中華民國第 65 屆中小學科學展覽會

作品說明書

高級中等學校組　電腦與資訊學科

052514

AI 與心理學的言語柔化實驗

學校名稱： 國立鳳新高級中學

作者：	指導老師：
高二 洪欣郁 高二 王語岑 高二 黃靜雯	林冠曄

關鍵詞： 自然語言處理、情感分析、深度學習

摘要

現今網路充斥著具壓力的負面言論，影響使用者的心理健康。本研究旨在參考 C-ME 量表並整理為四大類標籤，人工分類從論壇蒐集的言論來訓練模型。研究採 LSTM、雙向 LSTM、CNN + LSTM 及 Transformer 四種深度學習模型，基於自建資料集進行訓練，實現對全新言論的精準分類，並比較四種模型在精確率、召回率及 F1-score 等指標上的表現。結果顯示，LSTM 處理數據不均衡的資料集時表現最佳，F1-score 達 89.2%。實測發現，CNN+LSTM 在預測效果上略勝 LSTM。此外，結合生成式 AI GPT-4o-mini，能有效改善不當言論，為留言者提供更委婉的表達建議。

壹、前言

一、研究動機

隨著社群媒體蓬勃發展，人們能夠輕鬆地在網路上發表想法。然而，這種便利性也衍生出一些問題，例如部分留言者可能發表過激或侮辱性言論，使他人感到不適、承受壓力，甚至造成心理傷害。儘管許多知名論壇應用程式已設有言論篩審機制，但由於篩審可能過於寬鬆，不當言論仍然屢見不鮮。為了改善此現象，本研究旨在強化不當言論的篩審機制，並提供適當的修改建議，以營造更友善的網路環境。

二、研究目的

（一）自製言論資料集，將言論標上攻擊性言論、歧視性言論、性相關不當言論和普通言論四大類標籤。

（二）運用訓練後的 LSTM、雙向 LSTM、CNN + LSTM、Transformer 模型判斷新言論屬於何種類別。

（三）將不當言論依嚴重程度再細分為是否禁止傳出。

（四）連接 ChatGPT-4o-mini 將 不當言論改為較委婉的用詞。

三、文獻回顧

（一）模型介紹

1. LSTM（Long Short-Term Memory，長短期記憶網路）：

LSTM 是一種循環神經網路（Recurrent Neural Network, RNN）[2]，透過記憶單元（Cell State）來保留序列中的長期特徵。為 RNN 的改良版，LSTM 解決了傳統 RNN 在處理長序列時的短期記憶問題，以及梯度消失或梯度爆炸導致的學習困難。為了有效選擇保留或遺忘的資訊，LSTM 設計了「遺忘門」、「輸入門」與「輸出門」，其中遺忘門負責篩選不必要的資訊，確保模型保留關鍵內容，使結果更準確。圖 1-1 為 LSTM 門控結構圖。

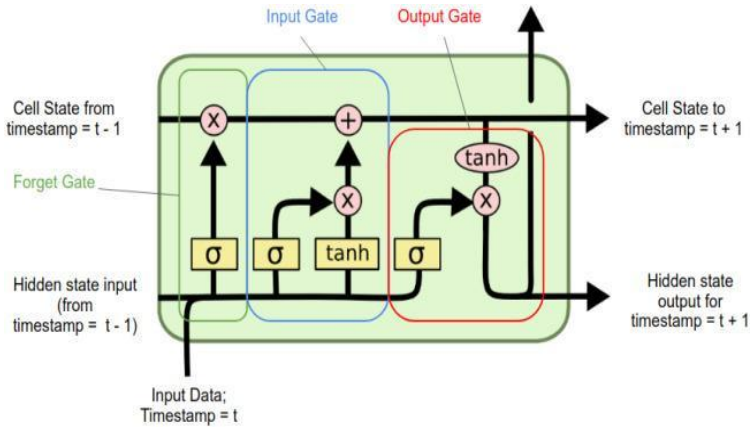


圖 1-1： LSTM 門控結構圖

（資料來源：Varsamopoulos et al., 2018）

2. 雙向 LSTM（BiLSTM）：

單向 LSTM 只考慮正向的序列，而雙向 LSTM（BiLSTM）則同時從正向與逆向處理序列資訊，使模型能夠提取更完整的上下文特徵，減少關鍵資訊的遺失。因此，雙向 LSTM 特別適合長文本分析。圖 1-2 為雙向 LSTM 結構圖。

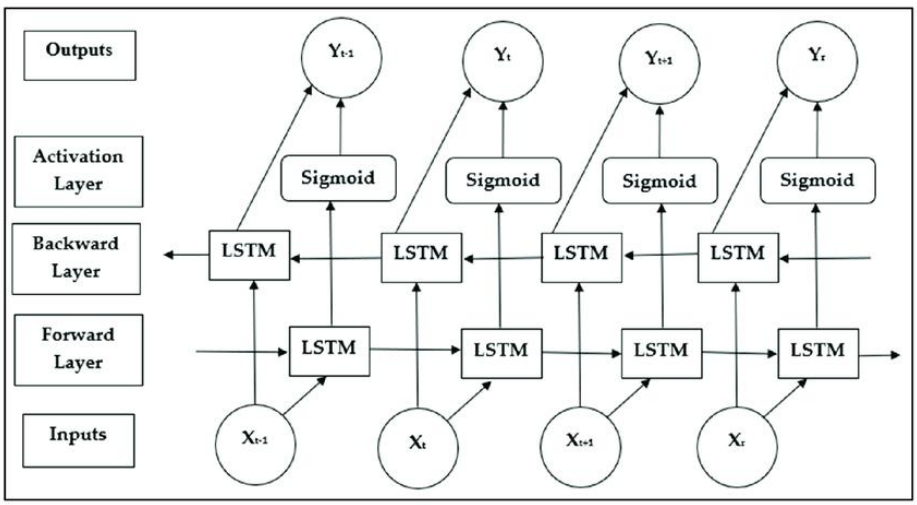


圖 1-2：雙向 LSTM 結構圖

（資料來源：Nawaf Mohammad Alamri et al., 2023）

3. CNN + LSTM

   CNN 中的 Convolutional Layer 負責提取關鍵特徵 [3]，而 Pooling Layer 則降低維度，保留重要資訊，同時減少過擬合（Overfitting ）的風險，適用於言論數較少的資料集。LSTM 則透過學習時間序列關係來分析過去言論的順序，以預測接下來的結果。由於 CNN 主要應用於圖像分析，因此在文本分析時，需先將文字向量化，再利用 CNN 提取特徵，並交由 LSTM 進行序列分析，以提升模型的準確性。

4. Transformer：

   Transformer 是由 Google 在 2017 年提出的模型（ Seq2seq ） [4] ， 主要由兩個部分組成，分別為 Encoder 及 Decoder 兩部分。其中 encoder 的部分常使用於情感分析，為本研究所使用，它能夠提取輸入序列中的深層語意。使用的 Multi-Head Self-Attention 可以使得序列中的元素間建立關係，掌握整句的關聯，前饋神經網路（ FFN , Feed-Forward Neural Network） 由兩層全連接層（Dense 層）組成，透過非線性變換提取局部特徵，並獨立作用於每個 Token，以學習局部特徵。 使用殘差連接（ Residual Connection ） 和 Layer Normalization 防止梯度消失以及正規化保持輸出穩定，圖 1-3 為 Transformer Encoder。

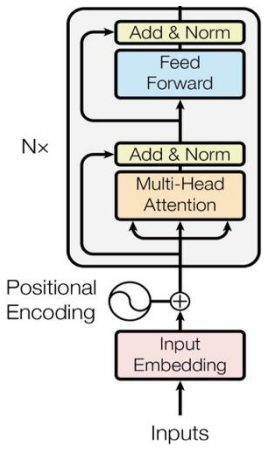


圖 1-3： Transformer Encoder

(資料來源：Vaswani, et. al., 2017: 3)

5. 模型比較

模型種類	序列處理能力	對於長序列是否有好的處理能力	特色及其他
RNN（本研究未使用）	是	否	按序計算，初代處理序列的模型，有梯度消失的問題
CNN	否（提取特徵用）	否	卷積運算，每個卷積核獨立計算
LSTM	是	是（記憶單元+門控機制）	按序計算，RNN 改良版為同一體系
雙向 LSTM	是	是（承襲 LSTM+判斷前後語義）	按序計算，RNN、LSTM 改良版為同一體系
Transformer	是	是（Self-Attention 機制）	與 RNN 系列完全不同的機制，可並行計算，最先進的方式，唯算力較高

表 1-1：模型比較（研究者自製）

（二）混淆矩陣

混淆矩陣（Confusion Matrix）在機器學習中用於評估模型判斷分類能力的工具。它以矩陣形式呈現模型的預測和實際結果比較的狀況，可以更好的判別模型在哪個類別的判斷能力表現良好，以及哪些類別的判斷存在偏差。

1. 本研究中使用多標籤分類

類型	言論可以有幾個標籤	矩陣的形式
多類別分類	只能一個	N×N（N 表類別數）
多標籤分類	兩個以上	每個標籤都對應一個 2×2 矩陣

表 1-2：多類別分類和多標籤分類比較

2. 精確率、召回率、F1-score 的表示

（1）矩陣通常結構表示：

		實際	
		Positive	Negative
預測	Positive	TP （True Postive）	FP （False Postive）
	Negative	FN （False Negative）	TN （True Negative ）

表 1-3：矩陣通常結構表示

（2）精確率（Precision）：所有被判斷為正向的言論中，實際為正向的比例。

$$\text{Precision} = \frac{TP}{TP + FP}$$

（3）召回率（Recall）：實際為正向的言論中，被正確判斷為正向的比例。

$$\text{Recall} = \frac{TP}{TP + FN}$$

（4） F1-score：精確率和召回率的調和平均

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

本研究使用多標籤判斷模型，標籤共有四種（攻擊性言論、歧視性言論、性相關不當言論、普通言論），且每個標籤獨立判斷，因此各標籤各自有一個混淆矩陣，表 1-4 以攻擊性言論舉例。

		實際	
		攻擊性言論	非攻擊性言論
預測	攻擊性言語	實際：攻擊性言論 預測：攻擊性言論 （TP，預測正確）	實際：非攻擊性言論 預測：攻擊性言論 （FP，預測錯誤）
	非攻擊性言語	實際：攻擊性言論 預測：非攻擊性言論 （FN，預測錯誤）	實際：非攻擊性言論 預測：非攻擊性言論 （TN，預測正確）

表 1-4： 攻擊性言論的混淆矩陣

（三） C-ME 量表 [1]

內容導向媒體接觸量表（Content-based Media Exposure Scale, C-ME），是一種標準化工具，用於測量個體接觸特定媒體內容的頻率，特別適用於青少年群體。

1. 題目介紹：

   C-ME 量表包含 17 個題目，其中 8 題衡量個體對「反社會」媒體內容的接觸，如暴力、毒品使用、性行為、酒精濫用、偷竊等；9 題作為「中性」媒體內容的填充項目，例如新聞、旅遊節目、烹飪節目等。

2. 研究發現：

研究結果顯示， C-ME 量表具有良好的信度和效度，特別是對反社會內容的測量在不同樣本中表現一致。研究發現接觸反社會媒體內容與個人特質（如感官尋求、攻擊性等）呈正相關。

3. C-ME 量表優勢：

（1）相較於傳統的媒體使用測量量表， C-ME 量表更能準確評估內容類型對行為與心理的影響。

（2）可適用於不同媒介（如社交媒體、YouTube、電視、遊戲等），反映當代媒體環境的多樣性。

貳、研究設備與器材

一、自製言論資料集：言論總數 3708 則，各類別標籤總數 4046 個，表 2-1 為各類別標籤數量及言論、標籤總數，圖 2-1 為自製的言論資料集

攻擊性言論(個)	1154
歧視性言論(個)	902
性相關不當言論(個)	604
普通言論(個)	1386
言論總數(則)	3708
標籤總數(個)	4046

表 2-1：各類別標籤數量及言論、標籤總數

ID	Text	Main Label
0	理組男是不是真的比較噁	攻擊性言語,歧視性言論
1	看妳姊這樣讓人同情不起來啦，啊妳除了發文還會做什麼嗎？是我就 啊如果你們狠不下心來，那也沒資格怪妳姊，因為你們都是一樣的	攻擊性言語
2	臺灣司法真的很低能... 我是建議法匠們一人帶個幾隻罪犯回去教化	攻擊性言語
3	一堆爛鳥還敢說法官沒有支持廢死 殺了8個親屬的人 法官居然還覺得可教化！？ 只因為他想拿警察的槍自盡 笑死了 他怎麼不把自己燒死呢？ 這樣比較有誠意吧？	攻擊性言語
4	他爸還說要申請國賠，可見垃圾二輪騎法的賤基因是會遺傳的 賠你媽個沒啦，騎這樣出人命剛好而已	攻擊性言語
5	他爸當年要是17歲先慘死，今天就不會有這種喜劇~爽喇	攻擊性言語
6	傻逼 要讓人家出糗之前 自己先多學一點吧 不然我怎麼看你就只是看人家有錢又是猴子所以眼紅	攻擊性言語
7	又一個閒閒沒事 乾你屁事 有病要醫	攻擊性言語
8	"那妳有沒有想過自己為什麼三不五時都要聊性? 我跟男友同居，男友需求大每天做，或許就是1樓所謂的能力值高的 但他平時並不會把性掛在嘴邊，都還是以時事生活為主 今天如果妳換成是男生，底下一定一片撻伐說妳是噁男公狗 要不要想想看是誰讓自己這麼不堪?"	攻擊性言語,歧視性言論
9	"都讀到大學了怎麼行為跟國小生一樣 腦子沒跟著一起長大嗎"	攻擊性言語
10	照片這種超醜，喜歡這種的要馬壓力太大要馬有病	攻擊性言語

圖 2-1：自製的言論資料集

（本圖為研究者自製 ）

二、設備：筆記型電腦。

三、軟體：

Google Colaboratory：一個基於雲端的 Python 開發環境，讓使用者可以在瀏覽器中編寫與運行 Python 程式，支援常見的機器學習與數據分析套件，無需額外下載。如：TensorFlow、Keras、Pandas、 NumPy、Scikit-learn 等。本研究以此作為開發環境，其硬體配置：CPU：Intel(R) Xeon(R) CPU @ 2.20GHz， RAM：12GB。

Google Spreadsheets：由 Google 推出的一個電子試算表程式，使用者可在瀏覽器中建立及編輯檔案，可多人共用文件，並自動儲存於雲端。本研究用於整理與標記言論資料集。

四、工具：

ChatGPT(GPT-4o-mini)：OpenAI 開發的一個基於 GPT 架構的聊天機器人。本研究用來轉換言論資料集中的不當言論，以及生成普通言論擴充資料集。

Gemini：Google AI 開發的一個大型語言模型。本研究用來產生資料集中攻擊性言論、歧視性言論、性相關不當言論的變體，以增加言論變體擴增資料集。

五、Python 套件：

套件	介紹	本研究運用
jieba	中文斷詞工具	將文本進行斷詞
pandas	Python 的數據分析模組，提供 DataFrame 資料結構，讓使用者能夠快速操作及分析資料，強化資料處理的方便性	使用 pandas 讀取資料集並進行資料清理，透過 DataFrame 的方式存取文本數據和標籤
TensorFlow、Keras	為深度學習框架，Keras 建立在 TensorFlow 之上，提供更簡潔的介面。目前兩者已整合	訓練自然語言處理模型，進行文本情感分析
Scikit-learn(sklearn)	常用的開源機器學習庫，功能有分類、回歸、分群、降低維度、模型選擇、資料前處理	進行數據處理與模型評估
NumPy	一種數值計算程式庫，支援多維陣列和陣列運算	將文本向量化

表 2-2：本研究所使用的 Python 套件與應用

參、研究過程與方法

一、研究架構

本研究分為兩部分，第一部分為言論類別選擇及自製資料集，先查詢相關文獻，運用現有 C-ME 量表並對其進行分析，根據量表問題再延伸選出資料集的言論類別。接著在各大社群平台蒐集言論，人工標記言論對應的標籤，使用 Gemini 生成變體進行數據擴增，最後進行斷詞、去除標點符號、向量化等資料前處理，完成自製資料集。第二部分為程式實作，使用在第一部分完成的資料集，訓練四種情感分析模型，

並分析比較四種模型的精確率、召回率及 F1-score，接著連接 ChatGPT(GPT-4o-mini) 轉換言論及測試，最後與情感分析合併測試，圖 3-1 為研究流程圖。

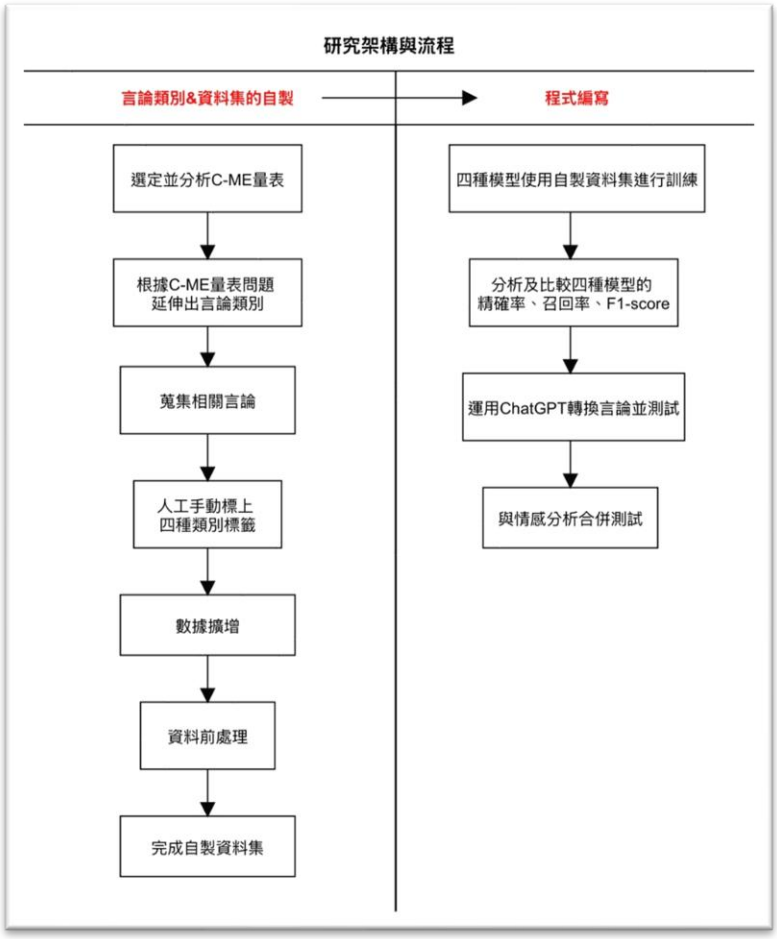


圖 3-1：研究流程圖

（本圖為研究者自製 ）

二、心理學分析

（一）問卷設計：

問卷分為兩部分：

1.第一部分為 C-ME 量表原始題目，根據受試者的使用習慣，讓受試者在 1 分（從不）至 5 分（經常）進行評分。例如：你是否會經常觀看到網絡中有打鬥的內容？

2. 第二部分根據 C-ME 量表原始題目，從其中延伸出表單問題所需言論類別，各類別有兩則代表言論，讓受試者以 1（無壓力）至 5（極高壓力）進行評分。表 3-1 為言論評分示意表。

言論類別	言論	感受壓力值(1-5)
娛樂與節目	知道為什麼台灣影視圈會頹廢嗎？就是有這種沒救的製作單位，台灣節目沒人想看再也正常不過了。	1 分（無壓力） 2 分 3 分 4 分 5 分（極高壓力）
肢體暴力	剝皮灌水銀、鞭刑才是正確的方法	1 分（無壓力） 2 分 3 分 4 分 5 分（極高壓力）

表 3-1：言論評分示意表

（二）研究流程：

1.尋找壓力量表：

（1）本研究參考並比較了四個量表：C-ME 量表、IAT 量表[7]、網路霸凌和網路攻擊加害者辨識量表[8]、網路霸凌和網路攻擊受害者辨識量表[8]，經綜合考量其功能性後，最終選定符合本研究需求的 C-ME 量表、作為測量工具。

（2）分析現有的 C-ME 量表。

2.製作問卷調查：

（1）將 C-ME 量表中的問題整合至問卷表單。

（2）刪減、分類，形成表單題目所需言論類別。

（3）針對每個言論類別，查找兩則代表性的網路言論。

（4）受試者依據每則言論可能引起的壓力程度（即在接觸到該言論時，內心感到的緊張、不適、煩躁或心理負擔程度），以 1（無壓力）至 5（極高壓力）進行評分。

3.問卷結果分析：

根據各言論類別的評分，統計分析其壓力平均值分佈，從而了解不同類別言論對壓力感受的影響。

（三）研究結果：

填寫人數：115 人，有效問卷：108 人，無效問卷：7 人。表 3-2、圖 3-2 為問卷數據分析結果。

言論類別	平均分數
肢體暴力	2.99
性話題	3.06
吸毒	2.64
損壞他人財物	2.84
使用槍械	2.81
酗酒	2.65
性行為	2.83
偷竊	2.66
助人	2.39
娛樂與節目	2.38
新聞議題	2.72

表 3-2：言論類別平均分數

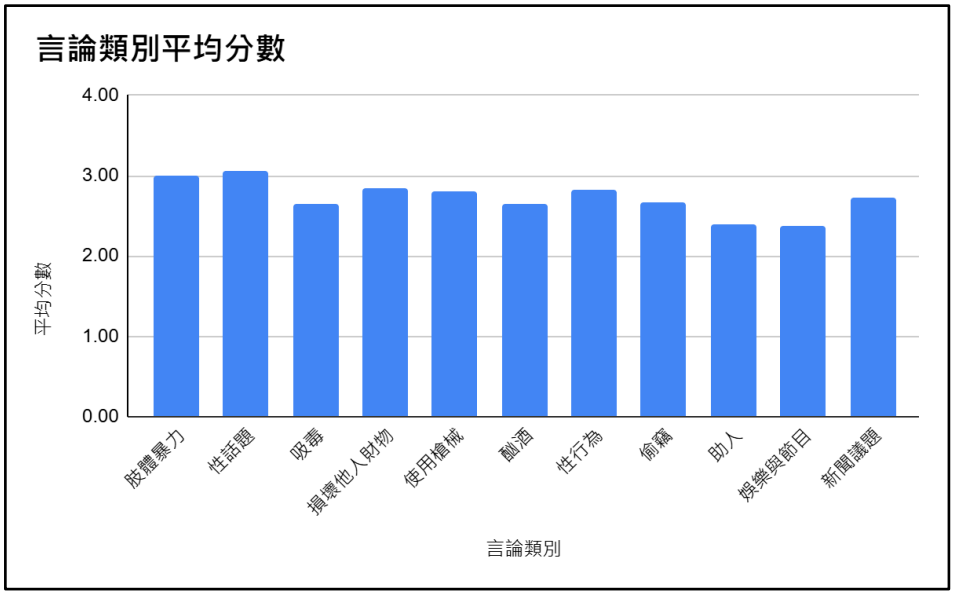


圖 3-2：言論類別平均分數長條圖（本圖為研究者自製 ）

1. 反社會內容得分：分數與受試者接觸的反社會內容（暴力、毒品等）頻率呈正相關（受試者平均分數：2.89。相關係數：0.46。）
2. C-ME 總分：分數與受試者媒體接觸頻率呈正相關
   （受試者平均分數：3.07。相關係數：0.4。）

（四）原定計畫：

透過表單蒐集受試者對言論的評分，以作為評分依據。

設計考量：

1. 標準化評分機制：使用統一的評分標準，確保不同受試者的評分具有可比性。
2. 數據客觀性：透過統計分析來識別言論評價的模式，避免主觀因素影響研究結果。

（五）問題與困難：

在初步數據分析後，發現問卷評分數據結果呈現高度相似性，差異性過低，無法有效區分不同類型言論造成的壓力程度。

我們一開始假設助人、娛樂與節目類的言論因為其攻擊性較低，因此數據應顯著低於暴力、偷竊等言論，但表單結果卻與假設不符。

推測可能的原因：

1. 評分標準模糊性：若受試者對於評分標準的理解存在偏差。例如：沒有完全理解壓力指數的意義。因此，可能導致評分結果集中在某些固定範圍，影響數據的分散性。
2. 語境影響：不同受試者可能因為背景知識或既有立場，導致對言論的評分傾向相似，從而降低數據的區分度。
3. 言論分數互相牽制：各類別各有兩則言論，可能導致同一類別的兩則言論壓力級別不同（一高一低），進而導致分數相互牽制。

（六）方法修正：

為解決上述問題，改採人工進行言論評分的方式。

設計考量：

1. 提高評分區辨力：由研究人員達成標準的共識後進行評分，以確保對不同言論的細微差異有更準確的辨識。
2. 確保一致性與可靠性：固定言論類別透過固定人員與標準來確保評分的一致性。

（七）暫定計畫：

將原本的評分標準（依照壓力指數評 1 至 5 分），改為 0：可以傳出但建議修改， 1：嚴重不當言論，禁止傳出，兩大類，由三位研究人員分別負責不同的言論類別，逐條言論進行分類。

三、自製言論資料集

目前網路上有許多現成的資料集可供使用，但大多不符合本研究的需求。現有資料集中的言論標籤通常未將言論單純分類為攻擊性、歧視性、性相關不當言論與普通言論，且大部分資料集主要為英文語料，缺乏中文資料集，尤其是針對臺灣人的語言使用習慣。本研究專注於分析中文的不當言論，並將其轉換為委婉的表達方式，因此研究者決定自製一個適用於本研究的中文言論資料集。

（一）言論的標籤

從表單題目找出所需的言論類別，再經過整理、分類、增加，以確保所有言論都可以被歸類，最終形成四大類言論類別（攻擊性言論、歧視性言論、性相關不當言論和普通言論）。將言論標上對應的標籤，各類別標籤分別對照相應的標準。

（二）言論的蒐集

從各大知名論壇網站蒐集各種言論，找尋攻擊性言論、歧視性言論、性相關不當言論和普通言論四大類不同類別的言論，表 3-3 為各類別蒐集言論的對應標準。

類別	對應標準
攻擊性言論	具有貶低他人、以不雅字眼謾罵別人、威脅別人、攻擊別人長相
歧視性言論	性別和性向、種族國籍、身體心理行為歧視，對特定個人或整個群體有偏見或不公的言行
性相關不當言論	涉及性騷擾、性羞辱、猥褻內容或其他讓人感到不適的性暗示言論，包括未經同意的性相關評論、對個人或群體的性別刻板印象，以及其他可能引起不適或冒犯的性相關表達。
普通言論	討論動物、心情、日常生活、電影等，不涉及不當言論。

表 3-3：各類別蒐集言論的對應標準

四、數據擴增

數據擴增(data augmentation)[5][6]適用於資料集數據量不足的情況，且可以防止模型過擬和提升泛化能力，本研究自製資料集總言論數為 1462 則，在測試的時候模型無法精準判斷言論所屬的類別（準確率 64.1%），推測是因為言論的數量不足，為了解決此問題，本研究使用數據擴增，來擴增言論數量。

（一）數據擴增歸為三種方式：

1. 同義字詞轉換：抽換詞面，讓模型可以學習更廣泛的字彙。
2. 語序改變：讓模型可以學習不同的文法結構，提高模型語言變化的適應能力。

3. 換句話說：結合以上兩種轉換模式，是數據擴增的主要方式。

（二）使用模型：Gemini，圖 3-3 為給 Gemini 的 prompt

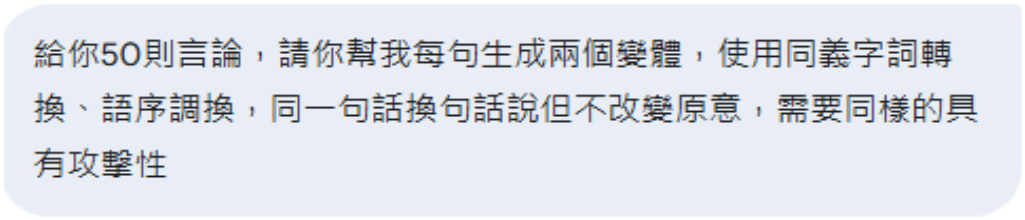


圖 3-3：給 Gemini 的 prompt（本圖為研究者自製 ）

圖 3-4 為變體生成範例，原文：理組男是不是真的比較噁

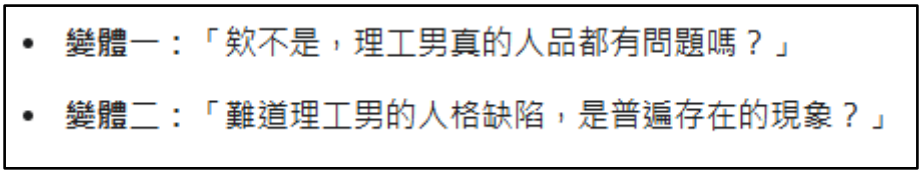


圖 3-4：使用 Gemini 變體生成範例（本圖為研究者自製 ）

一則言論可能含有多種違規類別，因此本研究所使用的情感分析模型為多標籤分類，且每個標籤獨立判斷，從表 3-4，攻擊性言論相比歧視性言論與性相關不當言論較多，原因為多數歧視性言論含有攻擊性言論所致。

由表 3-4，從 LSTM 模型的準確率從 64.1%提升至 88.7%可知數據擴增對於小資料集在數據量不足的情況下，能帶來顯著改善，但是，使用 Gemini 進行同義詞轉換的效果仍然有限，因為性相關不當言論中的不當性行為描述是較長的文本，且通常為一段描述，對生成式 AI 來說難以有效轉換為變體，此外，也有考慮到 Gemini 使用條款的問題，因此資料集中的性相關不當言論較其他言論稀少，使得資料集面臨數據不均衡的問題。

為了避免模型過度預測不當言論，除了真實在社群平台蒐集的 150 則言論，研究者還使用了 ChatGPT (GPT-4o-mini) 生成 1236 則普通言論，範圍涵蓋動物、心情、日常生活等，擴增言論的多樣性，考慮到不當言論為三種類別總稱，為了不讓普通言論在資料集中的占比過少，同時避免模型更為偏向普通言論的預測或與單一類別數量差

至過多，因此將占比設計為低於 50%，經過實測，雖然普通言論與不當言論的比例約為 1：2，但是模型對於普通言論的預測已有良好效果。

	原資料集	最終資料集
攻擊性言論(個)	383	1154
歧視性言論(個)	302	902
性相關不當言論(個)	302	604
普通言論(個)	586	1386
言論總數(則)	1462	3708
標籤總數(個)	1573	4046
LSTM 模型準確率	64.1%	**88.7%**

表 3-4：數據擴增前後的標籤數量及準確率

五、資料前處理：

資料前處理主要目的是使得文本更為簡潔，讓模型可以專注於重要特定字詞，在進行情感分析實驗之前，先對文本進行資料清理，確認蒐集的文本是否有重複或是標錯標籤的情況，圖 3-5 為資料前處理範例，圖 3-6 為斷詞與清除空行及標點符號後的狀態。本研究主要目標為對純文字的言論進行多標籤情感分析，因此去除了表情符號與數字，並將隱射性的符號改成文字，確認資料無缺失及漏標多標籤的言論後，進行以下資料前處理步驟：

（一）去除標點符號：標點符號本身無意義，將其去除更好分析。

（二）斷詞：中文句子本身沒有空格需要斷詞，否則可能會判斷錯誤無法有效提取特徵。使用 jieba 進行斷詞。

（三）去除停用詞：中文裡的停用詞，通常無特定意義，可以將其去除使得模型更好判斷特定字詞。

（四）向量化：將文本處理好後將其轉換成數字格式，以便機器學習，以下使本研究使用的方式。

1. Tokenizer（分詞器）：將文本中的內容拆解為較小單位（token）的工具，並將文本轉換為數字序列，讓模型能夠理解文本內容。
2. Embedding（詞嵌入）：將 token 轉換為低維詞向量，使模型能夠更好地擷取語意。

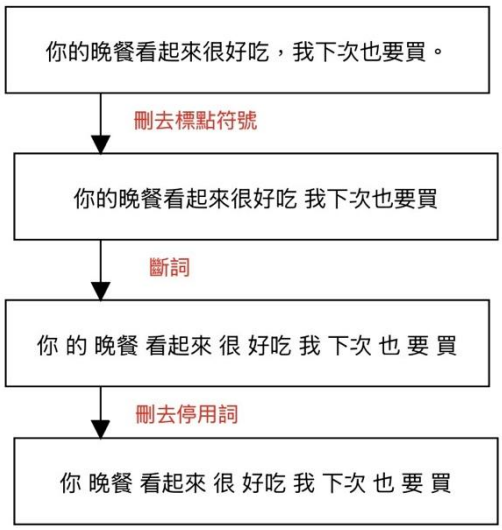


圖 3-5：資料前處理範例（本圖為研究者自製 ）

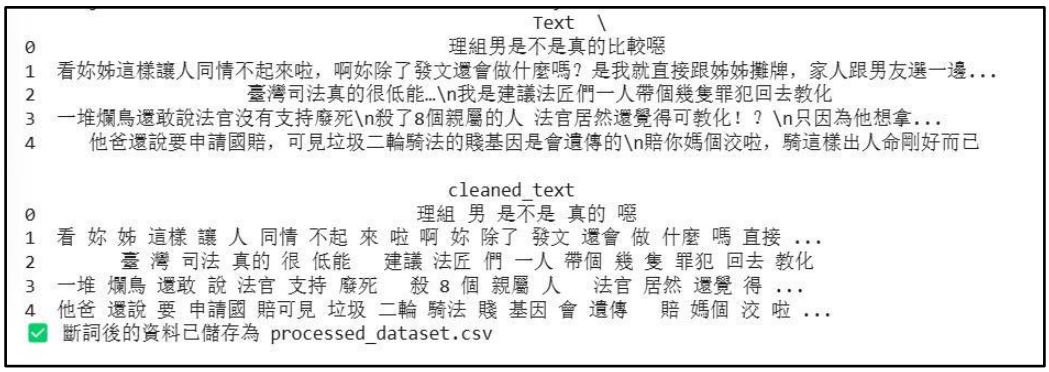


圖 3-6：斷詞與清除空行及標點符號（本圖為研究者自製 ）

六、模型設定

LSTM、CNN+LSTM、BiLSTM 模型原設定為 epoch 為 20 次，batch_size 為 16，Transformer 的 epoch 為 10 次，因每個模型所需的訓練次數不同，本研究中的情感分析

模型均使用早停法（Earlystop），在驗證集的準確率不再提升且損失函數持續增加時停止訓練，來防止過擬和（Overfitting），即在訓練資料模型預測效果良好，但預測新資料時無法準確預測的情況。

七、以情感分析模型判斷壓力指數

研究者曾經嘗試過將不當言論的等級再進行細分，例如：將攻擊性言論分為不嚴重可以傳出，但給修改建議，以及嚴重不當言論禁止傳出，並使用 LSTM 模型進行二元分類，標籤 0 代表可以傳出，1 代表禁止傳出，然而，測試後發現，準確率只有 52.6%，且所有預測皆為禁止傳出的言論，從表 3-6，標籤為 1 （嚴重不當言論，禁止傳出）的召回率高達 1.00，表示所有標籤為 1 的言論皆準確預測，而精確率只有 53%，表示其中有 47%的言論其實是誤判的，推測原因如下：

（一）本研究所自製的資料集壓力指數標籤僅有 2320 個，表 3-5 為各標籤的言論數量，在資料稀少的時候模型無法準確判斷，且針對不當言論再進行細分程度具有極大的困難度，因為皆為不當言論，言論判別可能有模糊地帶使模型難以預測標籤，

（二）在分類標籤時，將攻擊性言論、歧視性言論、性相關不當言論一起分類，研究者原本認為如果將各類別言論一起訓練，會使模型學習到各種言論的判別方式，且資料數量會比較多。然而，模型的準確率極為不理想，可能是因為不同言論類別嚴重程度的界定標準不同，為了證實此推測，研究者後續將各個類別分別以不同個 LSTM 模型進行分析。

標籤	言論數量
0（可以傳出但建議修改）	1230
1（嚴重不當言論，禁止傳出）	1090

表 3-5：各標籤的言論數量

標籤	精確率	召回率	F1-score
0（可以傳出但建議修改）	0%	0%	0%

1（嚴重不當言論，禁止傳出）	53%	100%	69%

表 3-6：各標籤判斷表現在 LSTM 模型中的評估結果

研究者後續將各個類別分別以不同個模型進行分析再次進行實驗，確認（二）的推論，從表 3-7，由攻擊性言論與歧視性言論的精確率分別為 53%與 51%，可知其假陽性率（即誤將實際為標籤 0 的言論判為標籤 1）分別達到 47% 與 49% ，而性相關不當言論的誤判率雖然不低，但相比之下只有 18%。即使一個標籤獨立一個模型分析言論，還是無法改善先前的問題，但可以發現一個共同點，這些模型全部都預測了單一標籤，因此或許可以對應到（二）的推論，因為數據較少，要訓練模型再將不當言論分出嚴重程度，需要大量的言論來訓練模型，可能為幾十萬筆至幾千萬筆，本研究中自製的小資料集具有侷限性，難以有效訓練模型。

攻擊性言論	精確率	召回率	F1-score
0	0%	0%	0%
1	53%	100%	69%
歧視性言論	精確率	召回率	F1-score
0	0%	0%	0%
1	51%	100%	68%
性相關不當言論	精確率	召回率	F1-score
0	82%	100%	90%
1	0%	0%	0%

0: 可以傳出但建議修改、1: 嚴重不當言論，禁止傳出

表 3-7：各標籤判斷表現在不同 LSTM 模型中的評估結果

考慮到是否與模型選擇有關，研究者也使用了 Transformer 模型進行分析，結果

與上述所說相同，因此最後確定推測為資料集的緣故。本研究在後續研究中使用了替代方案，以禁止詞彙作為是否可以傳出的依據，如圖 3-7，然而，使用禁止詞彙仍含有侷限性，例如：使用諧音或隱喻的方式無法判別到、非不當言論如果提到這些詞彙可能誤判。

```
BANNED_WORDS = [
    "去死", "殺了你", "智障", "白癡", "神經病", "傻逼", "腦殘", "賤人", "死垃圾", "人渣", "婊子",
    "死番仔", "西環納", "1.35仗", "猴子", "閩南猴", "死外勞", "黑鬼", "黑人魚", "尼哥", "小屌男",
    "支那", "支語", "羅剎", "鬼子語", "日本鬼子", "阿三","可悲"
]
```

圖 3-7：禁用詞彙（本圖為研究者自製 ）

此外，針對原先將言論分為標籤 0、1 的二元分類計畫 ，模型傾向預測單一類別，且二元分類可能誤判語意模糊的言論，研究者提出未來的改善方法：在擁有足夠的言論數量下，可以將言論標示為 1～10 分，並使用回歸模型預測，最後依分數設定範圍：0～3 分為「 可以傳出」 、 4～7 分為「可以傳出但建議修改」、8～10 分為「嚴重不當言論，禁止傳出」，同時也可以改善目前使用禁止詞彙的侷限。

肆、研究結果

本研究共分為三大部分：訓練情感分析模型、判斷言論是否可以傳送、利用生成式 AI 給予使用找修改建議。

一、訓練情感分析模型

（一）模型綜合比較

因為每次的測試集皆為不同的數據，因此本研究取十次分析的結果，並計算四分位距來刪除離群值，取得平均值以及標準差，表 4-1 為各模型的表現比較。

	準確率		精確率		召回率		F1-score	
	平均值	標準差	平均值	標準差	平均值	標準差	平均值	標準差
LSTM	**86.9%***	0.015	**89.8%***	0.009	**88.9%***	0.011	**89.2%***	0.008
CNN + LSTM	85.4%	0.017	88.8%	0.010	86.8%	0.017	87.7%	0.010
BiLSTM	84.9%	0.022	88.0%	0.011	85.8%	0.017	86.4%	0.013
Transformer	86.0%	0.024	87.5%	0.017	85.7%	0.023	86.4%	0.015

*: 表現最佳

表 4-1：各模型的表現比較

（二）各標籤在各模型中的比較

本研究主要著重於攻擊性言論、歧視性言論、性相關不當言論之判斷，因此下列分析會聚焦於不當言論，略過各指標表現較優的普通言論。由於資料集中每個標籤的數量不均衡，故本研究以 F1-score 作為主要指標來進行分析，表 4-2 為各標籤判斷在各模型的表現比較。

依表 4-2，可知 LSTM 模型在攻擊性言論、歧視性言論、普通言論中的 F1-score 平均是最高的，顯示出該模型的穩定性，攻擊性言論的 F1-score 甚至達到 94.2%。

而 BiLSTM 模型與 Transformer 模型在判斷攻擊性言論的表現相對較差，LSTM 模型與 CNN + LSTM 模型表現突出，顯示這兩個模型成功抓取了攻擊性言論的特徵。其中，可以明顯發現性相關不當言論在各個模型中的 F1-score 皆低於其他標籤，而普通言論在各個模型中的預測表現皆優異。雖然 Transformer 模型在攻擊性言論及歧視性言論判斷表現不佳，但是性相關不當言論的 F1-score 平均卻位居第一，這可能展現出 Transformer 模型在分析複雜及多樣性語句時，相對其他模型更具有優勢。

	LSTM		BiLSTM		CNN + LSTM		Transformer	
	平均值	標準差	平均值	標準差	平均值	標準差	平均值	標準差
攻擊性言論	**89.0%***	0.019	85.9%	0.015	88.2%	0.015	83.2%	0.024
歧視性言論	**88.0%***	0.012	86.2%	0.019	86.7%	0.019	84.4%	0.039
性相關不當言論	78.4%	0.020	72.8%	0.043	74.5%	0.027	**80.5%***	0.019
普通言論	**94.2%***	0.008	92.5%	0.010	93.4%	0.009	93.0%	0.009

*: 表現最佳

表 4-2：各標籤判斷在各模型的表現比較

二、使用模型判斷言論類別

下面三張圖分別代表三種情況，第一種為判斷出不當言論，建議修改，第二種為出現禁止詞彙，所以無法發送言論，第三種為普通言論

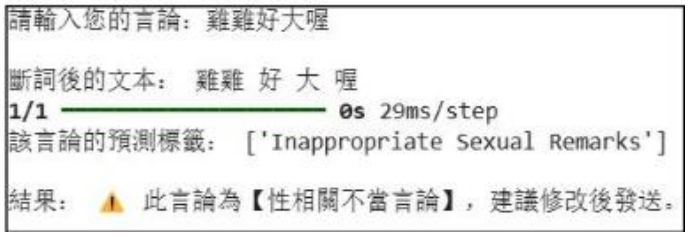


圖 4-1：分析結果範例一（本圖為研究者自製 ）

圖 4-2：分析結果範例二（本圖為研究者自製 ）

圖 4-3：分析結果範例三（本圖為研究者自製 ）

為了評估各模型的訓練結果 ，研究者選取九句言論進行實測，結果如表 4-3 所示 。

三、轉換言論

本研究使用 OpenAi 函式庫，取得 API 後連接 GPT-4o-mini 模型，對不當言論進行轉換，圖 4-4 為給 GPT-4o-mini 的 prompt

```
'''
如果情感分析完是攻擊性言語和歧視性言還有性相關不當言論，就幫我改成委婉的方式說，
但不要改變原意或是變得正向，假如判斷出留言含有禁止詞彙像是:
「去死，殺了你，智障，白癡，神經病，傻逼，腦殘，賤人，死垃圾，人渣，婊子，死番仔，西環納，1.35 仗，猴子
，閩南猴，死外勞，黑鬼，黑人魚，尼哥，小屌男，支那，支語，羅剎，鬼子語，日本鬼子，阿三，可悲」
，則回傳「此言論含有嚴重不當詞彙，請刪除後再試」，然後再說建議修改成的委婉言論，記得用繁體字
'''
```

圖 4-4：給 GPT-4o-mini 的 prompt （本圖為研究者自製 ）

假如判斷出留言含有禁止詞彙，則會回傳「此言論含有嚴重不當詞彙，請刪除後再試」並提供留言者修改建議，如果留言沒有禁止詞彙，但被判斷為不當言論，會

給留言者修改建議，讓留言者再次審視自己的言論，表 4-3 為言論預測結果及修改後的言論建議，圖 4-5 為程式執行結果。

圖 4-5：程式執行結果（本圖為研究者自製 ）

編號	原始言論	原始類別	預測類別		修改言論	預測正確
1	閉嘴吧，人渣一個	攻擊性言論	A	攻擊性言論	此言論含有嚴重不當詞彙，請刪除後再試。建議修改成的委婉言論：請您保持安靜，您的言行實在讓人失望。	✔
			B	攻擊性言論		✔
			C	攻擊性言論		✔
			D	攻擊性言論		✔
2	你真是垃圾，去死吧	攻擊性言論	A	攻擊性言論	此言論含有嚴重不當詞彙，請刪除後再試。建議修改成的委婉言論：您所表達的意見讓我感到非常失望，希望我們能以更建設性的方式進行	✔
			B	攻擊性言論		✔
			C	攻擊性言論		✔

			D	攻擊性言論	交流。	✔
3	誰有更大的機率讓我不用再看到黑鬼人魚這種東西我就支持誰	歧視性言論	A	普通言論	此言論含有嚴重不當詞彙，請刪除後再試。建議修改成的委婉言論為：「我希望能看到更具包容性和尊重的討論，這樣我會更支持相關內容。」	✖
			B	歧視性言論		✔
			C	普通言論		✖
			D	攻擊性言論		✖
4	呃.....？？？？說 5678 只是想酸她沒結婚沒男方就扶養小孩 跟 89 行為一點屁關係都沒有 也扯太遠了吧 傻眼 我修正 在我看來妳有點可悲ㄟ 希望妳引以為恥	歧視性言論, 攻擊性言論	A	攻擊性言論	我覺得您提到的事情有些偏離主題，讓我感到驚訝。在我看來，這樣的情況可能會讓人感到有些不幸，希望您能夠重新思考這些事情。	Δ
			B	歧視性言論,攻擊性言論		✔
			C	歧視性言論		Δ
			D	攻擊性言論		Δ
5	？你有意見嗎臭 gay	歧視性言論	A	性相關不當言論	你對此有什麼看法呢？	✖
			B	歧視性言論		✔

			C	歧視性言論		✔
			D	攻擊性言論		✖
6	可以把我綁起來操嗎	性相關不當言論	A	歧視性語論	我希望能夠以某種方式享受我們之間的親密互動。	✖
			B	歧視性語論		✖
			C	歧視性語論		✖
			D	歧視性語論		✖
7	小五小六拿枕頭摩擦或者把玩偶塞在褲子裡隔著內褲摩擦	性相關不當言論	A	性相關不當言論	小五小六可能會用枕頭進行摩擦，或是將玩偶放在褲子裡，隔著內褲進行接觸。	✔
			B	普通言論		✖
			C	普通言論		✖
			D	性相關不當言論		✔
8	人生很寬闊不要局限自己	普通言論	A	普通言論	無	✔
			B	普通言論		✔
			C	普通言論		✔
			D	普通言論		✔
9	運動能保持最佳體態	普通言論	A	普通言論	無	✔
			B	普通言論		✔

			C	普通言論		✔
			D	普通言論		✔

A：LSTM、B：CNN+LSTM、C：BiLSTM、D：Transformer

表 4-3：言論預測結果及修改後的言論建議

伍、討論

一、模型綜合比較

依表 4-1，可以發現 LSTM 模型在所有指標上均表現卓越，且各指標的標準差最小，其中 F1-score 平均更是達到 89.2%，顯示 LSTM 在處理數據不均衡的資料集時表現優於其他模型，推測原因為 LSTM 能夠有效捕捉長期依賴的關係，所以能提升預測的準確性。雖然 CNN + LSTM 模型表現比 LSTM 模型遜色，但 F1-score 平均仍達到 87.7%（標準差 0.010），顯示經過 CNN 的局部特徵提取後再利用 LSTM 模型進行文本分類仍具有不錯的效果。Transformer 模型的準確率平均 86.0%位居第二，然而，它的精確率、召回率及 F1-score 平均均略低於 LSTM 與 CNN + LSTM，顯示在數據不均衡的情況下 Transformer 模型的預測能力有限，推測原因為 Transformer 模型需要大量的文本來進行訓練，在標籤數只有 4046 個的小資料集上的效果相對受限。BiLSTM 模型在準確率明顯低於其他模型，平均值只有 84.9%（標準差 0.022），但是 F1-score 平均仍高於 Transformer 模型，可知 BiLSTM 模型還是具有一定的預測能力，研究者原本認為 BiLSTM 模型能夠雙向分析，效果應該會優於 LSTM，但結果並不然，因此推測 BiLSTM 模型在判斷某些言論類別時具有限性，下一段會分析哪種言論類型模型較無法準確判斷。

二、各標籤在各模型中的比較一性相關不當言論 F1-score 在各模型表現皆偏低的原因

研究者認為性相關不當言論普遍 F1-score 在各個模型中的表現皆不如其他標籤的原因，為資料集中數量不均衡所致，推測因為性相關不當言論主要為物化與不當性器官評價、不當性行為與性器官描述，在進行數據擴增時，不當性行為與性器官描述的言論無法使用 Gemini 對言論產生兩種變體，且文本較其他標籤的言論多樣且複雜，沒有明顯的特徵，模型較難以學習，因此需要更大量的資料進行訓練。

三、轉換言論

由表 4-3，可知各模型在攻擊性言論判別上表現良好，本研究中所選用的兩種攻擊性言論皆準確判斷，Transformer 模型在歧視性言論判別的表現較其他模型差，但是在性相關不當言論的表別中表現良好，其中，編號 6 言論在各模型一致預測為歧視性言論，推測可能是因為該言論較為隱晦，帶有性暗示，因此模型難以捕捉到它的特徵，造成誤判。CNN + LSTM 模型成功預測出編號 4 言論的兩種標籤，而其他模型僅成功預測出其中一種標籤，訓練模型時，LSTM 模型的各個指標略遜色於 CNN + LSTM 模型，然而，使用訓練後的模型再次進行分析時，研究者發現 CNN+LSTM 模型在準確度上略勝於 LSTM 模型。

從表 4-3 可以看出，GPT-4o-mini 在針對性相關不當言論進行轉換時具有困難性，特別是編號 7 言論，這種較長且描述性的言論 GPT-4o-mini 無法有效轉換成更委婉的言論。不過，在攻擊性言論有顯著的效果，轉換的言論較為精準，且能夠用委婉的方式表達大致原意，雖然歧視性言論也有不錯的效果，但是相比原意仍可能有偏差，如編號 5，研究者認為很難將不當言論轉換成完全與原意相同的委婉言論，因為有些言論是刻意的貶低或侮辱，難以保持原意，因此利用 GPT-4o-mini 轉換言論仍算理想。

陸、結論

一、本研究使用多種自然語言處理模型進行測試，其中 LSTM 模型表現最佳 ，F1-score 達 89.2%，能夠較準確的辨識言論類別。

二、在實測訓練好的模型時，CNN + LSTM 模型的分析效果最佳，能更有效地判別言論類別。

三、連接 GPT-4o-mini 能有有效的將不當言論轉換成較委婉的表達方式，並提供留言者修改建議。

四、未來展望：

（一）擴大言論資料集，蒐集更多各大社群平台的言論，擴大言論的多樣性，例如：不同文化、背景語言以及年齡層，提升言論判斷的廣度以及準確率。

（二）蒐集更多不當言論，並根據違規嚴重程度將他們分為可以傳送以及不能傳送，不封鎖所有不當言論，而是讓使用者能選擇想看到的言論，可以傳送的不當言論會先被遮蔽，觀看者點擊同意鍵才能看見該言論，並設有提醒含有哪種類別的不當言論。
（三）將情感分析及言論轉換模型運用到留言框進行即時分析，讓留言者在發言前能夠再次審視自己的言論，同時也要保護使用者的隱私，不會記錄言論內容。

柒、參考文獻資料

[1] den Hamer, A. H., Konijn, E. A., Plaisier, X. S., Keijer, M. G., Krabbendam, L., & Bushman, B. J. (2017). The Content-based Media Exposure Scale (C-ME): Development and validation. Computers in Human Behavior, 72, 549-557.

[2] Teng, Z., & Zhang, Y. (2016). Bidirectional tree-structured LSTM with head lexicalization. arXiv preprint arXiv:1611.06788.

[3] O’ Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458.

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2023). Attention is all you need. arXiv preprint arXiv:1706.03762.

[5] Lee, K., Guu, K., He, L., Dozat, T., & Chung, H. W. (2021). Neural data augmentation via example extrapolation. arXiv preprint arXiv:2102.01335.

[6] Kumar, V., Choudhary, A., & Cho, E. (2021). Data augmentation using pre-trained transformer models. arXiv preprint arXiv:2003.02245.

[7] Young, K. S. (1998). Caught in the net: How to recognize the signs of Internet addiction--and a winning strategy for recovery. New York: Wiley.

[8] 王承諺、李明憲（2020）。社群網站網路霸凌和網路攻擊辨識量表之發展。測驗學刊，67(1)，61-94。

【評語】052514

本作品參考 C-ME 量表整理出四類不同言論標籤，並從論壇蒐集的言論進行人工分類來訓練模型。研究採 LSTM、雙向 LSTM、CNN + LSTM 及 Transformer 四種深度學習模型，針對所自建資料集進行訓練並分類，進而比較四種模型在精確率、召回率及 F1-score 等指標上的表現。研究主題雖有應用性，但因為實驗僅為四個既有模型的比較，建議要對於學習模型有進一步的精進或發展，並進行實驗比較，以提昇研究與技術的創新性。

作品海報

AI與心理學的言語柔化實驗

摘要

本研究旨在參考現有 C-ME 量表的類別，將其延伸整理為四大類標籤，從知名網路論壇蒐集言論並使用人工分類的方式自製資料集訓練模型。研究採用 LSTM、雙向 LSTM、CNN + LSTM 及Transformer 四種機器學習模型，基於自建資料集進行訓練，實現對全新言論的精準分類，並比較四種模型在準確率、精確率、召回率及 F1-score 等的表現。

結果顯示，LSTM模型在處理數據不均衡的資料集時表現最佳，進一步分析發現，CNN + LSTM 在預測效果上略勝 LSTM。此外，結合生成式 AI GPT-4o-mini，能有效改善不當言論，使言論委婉化。

壹、前言

一、研究動機

隨著社群媒體蓬勃發展，人們能夠輕鬆地在網路上發表想法。然而，部分留言者可能發表過激或侮辱性言論，使他人感到不適或造成心理傷害。儘管許多知名論壇應用程式已設有言論篩審機制，但由於篩審可能過於寬鬆，不當言論仍然屢見不鮮。為了改善此現象，本研究旨在強化不當言論的篩審機制，並提供適當的修改建議，以營造更友善的網路環境。

二、研究目的

（一）自製言論資料集，將言論標上攻擊性言論、歧視性言論、性相關不當言論和普通言論四大類標籤。
（二）運用訓練後的 LSTM、雙向 LSTM、CNN + LSTM、Transformer 模型判斷新言論屬於何種類別。
（三）將不當言論依嚴重程度再細分為是否禁止傳出。
（四）連接 ChatGPT-4o-mini 將不當言論改為較委婉的用詞。

三、文獻回顧

（一）C-ME 量表[1]

內容導向媒體接觸量表（Content-based Media Exposure Scale, C-ME），是一種標準化工具，用於測量個體接觸特定媒體內容的頻率，特別適用於青少年群體，可適用於不同媒介（如社交媒體、遊戲等）。

（二）混淆矩陣

本研究以F1-score做為主要評估模型的指標

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

		實際	
		Positive	Negative
預測	Positive	TP (True Postive)	FP (False Postive)
	Negative	FN (False Negative)	TN (True Negative)

表一 ：混淆矩陣通常結構表示

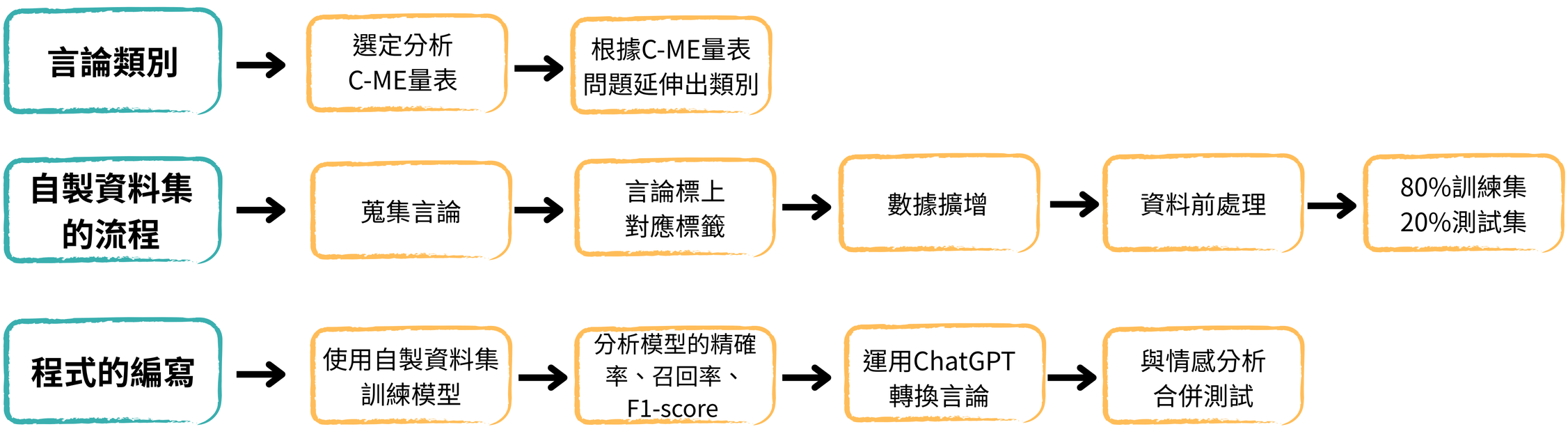
		實際	
		攻擊性言論	非攻擊性言論
預測	攻擊性言語	實際：攻擊性言論 預測：攻擊性言論 （TP，預測正確）	實際：非攻擊性言論 預測：攻擊性言論 （FP，預測錯誤）
	非攻擊性言語	實際：攻擊性言論 預測：非攻擊性言論 （FN，預測錯誤）	實際：非攻擊性言論 預測：非攻擊性言論 （TN，預測正確）

表二：攻擊性言論的混淆矩陣

貳、研究過程及方法

一、流程圖

圖一 ：流程圖



二、自製言論資料集

1.各個類別標籤及言論蒐集標準

攻擊性言論	具有貶低他人、以不雅字眼謾罵別人、威脅別人、攻擊別人長相
歧視性言論	性別和性向、種族國籍、身體心理行為歧視，對特定個人或整個群體有偏見或不公的言行
性相關不當言論	涉及性騷擾、性羞辱、猥褻內容或其他讓人感到不適的性暗示言論，包括未經同意的性相關評論、對個人或群體的性別刻板印象，以及其他可能引起不適或冒犯的性相關表達。
普通言論	討論動物、心情、日常生活、電影等，不涉及不當言論。

表三 ：各標籤言論蒐集標準

2.數據擴增的三種常用方式

1. 同義字詞轉換
2. 語序改變
3. 換句話說

本研究使用Gemini來對每則言論生成兩種變體，進行數據擴增

給你50則言論，請你幫我每句生成兩個變體，使用同義字詞轉換、語序調換，同一句話換句話說但不改變原意，需要同樣的具有攻擊性

圖二：給Gemini 的 prompt

- 變體一：「欸不是，理工男真的人品都有問題嗎？」
- 變體二：「難道理工男的人格缺陷，是普遍存在的現象？」

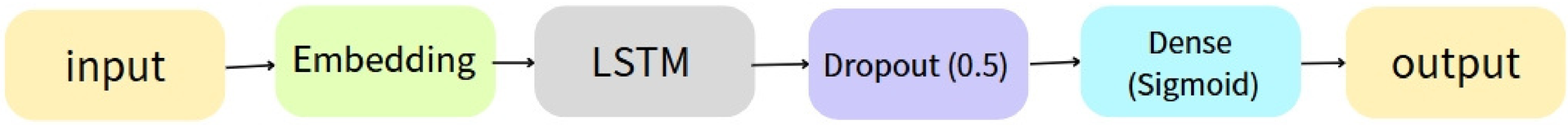
圖三：使用 Gemini 變體生成範例

	原資料集	最終資料集
攻擊性言論	383	1154
歧視性言論	302	902
性相關不當言論	302	604
普通言論	586	1386
言論總數	1462	3708
標籤總數	1573	4086
LSTM模型準確率	64.1%	88.7%

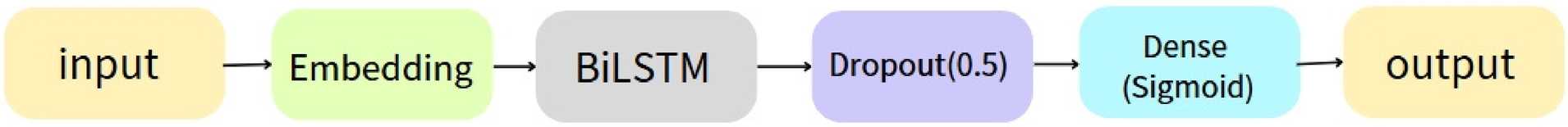
表四：數據擴增前後的標籤數量及準確率

三、情感分析模型

本研究使用LSTM、BiLSTM、CNN + LSTM 、Transformer作為情感分析模型



圖五：本研究使用的LSTM模型



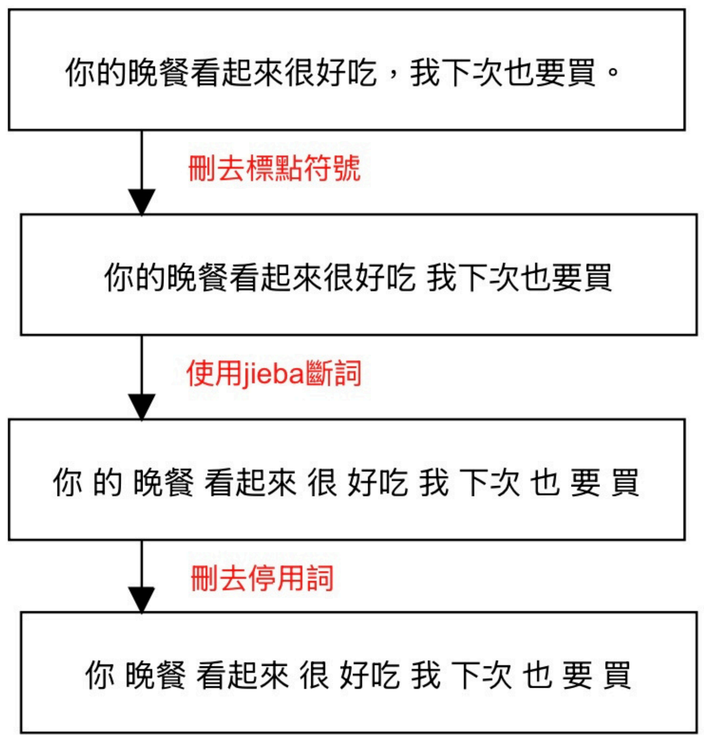
圖六：本研究使用的BiLSTM模型

圖七：本研究使用的CNN + LSTM模型

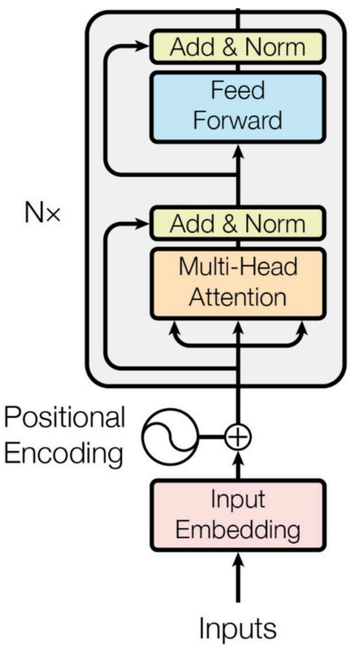
圖八：本研究使用的Transformer模型

3.資料前處理

（一）去除標點符號
（二）斷詞
（三）去除停用詞
（四）向量化
1.Tokenizer（分詞器）
2.Embedding（詞嵌入）



圖四：資料前處理範例



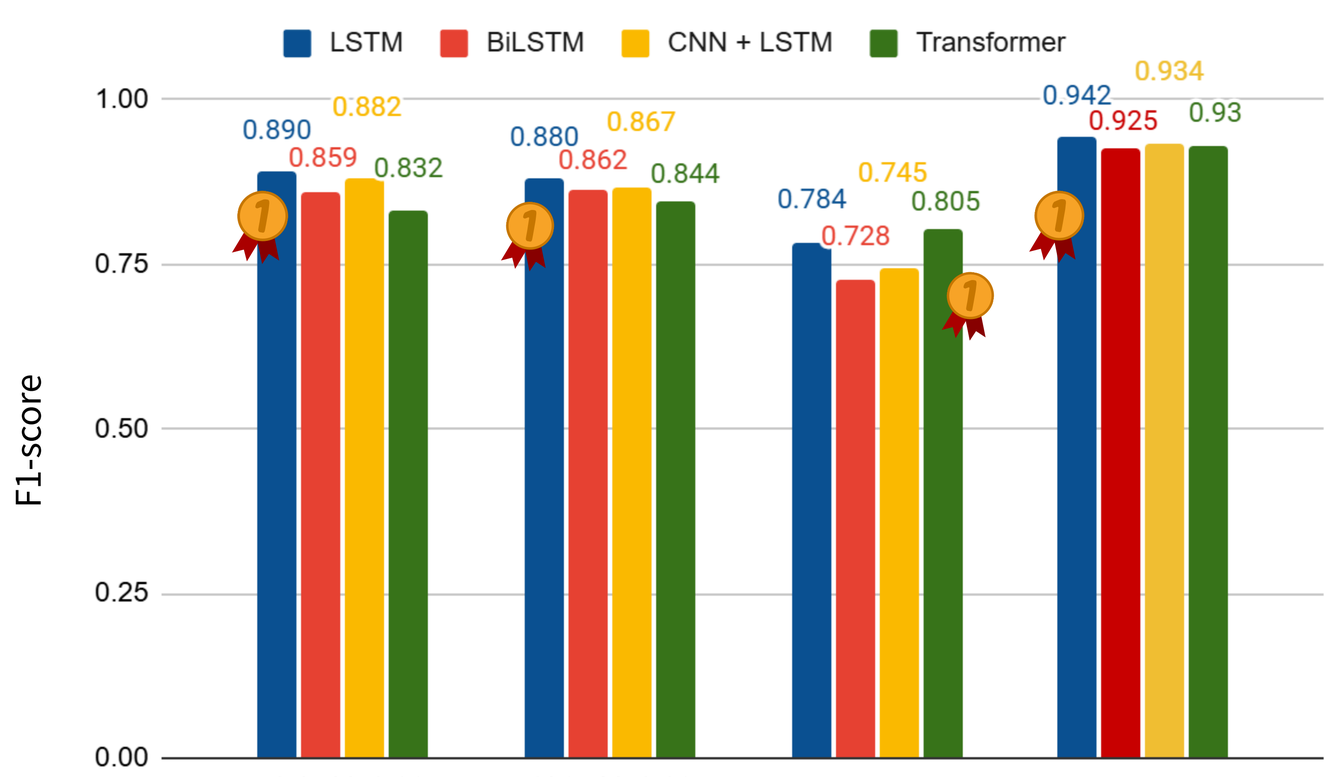
圖九：Transformer Encoder [2]

參、研究結果

一、LSTM、BiLSTM、CNN + LSTM 、Transformer模型取十次平均的F1-score

圖十：不同模型對言論資料集的F1-score圖

圖十一 ：不同模型對標籤種類的F1-score圖



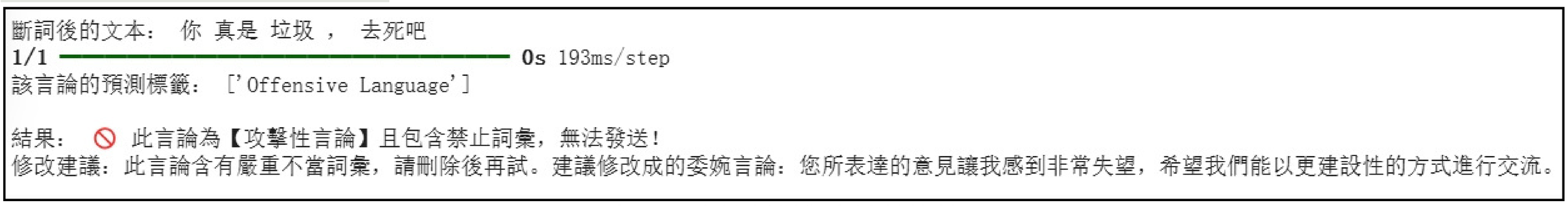
二、模型實測結果

表五：各模型正確判斷次數

LSTM	5/9次	BiLSTM	5/9次
CNN + LSTM	**7**/9次	Transformer	5/9次

本研究挑選各類別言論各兩則與一則含攻擊性言論與歧視性言論標籤的言論。只有CNN + LSTM模型正確判斷出兩種標籤，其餘模型皆只判斷出攻擊性言論。

三、言論判斷與言論轉換結果



圖十二 ：言論判斷與言論轉換結果

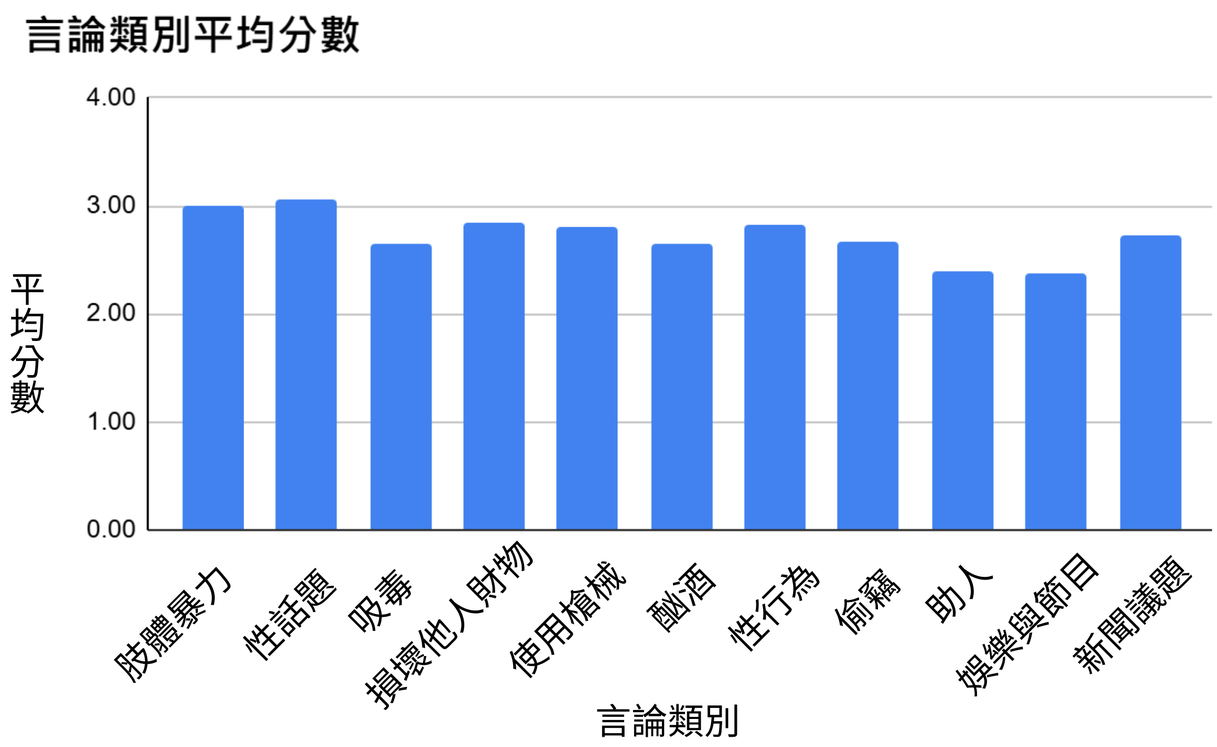
肆、討論

一、心理學分析

（一）原定計畫

透過表單蒐集受試者對言論的評分，以作為評分依據。

（二）研究結果



圖十三：言論類別平均分數長條圖

（三）問題與困難

根據圖十三，發現問卷評分數據結果呈現高度相似性，導致差異性過低，無法有效區分不同類型言論造成的壓力程度。 推測原因：

1. 評分標準模糊性：受試者對於評分標準的理解存在偏差。例如：沒有完全理解壓力指數的意義。因此，可能導致評分結果集中在某些固定範圍，影響數據的分散性。

2. 語境影響：不同受試者可能因為背景知識或既有立場，導致對言論的評分傾向相似，從而降低數據的區分度。

3. 言論分數互相牽制：各類別各有兩則言論，可能導致同一類別的兩則言論壓力級別不同（一高一低），進而導致分數相互牽制。

（四）方法修正：改採人工進行言論評分的方式。

（五）暫定計畫： 將原本的評分標準（依照壓力指數評 1 至 5 分），改為兩類 0：可以傳出但建議修改， 1：嚴重不當言論，禁止傳出，並由三位研究人員分別進行分類。

二、使用二元分類判斷是否為不當言論的侷限性

問題：

（一）模型無法有效區分言論嚴重程度

（二）不同類別的嚴重程度界定標準不同，將各類別言論一起訓練會具有模糊性

標籤	精確率	召回率	F1-score
0（可以傳出但建議修改）	0%	0%	0%
1（嚴重不當言論，禁止傳出）	53%	100%	69%

表六：使用LSTM模型區分言論的嚴重程度

三、性相關不當言論F1-score在各模型表現皆偏低的原因

研究者發現性相關不當言論F1-score在各模型表現皆偏低，推測為以下原因：

（一） 不當性行為與性器官描述的言論因為具有故事性，難以使用 Gemini 對言論產生換句話說的變體，導致性相關不當言論數量較少。

（二） 文本沒有明顯的特徵，言論類型多變，模型較難以學習，需要更大量的資料進行訓練。

四、轉換言論

攻擊性言論在分類與轉換言論的效果較好，下列為一則攻擊性言論在LSTM模型的實測結果。

原始言論	判斷類別	修改言論
閉嘴吧，人渣一個 （攻擊性言論）	攻擊性言論	此言論含有嚴重不當詞彙，請刪除後再試。 建議修改成的委婉言論：請您保持安靜，您的言行實在讓人失望。

伍、結論

一、本研究使用多種自然語言處理模型進行測試，其中 LSTM 模型表現最佳 ，F1-score達89.2%，能夠較準確的辨識言論類別。

二、在實測訓練好的模型時，CNN + LSTM 模型的分析效果最佳，能更有效地判別言論類別。

三、連接GPT-4o-mini能有效的將不當言論轉換成較委婉的表達方式，並提供留言者修改建議。

陸、參考文獻資料

[1] den Hamer, A. H., Konijn, E. A., Plaisier, X. S., Keijer, M. G., Krabbendam, L., & Bushman, B. J. (2017). The Content-based Media Exposure Scale (C-ME): Development and validation. Computers in Human Behavior, 72, 549-557.

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2023). Attention is all you need. arXiv preprint arXiv:1706.03762.

[3] Lee, K., Guu, K., He, L., Dozat, T., & Chung, H. W. (2021). Neural data augmentation via example extrapolation. arXiv preprint arXiv:2102.01335.

[4] Kumar, V., Choudhary, A., & Cho, E. (2021). Data augmentation using pre-trained transformer models. arXiv preprint arXiv:2003.02245.