# 基於 DistilBERT 和 Group Query Attention 之惡意言論分類研究與應用

專題編號：113-CSIE-S024
執行期限：112 年第 1 學期至 113 年第 1 學期
指導教授：王正豪
專題參與人員： 110820038 史學文
110820044 黃軒和
110820034 林芳成

## 一、摘要

本專題旨在利用深度學習技術建立一個能針對不同類型之惡意言論進行偵測與分類的模型。本研究以 Transformer 架構為基礎，使用自建的模型架構。透過 PTT 網路爬蟲搜集資料，結合 Google ShieldGemma 標注 4 種類別的惡意言論，並透過知識蒸餾 (Knowledge Distillation) 技術訓練自製模型。另外，我們也製作了 Discord (社群通訊軟體) 機器人，能刪除使用者發送的惡意言論，並提示觸犯了哪個類別。
**關鍵詞：**惡意言論偵測、多標籤分類、深度學習、自然語言處理

## 二、緣由與目的

隨著社群媒體與網路平台的興起，惡意言論在網路上廣泛傳播，傳統的偵測方法（如關鍵字過濾）難以精確辨識隱含或複雜的攻擊性語句。

本研究旨在利用深度學習模型有效偵測和分類網路上的惡意言論，提升網路社群的健康環境。

## 三、研究範圍

(一) **模型開發：**建立基於 Transformer 架構的惡意言論分類模型。

(二) **資料收集與標註：**透過 PTT 網路爬蟲收集資料，使用 Google ShieldGemma 標註四種類別的惡意言論（危險內容、騷擾、仇恨言論、色情內容）。

(三) **應用實作：**開發一個 Discord 機器人，實現即時偵測與處理惡意言論。

## 四、使用技術方法

(一) **知識蒸餾 (Knowledge Distillation)：**

ShieldGemma 2B 標注不同種類惡意言論為 Positive 和 Negative 的機率。

(二) **資料取得：**

用 Python 取得文章以及留言內容，並使用正則表達式 (Regular Expression) 對資料進行清理。

(三) **模型連接方式：**

實驗 Transformer 架構不同 Encoder-Decoder 連接方式，取表現最佳者。

## 五、架構流程

(一) **資料集準備:**

使用網路爬蟲由 PTT 爬取資料，再用 ShieldGemma 標記該文本帶有惡意言論的機率，最

後對資料集做資料平衡。

(二) **模型設計：**
構建以 DistilBERT 為 Encoder，Group Query Attention 為 Decoder 的模型架構。

(三) **訓練模型：**
透過知識蒸餾技術訓練模型。

(四) **部署至 Discord 機器人**

## 六、工具說明

(一) **PyTorch**：進行模型的構建、訓練與評估。

(二) **Discord.py**：開發 Discord 機器人框架。

(三) **ShieldGemma**：大型語言模型，用於自動化資料標註。

(四) **網路爬蟲工具**：Requests、BeautifulSoup，用於資料收集與解析。

## 七、實驗結果

(一) **模型性能**：在使用僅 1/10 參數量的情況下，模型在「危險內容」和「色情內容」兩類的 F1 分數分別達到 0.79 和 0.80。

(二) **分類效果**：模型在「騷擾」和「仇恨言論」類別上的表現有待提升，可能受限於資料量或類別定義的複雜性。

(三) **應用成果**：開發的 Discord 機器人能即時偵測並刪除惡意言論，並提示使用者違規的類別，可設定各類別的靈敏度。

## 八、結論

(一) 本研究成功地開發了一個基於 DistilBERT 和 Group Query Attention 的惡意言論分類模型，我們在減少模型參數的情況下，部分類別上達到與大型模型相近的效果。

(二) Dangerous 和 Sexually 類別的分類效果表現較為穩定且準確。然而，在 Harassment 和 Hate Speech 類別上的分類效果相對不如預期，這部分可能是資料數量的不足或是定義過於模糊導致。

(三) 我們開發了基於該模型的 Discord 機器人，實現惡意言論的即時偵測與處理，展示了該技術的實際應用潛力。

## 參考文獻

[1] Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” arxiv, 1910.01108, pp.1-5, 1 Mar 2020.

[2] Joshua Ainslie, James Lee-Thorp*, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, Sumit Sanghai, “GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints,” arxiv, 2305.13245, pp. 1-7, 23 Dec 2023.

[3] Fenglin Liu, Xuancheng Ren, Guangxiang Zhao, Chenyu You, Xuewei Ma, Xian Wu, Xu Sun, “Rethinking and Improving Natural Language Generation with Layer-Wise Multi-View Decoding,” arxiv, 2005.08081, pp.1-25, 29 Aug 2022.