**Natural Language Processing**

**Final Project: LLM – Detect AI Generated Text**

**Team 81**

**113062624張瑋倫**

**113061536沈韋翔**

**113061595莊翔凱**

**113064534林士翔**

**1 Task Introduction**

隨著AI的快速進展，大型語言模型(Large Language Models, LLMs)已經可以根據指令，生成近似人類書寫的文章，但是這樣的技術也為教育或學術界帶來潛在的隱憂。學生若是過於依賴AI生成文章來完成工作，可能影響到學生的學習與反思能力；學術界也面臨著如何確保研究成果真實性的挑戰。在這樣的背景下，開發可靠的AI生成文本檢測工具變得越來越重要。

此次的project任務旨在針對上述問題，提出一套能有效識別AI生成文章的模型，並透過一些方法的優化以提高分識別的準確度與可靠性。

**2 Dataset Overview**

這次kaggle任務中所提供的dataset為train\_essays.csv，說明如下：共有1378篇文章，其中人類撰寫為1375篇，AI生成為3篇。文章是根據prompt\_id所對應的主題寫成：prompt\_id = 0：Car-free cities，共有708筆；prompt\_id = 1：Does the electoral college work? 共有670筆。並且由Generated標籤表示該文章是由學生撰寫(0)或是由AI生成(1)。

|  |  |  |  |
| --- | --- | --- | --- |
| ID | Prompt\_id | Text | Generated |
| 0059830c | 0 | Cars. Cars have been around since they became famous in the 1900s, when Henry Ford created and built... | 0 |
| 011dc2bc | 1 | Dear state senator, It is the utmost respect that I ask for the method for presidential election be ... | 0 |

由上述可見，此次任務提供的training data十分不平衡，由於Human-written的樣本數量遠超過AI-generated，模型可能傾向於學習多數類別(Human-written)的特徵。因此在開始訓練模型前，需要先解決這個問題。

**3 Methodology**

我們的目標是使用LLM進行AI文本的偵測，因此在瀏覽kaggle上的code後，列出了一些參考做法。

**3.0 Replace training set**

資料集對於模型訓練至關重要，然而上述的標籤不平衡問題，需要透過資料加強等方式解決。然而在論壇上的諸多方法均直接替換訓練資料，尤以train\_v2\_drcat\_02.csv被廣為使用。

該訓練資料集多達近45000筆資料，其中人類撰寫和AI生成幾乎佔各半，適合不須平衡數據直接拿來使用。在往後我們研究自己的新方法時，使用替代訓練資料集也被我們納為重要步驟。

**3.1 Fine-tune DeBERTa as a classifier**

此節描述如何使用DeBERTa-v3-large模型進行微調。首先，只有使用競賽所提供的training set是不夠的，因此在訓練上另外使用了kaggle上額外的文本資料集：SlimPajama [1] 和Persuade 2.0 [2] 的文章，前者提供優秀的AI生成文章，後者則是由人類撰寫的文本。

在前處理過程中，移除了缺失行及重複文本，接著透過AutoTokenizer和AutoModelForSequenceClassification載入DeBERTa-v3-large的tokenizer和model，並預設binary classification(區別人類或AI文章)。一些超參數設置為：MAX\_LEN = 1024，learning rate = 2e-5，batch\_size = 4，grad\_accum\_steps = 8，epochs = 1，並且將dataset的95%設為training set，剩下5%為testing set。關於模型評估的部分，首先通過softmax層後，將preds結合了多種metrics來計算訓練結果，包括AUC-ROC、Accuracy、Precision、Recall以及F1 score。

**3.2 Ensemble Method**

在此介紹Kaggle上3rd method的一些技巧。該方法的核心理念即為Ensemble learning，已有許多研究和多個競賽表現證實使用集成學習的方式效果優異，並可大幅提升泛化效果。

3rd method使用12個deberta v3-large的預訓練模型以及4個使用TF-IDF做word embedding來訓練的模型。這邊特別提到該方法使用pseudo labeling技術，以其中4個deberta v3-large預訓練模型先行預測測試集並將高確信的結果重新加入訓練資料集做訓練，擴大數據集規模，從而提高模型的泛化能力。

至於為何額外加入4個使用TF-IDF來做預測的模型加入集成學習，原因在於有人研究出目前最成功的方法是結合基本的 tf-idf ngram 與 sklearn 中的強大模型。並結合了簡單性、高效性和穩定性的特點，能夠在大多數文本挖掘和自然語言處理任務中快速取得可靠的結果。

集成學習是要達到良好效果不可避免的方法，也是我們在研究新方法中關鍵的目標。

**3.3 Our Method**

**[Our method (combines the ideas from above)]**

**4 Experiments and Results**

**[Present all the experiment results (the scores) with brief explanation]**

**4.1 Deberta-v3-large Pipeline**

Score: 0.965134

**4.2 Ensemble Pipeline**

|  |  |
| --- | --- |
| **MNB** | 0.824018 |
| **SGD** | 0.805608 |
| **LGB** | 0.873032 |
| **MNB+SGD** | 0.823884 |
| **MNB+LGB** | 0.877878 |
| **SGD+LGB** | 0.868606 |
| **ALL(MNB+SGD+LGB)** | 0.872401 |

**4.3 Hybrid Method**

Score: 0.967799 (30%-70%)

[undergoing work: other thresholds]

**5 Discussion**

**[TBD]**

**6 Conclusions**

**[TBD]**

**7 References**

[1] <https://www.kaggle.com/datasets/tailen/slimpajama-ai-generated-parallel-dataset>

[2] <https://www.kaggle.com/datasets/tailen/persuade-corpus-ai-generated-dataset>