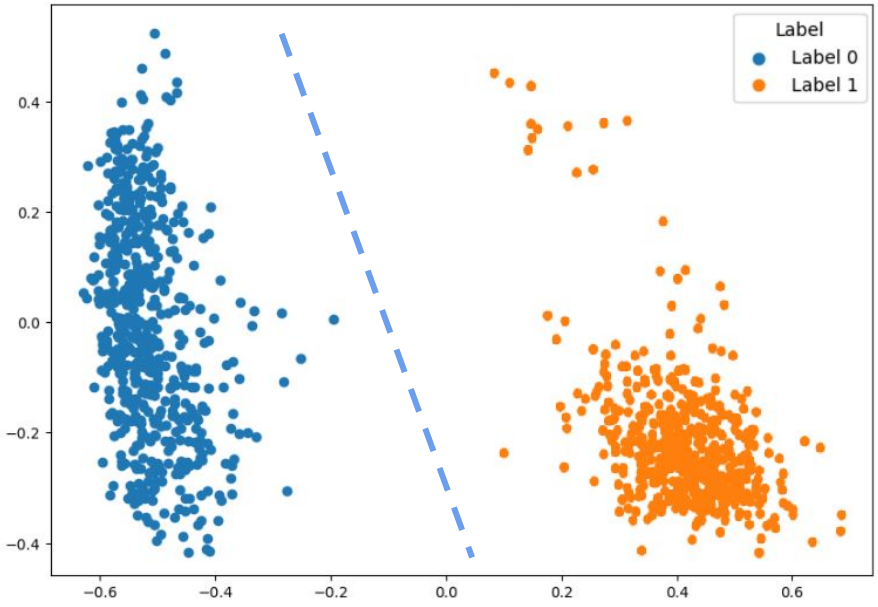
# 2 改良成果



# D PCA with data correction based on Plan A



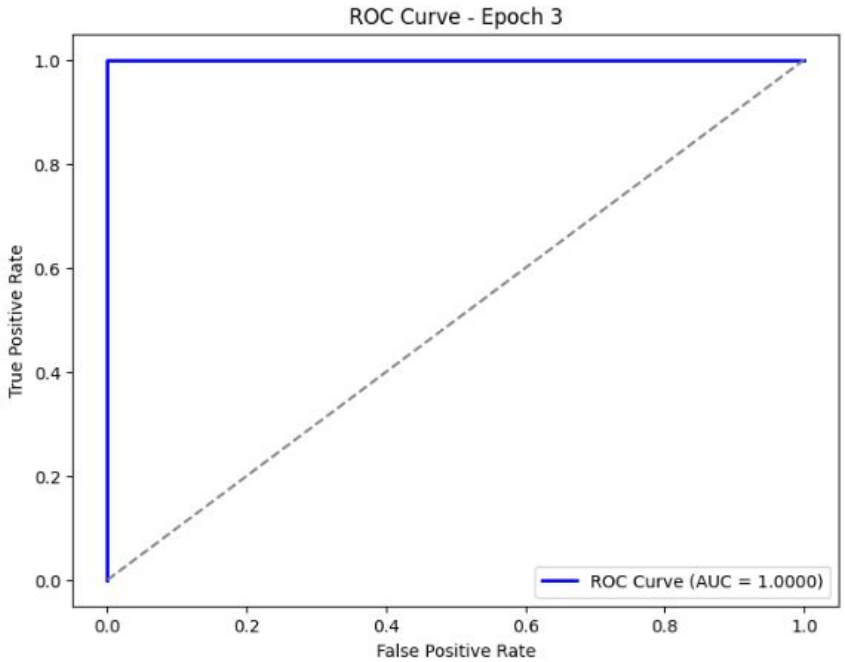

Plan A: PCA 訓練原始結果




Plan D: PCA 訓練改進結果

PCA (主成分分析) 是透過降維處理使得資料變異量降低的方法。

# D Data-corrected ROC based on Plan A




- **横軸 (False Positive Rate)**: 假陽性率
- **縦軸 (True Positive Rate)**: 真陽性率
- **ROC (Receiver Operating Characteristic)**: 1.00

**ROC 曲線** 是透過分類器進行效益與成本之相對關聯成數的分類法，結果越靠左邊及上面表達越好成效。


# 3 優勢解析

# B Evaluate the advantages of Plan B

## 較多的訓練層數

多工學習、注意力機制等，提高表現。

## SMOTE 類別與損失函數

處理不平衡資料集。


## 均分或注意力機制

使用特定資料集上訓練的微調嵌入或替代池化策略。

SMOTE (合成少數取樣技術) 是針對少數樣本位置相近處增加人工樣本的方法。

# C Evaluate the advantages of Plan C


**預訓練模型**

將 DistilBERT 和 BLSTM + Word2Vec 結合。

**大量資料取樣穩定**

組合多個資料集，並採用下取樣來解決類別不平衡問題。

**資料正則化處理**

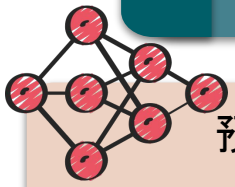
使用 Dropout 層來減少過度擬合。

**多模型整合**

基於 DistilBERT 與 BLSTM + Word2Vec，整合並實踐出良好的模型。

# 4 未來規劃

# D Advanced Architecture Integration: Plans B and C and D



## 預訓練模型

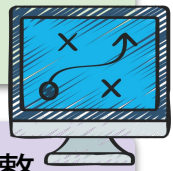
結合 DistilBERT 與 BLSTM + Word2Vec，以加強語義理解能力。

## 分詞處理機制修改

利用 HuggingFace 的 Tokenizer 實現精準的文字預處理。

## 資料正則化處理

透過引入 dropout 層，有效抑制過擬合，同時顯著提升模型的泛化性能。

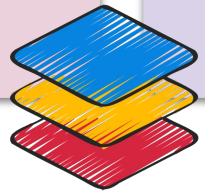


## 類別與損失函數調適

結合過取樣技術與加權隨機樣本選取器，有效提升少數類別的覆蓋率與模型表現。

## 資料集優化策略

整合多個資料集，以擴展數據多樣性，從而提升模型的泛化能力。



## 池化機制與策略統整

整合基於注意力的池化，以更好地捕捉序列的重要部分。


# Thank you