The 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)
Taipei City, Taiwan, October 20-21, 2023. The Association for Computational Linguistics and Chinese Language Processing

生成模型是否能用於偵測身體羞辱仇恨言論?
Can generative models be used to detect hate speech related to body shaming?

蔡元翔
國立臺北商業大學企業管理學系
e10851023@ntub.edu.tw

張瑜芸
國立政治大學語言學研究所
yuyun@nccu.edu.tw

摘要

本研究實驗 Encoder 和 Decoder 兩者架構下的預訓練語言模型是否能很好的判斷身體羞辱的仇恨言論。先前研究多針對大型語言模型中的生成模型是否會生成歧視言語進行討論和防範，但尚未有研究進一步判別生成模型是否可應用於自動分類判斷歧視言論。因而本研究採用零樣本分類方式並提供完整身體羞辱定義，觀察以 Decoder 架構爲主的生成模型 (ChatGLM-6B 和 Chinese-Alpaca-Plus-7B 模型) 是否適用於自動判別歧視言論。此外，爲了更完整的了解大型語言模型不同架構下對於仇恨言論判斷結果，也採用以 Encoder 架構爲主的 BERT 模型進行分類判斷，並將兩架構下的結果做進一步分析比對。最終結果顯示 BERT 經過少量微調資料下就能獲得相對好的性能，生成模型在零樣本分類上確實是有些困難，需要進一步改善提示模板工程，提供多種句型結構的句子詳加解釋後，在少樣本分類上觀察生成模型表現是否能進一步提升。

Abstract

This study experiments with both Encoder and Decoder architectures of pre-trained language models to determine their effectiveness in identifying hate speech related to body shaming. Previous research has largely focused on discussing and mitigating the automatic generation of discriminatory language in generative models within LLMs. However, there hasn't been research investigating the further application of generative models in automatically classifying and identifying discriminatory language. Therefore, this study employs a zero-shot classification approach and provides a comprehensive definition of body shaming to examine whether Decoder-focused generative models (ChatGLM-6B and Chinese-Alpaca-Plus-7B) are suitable for automatically identifying discriminatory language. Furthermore, to gain a more comprehensive understanding of how different architectures within LLMs perform in hate speech detection, a BERT model with an Encoder architecture is also employed for classification. The results from both architectures are then further analyzed and compared. BERT shows good performance with minimal fine-tuning data, while generative models struggle with zero-shot classification. Thus we aim to explore the potential for improving the performance of generative models by providing detailed explanations for sentences with various structures.

關鍵字：零樣本分類、提示模板工程、仇恨言論檢測、身體羞辱　***Keywords:*** zero-shot classification、prompt engineering、detection of hate speech、body shaming

1 緒論

仇恨言論的偵測在自然語言處理的範疇中，一直以來都是一個很重要的任務，儘管自動偵測仇恨言論已取得多年的進步，但在這項任務上仍然面臨許多困難。其一是仇恨言論涵蓋的範圍太大，多數文獻專注探討仇恨言論的大方向，想訓練出涵蓋所有類型的模型，但 Gambäck and Sikdar (2017) 認爲這樣會讓許多推文被錯誤分類，也很容易忽視較少被關注的群體；其二是過往的研究中，有 73% 建立在監督式機器學習上，使資料集的取得在此研究變得更加重要 (Jahan and Oussalah, 2023)。且普遍監督式學習遇到的重大挑戰是缺乏各種少數語言、上下文不連貫或大量無標記數據 (Poletto et al., 2021; Jahan and Oussalah, 2023)；其三是仇恨言論很主觀的取決於環境，受到各種因素影響其定義，例如地理位置、社會規範和文化背景等等 (Waseem and Hovy, 2016)。

爲了避免有限資料集的影響，出現使用基於大型語言模型 (Large Language Model，以下

簡稱 LLM) 的基礎知識來做零樣本分類任務的研究 (Del Arco et al., 2023)。雖然 LLM 在各種任務中的表現都非常傑出，不過也有研究指出 LLM 較難理解攻擊性言論 (Wang and Chang, 2022)，且歧視文本的特性多是用譬喻或諷刺手法呈現，例如：「鳥仔腳像兩根竹籤插在貢丸上 (意指腿太細)」、「自稱棉花糖女孩眞的是很樂觀呢 (諷刺女性利用棉花糖來修飾自身體態)」，人類在遇到相似句型時，會聯想到單詞在該情境下的畫面進而理解完整含義，對模型來說需要具體了解到單詞意思之外，還要能夠把單詞所代表的涵意跟該情境融合，這是一件很困難的挑戰。

由於許多文獻指出 LLM 在文本生成上可能會產出歧視言語 (Hacker et al., 2023; Nadeem et al., 2020)，但沒有提到如果反向運用生成模型做歧視判斷的話，會不會有相同情況。因此本研究以仇恨言論中的身體羞辱爲主題，探索此類敏感性話題判斷較適合採用哪種類型的 LLM 模型做後續判斷。

大多 LLM 都是採用大量語料訓練下將 Embeddings 訓練出能模擬日常語言用法，模型應用趨勢也流行零樣本或少樣本進行下游任務調整。因應目前 LLM 模型於下游任務的使用趨勢，本研究依循此方式，觀察生成模型除了可能產出歧視言論之外，是否適合歧視文本判斷。此外生成模型爲 Decoder 架構，其模型判斷上可不使用任何標記資料即可進行，爲了更清楚了解 LLM 於判斷歧視言論此類相關敏感議題時是否還是需要仰賴少樣本進行微調，因此我們也採納 Encoder 架構所搭建的 BERT 模型作爲對比。

2 文獻回顧

2.1 身體羞辱

也就是基於外表的羞辱，皆與身體有關。是一種負面社交互動的術語，且經常出現在社群媒體中。Schlüter et al. (2021) 提出明確的科學定義「身體羞辱是一種非重複性行爲，跟一般網路霸凌 (長期、特定目標) 有所區別，在這種行爲中加害者不請自來的對目標身體表達負面意見或評論。加害者不一定有意傷害被害者，但被害者認爲該評論是負面、具有冒犯性或使其產生身體羞恥感。因此身體羞辱的範圍從善意的建議 (例如來自朋友基於醫學上的建議：你應該去減肥，以預防有高血壓) 到惡意的羞辱 (例：你的腿看起來很粗)」。此研究提及，身體羞辱似乎是隨著社群媒體而快速發展的，在社群媒體中過多修飾的極端照片，會成爲想模仿此形象的觸發因素，怕身體形象偏離規範而被社會排斥或處於不利的地位。

2.2 大型語言模型

在最簡單的形式中語言模型會爲特定的單詞序列分配相對應的機率，例如「在樹上的貓」會比「在樹上的魚」出現的機率更高。隨著時間的推移，許多研究人員已經創建各種語言模型來實現預測，其中最具代表性的是 Google 在 2017 年提出的 Transformer 架構。Transformer 是一種神經網絡的網絡架構，比起其他許多方法更快的進行訓練 (Vaswani et al., 2017)。

Transformer 基於注意力機制（Attention mechanism），並採用 Encoder - Decoder 結構。它們都由多層堆疊的自注意力層和前饋神經網絡（Feed-Forward Neural Network）層組成。Encoder 負責將輸入序列編碼成一個固定長度的上下文向量，而 Decoder 則通過關注 Encoder 輸出來生成目標序列。Transformer 神經網路架構開啓語言模型的新時代，出現許多在此架構上被廣泛使用的預訓練語言模型，其中以 Encoder 爲主的預訓練模型有 BERT (Devlin et al., 2018)、Electra (Clark et al., 2020)，Decoder 則包括 Bard (Thoppilan et al., 2022)、GPT-3 (Brown et al., 2020) 和 LLaMA (Touvron et al., 2023) 等等，在許多領域中都展現了卓越的性能。

2.3 零樣本分類

傳統的文本分類任務需要帶有標籤的資料進行訓練，讓模型學習不同標籤之間的關係，並在測試時對未知文本進行已知標籤的分類。零樣本分類則是在沒有該標籤的訓練集情況下，對未見過的標籤進行分類，目標是讓模型能夠將已學習到的知識應用於新標籤上。

據我們所知，最先提出零樣本文本分類任務方法的是 Pushp and Srivastava (2017)，作者將文本分類的任務建模爲二元分類問題，查找句子與類別之間的相關性。以這種方式訓練的模型會分別學習每個類別中給定句子的相關性，而不是像多類別多標籤分類中預測給定的類別。使模型學會句子和單詞標籤之間的概念，並且可以在資料集之外進行擴展。

隨著 LLM 的快速發展，有許多擁有大量基礎知識的預訓練語言模型釋出，各種基於 LLM 零樣本分類任務的方法相繼出現，Yin et al. (2019) 提出將零樣本分類任務視爲文本蘊涵問題，他認爲常規文本分類將標籤表示爲

數字，模型不了解標籤所代表的含義。通常人類理解文本和標籤候選詞的含義時，會在心中構建一個假設，將標籤候選詞填入後，判斷這個假設在給定文本下是否成立，此方法是爲了模仿人類判斷，使模型可以從蘊涵資料集中獲取知識。

此外也出現利用 LLM 的提示對仇恨言論做零樣本分類，這種方法是使用提示模板來處理原始文本和類別標籤。Del Arco et al. (2023) 比較 3 種基於 Encoder 的語言模型和兩種指令微調語言模型使用提示模板在八個數據集上做二元分類，其中指令微調語言模型之一 FLAN-T5，在建模階段的訓練資料中包含有害語言的資料集，發現該模型可以從有害語言學到的知識遷移至其他資料集，幫助模型做仇恨言論的分類。Chiu et al. (2021) 則使用 GPT3 做零樣本學習、一次性學習、單一類別少樣本學習和混合類別少樣本學習，四種類型，發現平均準確率在 50-70% 之間，認爲模型可能不太適合前 3 種學習方式，在混合類別學習中表現最佳。

2.4 基於提示的學習

不同於傳統機器學習，傳統機器學習接受輸入 x 並將輸出 y 預測爲 P(y|x)，基於提示的學習是創建一個提示函數 fprompt(x) 的過程，直接對文本概率進行建模，爲了使這些模型執行預測任務，將模板原始輸入 x 修改爲具有一些未填充槽的文本字串提示 $\acute{a}$，語言模型概率性的填充空白訊息以獲得最終字串 $\hat{a}$，由此可獲得最終輸出 y (Liu et al., 2023)。

提示主要有兩種變體，填空和前綴提示。

- 填空提示是填補文本字串中的空白處
  (例如：剛才電影的畫面也太精彩，這是一部 [X] 片)
- 前綴提示是繼續一個字串的前綴
  (例如：剛才電影的畫面也太精彩。這部電影的分類是什麼？[X])

選擇哪一種提示取決於任務和解決任務的模型。對於使用遮罩語言模型來解決的任務，填空提示是一個很好的選擇，因爲它們非常接近預訓練任務的形式。對於生成型任務，前綴提示往往更適合，因爲它們與模型從左到右的特性相融合。

提示模板很大程度上影響 LLM 的表現，在不同類型和風格的提示下會使性能發生巨大變化，爲了使 LLM 發揮出最大的性能，需要詳細的提示模板工程 (Zhou et al., 2022)，提示模板工程的關鍵是有效地將所需任務或意圖傳達給模型，因爲語言模型缺乏眞正的理解能力，在訓練資料中高度依賴模式和聯想 (Brown et al., 2022)。

儘管在許多領域中 LLM 都展現了卓越的性能，然而這些模型很難執行多步計算，尤其是那些需要精確推理的任務，因此 Nye et al. (2021) 建議將複雜的任務拆解成多個步驟和子任務，允許模型在生成最終答案之前，產生任意序列的中間標記，幫助模型完成最終目標。

LLM 在文本分類的任務中，Santu and Feng (2023) 認爲應提供清晰和結構良好的提示，可以引導模型到正確的方向，且避免含糊或籠統的提示，可能導致模型不明確響應。Luo et al. (2023) 則提出添加自定義內容，讓 LLM 完成自動檢測，可以使檢測更緊密符合各種社會的文化。但模型在編寫評語或進一步判斷時，可能會發現錯過的缺陷，並不會評估自己的預測，這意味著他們可以做出預測，但不能提供預測背後的推理 (Saunders et al., 2022)，Wang et al. (2023) 也指出模型生成的解釋有可能導致文本的標籤被錯誤分類，因此在文本分類上，使用二元分類方法會較合適。

3 研究方法

前面的章節概述仇恨言論的零樣本分類、LLM 任務設計和提示等，本研究將採用兩個 Decoder 生成模型使用前綴提示模板，並添加自訂義內容幫助生成模型判斷，作爲比較 Encoder 模型則用少樣本分類當基線，最終觀察各模型之間的差異。

3.1 資料集

本研究的數據來自台灣規模最大的 BBS 論壇 PTT，PTT 是台灣著名的網路社群，與其他社群媒體（如 Facebook、Twitter、Instagram 等）有很大的差異。PTT 的使用者可以使用相對去識別化的方式發表言論，會有更多激進或攻擊性言論出現，其他社群媒體具有更多個人資訊與相關朋友圈的連結，使得使用者在發表言論時更加顧慮，有助於降低仇恨言論的比例。其次，PTT 的分類看板非常多樣，有各式各樣的主題，使用者可以根據自己感興趣的內容到該板進行討論，而其他社群媒體討論的範圍更大許多，大多都沒有進行主題性的分類，很難區別該文章討論的內容。因此選擇 PTT 論壇，其中活躍人數最多的看板「八卦板」，內容大多關注名人八卦、網路熱門話題

標籤	名稱	數量	例句
0	沒有歧視	559	有肥豬肉可以先乾煸出油再爆香佐料
1	歧視爲胖	5,708	那個水桶腰別騙我這隻是豬吧
2	歧視爲瘦	829	台灣島上自己就一堆超越鳥仔腳的吸管腳

表 1: PTT 身體羞辱資料集

等。從該板創立截至 2022/12/20，所有文章底下的留言，共蒐集 34,715,860 筆留言。

儘管 PTT 八卦板的主題性限縮一定程度的範圍，但收集下來的資料集還是非常的發散且龐大，因此在使用前需要萃取相關的文本，以提高文本的相關和準確性。本研究根據分析身體羞辱的語言架構和特徵 (Samarin, 1969; Kraska-Szlenk, 2014)，並參考其國外研究資料集 (Hua, 2018; Reddy et al., 2022; Roodt, 2015)，結合台灣通俗罵法的用詞 (Su et al., 2017)，用自定義辭典的方式，篩選出含有特定關鍵字詞 (例：胖子、肥婆) 的留言，該留言可能爲身體羞辱的文本。根據自定義辭典共篩選出 7,096 筆留言。

在數據標記上，參考 Schlüter et al. (2021) 提及身體羞辱，更具體的現象總稱，包含體重、肥胖或骨瘦如柴的羞辱並且與外表羞辱有一定差別，根據此定義，我們將每個句子的內容加以理解後歸整爲三類，分別爲沒有歧視、歧視爲胖、歧視爲瘦。表 1 爲人工標記後的數量和例句。

考慮到本研究僅專注探討在少數語言或缺乏大量資料集的特定仇恨言論上，不需要大量的標記資料微調，且爲了平衡各標籤數量，從三個類別中各隨機抽取 500 筆資料，共 1,500 筆的資料集，去做後續模型判斷及分析。

3.2 模型及參數調整

爲了比較 Encoder 和 Decoder 模型之間的差異，且避免遇到上述監督式學習的問題 (Poletto et al., 2021)，我們採用少數樣本和零樣本分類兩種方法。且兩種分類方式，皆只使用隨機抽取後的 1,500 筆資料集，這兩種方法旨在減少對標記數據的依賴，以提高模型的泛化能力和應用範圍。

在 Encoder 模型的少數樣本分類中，基線選擇 Hugging Face 用中文資料集微調訓練的 bert-base-chinese，把資料分成 2：8 的訓練集和測試集，BERT 微調參數設定爲 learning rate=2e-5、epochs=5、batch size=64。我們利用有限的標記數據來進行模型的微調，使其能夠更好適應特定的仇恨言論檢測。此外我們還測試在各種比例變化下的訓練集，觀察在本研究資料集中，微調的訓練集數量和最終測試分數分別可能會有的關聯性和結果。

在 Decoder 生成模型上，選擇中國清華大學推出的 GLM (Du et al., 2022a) 和美國史丹佛大學推出的 Alpaca (Taori et al., 2023)，兩者架構下，使用中文資料集訓練和微調的兩種模型，ChatGLM-6B (Du et al., 2022b) 和 Chinese-Alpaca-Plus-7B (Cui et al., 2023)。

ChatGLM-6B 在 1：1 比例的中英資料集上訓練了 1T 的 token 量，兼具雙語能力。且具備 62 億的參數大小，也使得研究者和個人開發者可以自己微調和部署。

Chinese-Alpaca-Plus-7B 通過額外增加 20,000 個中文標記擴展原始 Alpaca 詞彙量，增強中文的編碼和解碼效率，並在 7B-plus 中額外使用 120GB 的資料集，提高中文的理解能力。

在模型中有多種參數可以設定，調整如下

- temperature (溫度)：可以控制文本變化的程度，溫度是介於 0 到 1 之間的參數，降低溫度表示模型會將更多權重放在機率較高的標記上。爲了保持模型判斷回答的一致性，我們把溫度設定在 0.3。
- top_k：這個參數控制模型在生成文本時考慮的詞彙選項數量。較小的 top_k 值會限制模型選擇的詞彙範圍，爲了使模型生成的文本更加精確和可控，我們設定 top_k = 20。
- top_p (或稱爲 nucleus sampling)：這個參數控制模型在生成文本時考慮的累積機率分佈範圍。由於解碼詞還是從頭部候選集中篩選，這樣的動態調整可以使生成的句子在滿足多樣性的同時又保持通順。在文本分類任務中，不需要使生成的句子富有多樣性，因次在這裡設定 top_p =

0.5，表示模型在生成文本時會考慮機率分佈的前 50% 範圍內的詞彙選項。

- num_beams：這個參數決定生成文本時使用的束搜索（beam search）。當設置爲 1 時，表示只保留最有可能的一個文本序列。在這裡，num_beams 設定爲 5，在維持一定程度的計算量上，保持選擇機率最高的五個序列。

3.3 生成模型提示

在生成模型中 Liu et al. (2023) 認爲對於單一樣本執行多標籤任務，直接針對輸入文本定義整體提示是具有挑戰性的。解決問題的直觀方法是將整體提示分解爲不同子提示分別回答，以命名實體識別任務爲例，「有很多隻猴子在動物園裡」這句話中，猴子是生物實體、動物園是地點實體，如直接讓模型判斷兩種類別，原始提示爲：「猴子和動物園各別是什麼實體類型？」，這種方式可能使模型混亂造成精度降低，因此該文獻建議分成兩種子提示，一個提示專注標記一種標籤類別，例如兩個子提示分別爲，「猴子是什麼實體類型？」、「動物園是什麼實體類型？」，爲每個範圍創建不同的提示，分別進行預測提高模型表現。另篇研究中也指出把複雜的任務拆分成更多子任務，會使模型更精準的判斷 (Santu and Feng, 2023)。並且考慮到 Santu and Feng (2023) 提及清晰和結構良好的提示，可以引導模型到正確的方向，綜上所述應在提示中避免使用含糊不清或籠統的字眼，多標籤任務也需逐一分解，因此本研究將分類標籤拆成兩個二元分類子任務，並用前綴提示完成。

第一個二元分類任務是，提供 Schlüter et al. (2021) 完整的身體羞辱定義後，先判斷該文本是否有身體羞辱的歧視，該任務以下簡稱 Q1，如 Q1 判斷有身體羞辱歧視的內容，則會進入到第二個二元分類問題，歧視爲胖還是歧視爲瘦，該任務以下簡稱 Q2。下列爲兩者子任務的完整提示模板。

Q1 身體羞辱定義：[是一種非重複性行爲在這種行爲中加害者不請自來的對目標身體，表達負面意見或評論。加害者不一定有意傷害被害者，但被害者認爲該評論是負面、具有冒犯性或使其產生身體羞恥感。因此身體羞辱的範圍從善意的建議到惡意的羞辱。]
下列文字是一段句子，根據身體羞辱定義判斷該句子是否涉及有關於身體羞辱的歧視? 回答涉及歧視或未涉及歧視
{句子}

原始標籤	**ChatGLM-6B**
無歧視	30
歧視爲胖	25
歧視爲瘦	12
總和	67
	Chinese-Alpaca-Plus-7B
無歧視	123
歧視爲胖	28
歧視爲瘦	21
總和	172

表 2: 在原始標籤中各別判斷錯誤筆數

Q2 身體羞辱可分爲歧視瘦子，跟歧視胖子兩種類型，且兩種類型不會相互重複
下列文字是一段有身體羞辱歧視的句子，判斷該句子是哪種類型，只回答其中之一
{句子}

3.4 模型結果與評估

在生成模型進行文本分類時，由於文本長度、內容等因素，可能會導致回覆內出現更複雜的句子，使其無法明確歸類在現有的標籤之中。這些句子可能包含一些模棱兩可的詞語，詞語包含「可能是、不確定、無法判斷」等含糊不清的內容。這些句子很難被歸類到預先定義的標籤中，因此，在這種情況下，我們將其歸類爲特定的標籤 label：3 (無法判斷)，以表示這些回覆並未明確屬於現有的分類之一。表 2 爲在各標籤中，無法判斷的例子。

我們使用準確率、精確率、召回率和 F1-score 來評估模型在所有情況下的表現。準確率是指在所有樣本中 (包含仇恨言論和非仇恨言論)，被正確分類的樣本數佔總樣本數的比例。精確率是指在所有被預測爲正例的樣本中，有多少是眞正爲正例的比例。召回率是指在所有眞正爲正例的樣本中，有多少被預測爲正例的比例。F1-score 是精確率和召回率的調和平均數，以上指標用於評估模型的性能。表 3 上半部是 Q1 判斷有無身體羞辱，下半部是 Q2 判斷歧視胖瘦。BERT 只需進行文本分類任務，可以直接分類出三種不同類別的標籤，因此在上半部 Q1 空缺。

評估 BERT 模型在不同訓練資料比例下的分數分佈，可以幫助我們了解訓練資料比例對模型性能的影響，找到最低限度的訓練集，讓模型在性能和事前工作（例如資料收集、數據標記等）之間，找到相對平衡的比例，我們使用折線圖將評估結果視覺化，可以清楚地看到模型在不同資料比例下的表現差異。圖

標籤	評估指標	bert-base-chinese	ChatGLM-6B	Chinese-Alpaca-Plus-7B
無歧視	精確率 (precision)	68.21	52.08	26.90
	召回率 (recall)	64.03	33.60	32.60
	F1-score	66.05	40.10	29.48
身體羞辱	精確率 (precision)		72.48	66.06
	召回率 (recall)		84.00	54.30
	F1-score		77.81	59.60
Q1	準確率 (accuracy)		66.87	47.07
歧視爲胖	精確率 (precision)	55.95	46.16	63.32
	召回率 (recall)	68.02	80.60	40.40
	F1-score	61.40	58.70	49.33
歧視爲瘦	精確率 (precision)	71.67	79.35	46.65
	召回率 (recall)	61.11	39.20	37.60
	F1-score	65.97	52.48	41.46
Q2	準確率 (accuracy)	64.33	50.80	36.87

表 3: 所有結果，分數單位皆爲百分比制 (BERT 分數爲 2：8 的訓練集和測試集)

1 爲訓練集從 10% - 90% 之間比例和各標籤 F1-score 之間的關係。

最後，我們將兩個生成模型的混淆矩陣進行比較，提供有關模型預測結果和實際結果之間的對比，幫助我們更全面理解模型的表現，混淆矩陣以四個不同的結果類別爲基礎，分別爲眞陽性：模型正確預測爲正例的樣本數、眞陰性：模型正確預測爲負例的樣本數、假陽性：模型錯誤將負例預測爲正例的樣本數和假陰性：模型錯誤將正例預測爲負例的樣本數，透過混淆矩陣觀察模型在不同類別間的預測情況，我們可以瞭解哪些標籤在模型中容易被判斷錯誤，找出其規律，讓我們在後續做提示模板工程時，有更好的調整方向。圖 2 和圖 3 爲兩者模型的混淆矩陣，x 軸爲預測標籤，y 軸爲實際標籤。

4 結果分析

在本研究的仇恨言論文本分類任務上，使用 20% 訓練集（即 300 筆資料）微調 BERT 獲得最佳表現，相對於另外兩個生成模型，其性能優勢高達 15% 至 30%。唯一在判斷歧視爲胖這個特定類別上，ChatGLM-6B (以下簡稱爲 GLM) 的 F1-score 與 BERT 相差較小。如圖 1 所示，當 BERT 在 10% (150 筆) 資料微調時，表現與 GLM 相當；在 20% (300 筆) 資料時，BERT 稍微勝出；而訓練資料增加到 30% (450 筆) 時，性能顯著提升；並在達到 60% (900 筆) 資料後趨於平緩的上升。

生成模型在兩個任務上都是 GLM 表現較佳，在 Q1 判斷有歧視的 F1-score 爲 77.81%，

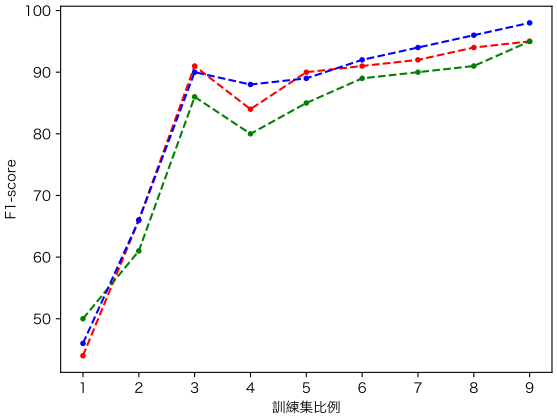

圖 1: F1-score 分佈圖 (三個顏色標籤分別爲，紅：無歧視、綠：歧視爲胖、藍：歧視爲瘦)

Q2 判斷歧視胖瘦分別爲 58.70% 和 52.48%，且該模型在 Q1 判斷無歧視上有著較低的召回率，在圖 2 中也可以明顯看到該模型將多數文本分類爲有身體羞辱的歧視，可能是由於測試資料的特殊性所致。在測試資料中，使用自定義辭典來萃取出特定資料，其中可能包含形容外觀的形容詞，如胖、瘦、肥、乾扁等，而這些形容詞的主詞不一定是指人類。這樣的句構可能導致生成模型在判斷是否有歧視時傾向將這些句子分類爲具有歧視性，因爲它們包含可能被歧視的特定形容詞。GLM 在 Q2 判斷胖瘦的分類裡，對歧視爲瘦的分類採用較嚴格的判斷標準，精確率來到 79.35%，相較於歧視爲胖精確率只有 46.16%，但也因爲判斷較嚴格，該分類的召回率只有 39.20%，且多數判斷錯誤的類別都集中在歧視爲胖的標籤。

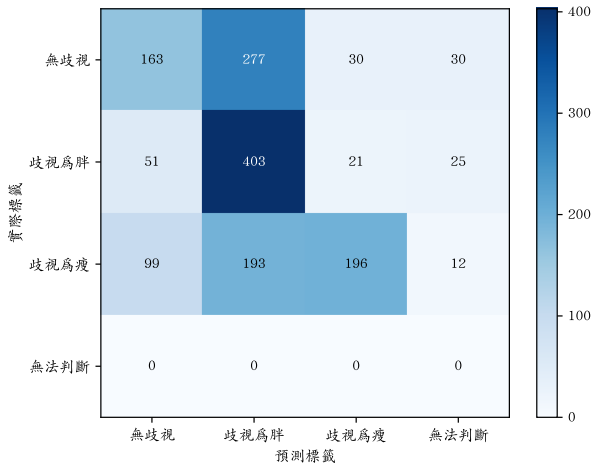

圖 2: ChatGLM-6B 混淆矩陣

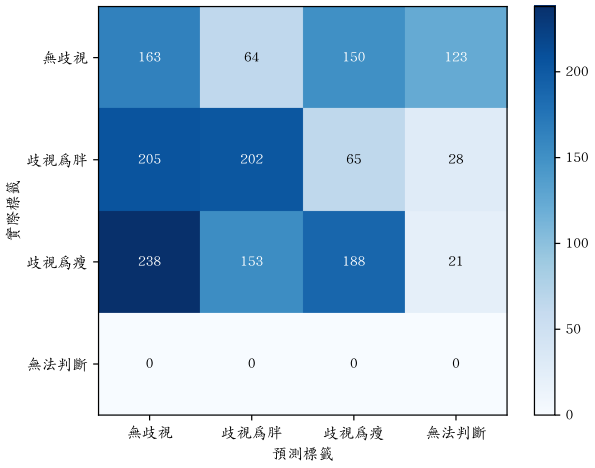

圖 3: Chinese-Alpaca-Plus-7B 混淆矩陣

而在 Chinese-Alpaca-Plus-7B (以下簡稱爲 Alpaca) 模型中，約有 10% 無法判斷的例子，比起 GLM 多出 2.5 倍，推測此模型需要更長的文本內容才能進一步判斷，跟 GLM 不同，Alpaca 偏向把多數文本分類爲無身體羞辱的歧視，接近三分之一的筆數被錯誤分類，進而拉開兩個模型在 Q1 任務上的差距。在 Q2 的任務中，從圖 3 看出，兩者模型呈現相反的判斷趨勢，Alpaca 在判斷歧視爲胖的分類中有著較高的精確率，也因爲 Q1 的判斷不佳，導致歧視胖瘦類別的召回率都很低。很明顯可以看出兩個模型之間在不同任務上的差異，以及它們在特定標籤上的性能表現。

在錯誤判斷的例子中，我們發現一些句子只是在做事實陳述，但模型卻誤判爲具有歧視性。例如：「很瘦骨頭會比較突出就有陰影」，雖然這只是描述骨頭瘦會有陰影的事實，但模型卻認爲有歧視性，且兩者的模型難以抓住句子的主詞，無法判斷該形容詞是在評論物體還是生物，只要句子中出現有胖或瘦等形容詞，例如：「肚子吃不飽，荷包瘦扁扁」，這句話可能只是描述作者自己感到飢餓和荷包空虛，但模型卻無法理解句子的眞正意思，因此錯誤的將其歸類爲歧視性言論。另外，文本中出現譬喻法的句型結構也是模型難以處理的情況。模型無法識別譬喻法是基於外觀上對物品的遷移，這可能是在諷刺受害者的外觀。例如：「甜不辣手指才需要」，這句話在暗示某人的手指很像甜不辣又肥又短，很明顯的具有嘲諷意味。類似的像是「你去花火節會想找龍妹嗎？同理可証！」這樣的句子也出現了相同的情況，僅有使用譬喻法的歧視詞彙無法被模型識別。

此外，我們還觀察到在 Q1 中，明確說出該句是有關於肥胖描述的情況下，模型卻在 Q2 中將其判斷歧視爲瘦。例如句子「死肥婆去減肥」，在 Q1 中模型回應爲「對於死肥婆這個詞語，它被用來形容一個肥滿的女性，並且該詞語被用來對待她，形容是負面評論的一部分。」，然而在 Q2 中卻回覆「這篇文章是歧視瘦子的」。由此可見，模型在解釋和分類判斷上可能會產生出意見相互矛盾的部分。而在處理不齊全的文本時，人類通常會根據個人的經驗和背景知識來補齊文本上下文的大概語意。例如在句子「有個女人願意辛苦懷胎十月，變胖變醜，只爲了...」中，「變胖變醜」只是撐托出懷孕這件事情的困難程度，但生成模型卻可能只會依照文本現有的句子去判斷，而無法理解句子的眞正意圖。

5 結論和未來工作

在本研究的資料集中，使用生成模型並添加詳細定義內容進行零樣本分類的任務，效果似乎不那麼顯著，儘管拆分成兩個二元分類的子任務，使任務簡單化，最終歧視爲胖瘦的 F1-score 也只有 58.70% 和 52.48%。就算只執行 Q1 任務，有歧視表現較好的 F1-score 雖然有 77.81%，但在無歧視的分數卻只有 40.10%。此外我們的資料集是來自 PTT 社群平台，該平台有一個特性是會限制每段留言長度，超過長度的留言會被自動換行，導致收集下來的文本長度過短，產生上述結果分析時的問題，這些結果表明生成模型在此任務中處理具有多義性、複雜結構或隱含意味的文本時，容易出現錯誤判斷。在處理含有歧義性、不完整的文本，可能出現意見相互矛盾的情況，且無法像人類一樣根據個人經驗進行推理和理解。

這些問題說明在零樣本分類上，對生成模型來說確實是有些困難，跟 Chiu et al. (2021) 研究中提到使用 GPT-3 零樣本分類結果大致相同，不過在混合標籤少樣本上有優於零樣本分類的表現，然而此作者的提示裡並未提供更詳細的分類定義，這可能導致模型在處理文本時表現受限。未來的方向可以進一步探討定義分類標準，將其結合少樣本學習的方法，在提供身體羞辱定義之外，新增不同句構的句型並詳加解釋，研究不同類型的樣本來觀察模型表現的變化。這些挑戰突顯生成模型在歧視言論辨識任務上的限制，並顯示改進提示模板工程的重要性，可以進一步優化將有助於提高模型的理解能力和鑑別能力，以更好處理現實中的語言。

Encoder 和 Decoder 兩者架構下的預訓練語言模型在本研究的任務中，BERT 經過微調後的性能顯著優於生成模型，即便在訓練資料只有 150 筆時，BERT 已經達到與生成模型相當的水準，在文本長度受限的情況下，BERT 較能從中獲取具有歧視性的關鍵字詞。對於少數語言或缺乏相關大量資料集的主題來說，或許重點可以放在研究 BERT 最少需要多少訓練資料進行微調，才能在性能上大幅提升。我們可以找到一個平衡點，在適度的資料量下獲得最佳的性能。

根據 ChatGPT 開發者表示，生成模型會產生歧視言論的問題，解決方式是通過在不適當內容的資料集上訓練另一個機器學習模型，來刪除或檢測攻擊性言論，此外也有由人工審核團隊在內容公開之前對其進行審查和批准 (Meyer, 2022)。但由於 OpenAI 缺乏透明度，我們無法驗證該說法是否屬實。不過該想法也是建立在需要大量的仇恨言論標記資料集上，是否能把這特定資料集的仇恨言論知識遷移到其他少數語言或其他未被包含在內的主題需要更進一步研究。

References

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. What does it mean for a language model to preserve privacy? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2280–2292.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ke-Li Chiu, Annie Collins, and Rohan Alexander. 2021. Detecting hate speech with gpt-3. *arXiv preprint arXiv:2103.12407*.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Flor Miriam Plaza Del Arco, Debora Nozza, and Dirk Hovy. 2023. Respectful or toxic? using zero-shot learning with language models to detect hate speech. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 60–68.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022a. GLM: general language model pretraining with autoregressive blank infilling. pages 320–335.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022b. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online*, pages 85–90.

Philipp Hacker, Andreas Engel, and Marco Mauer. 2023. Regulating chatgpt and other large generative ai models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1112–1123.

Tan Kim Hua. 2018. Cyberbullying: A cursory review. *Tan Kim Hua*, page 17.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, page 126232.

Iwona Kraska-Szlenk. 2014. Semantic extensions of body part terms: Common patterns and their interpretation. *Language Sciences*, 44:15–39.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Chu Fei Luo, Rohan Bhambhoria, Xiaodan Zhu, and Samuel Dahan. 2023. Towards legally enforceable hate speech detection for public forums. *arXiv preprint arXiv:2305.13677*.

Patrick Meyer. 2022. Chatgpt : How does it work internally?

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. 2021. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523.

Pushpankar Kumar Pushp and Muktabh Mayank Srivastava. 2017. Train once, test anywhere: Zero-shot learning for text classification. *arXiv preprint arXiv:1712.05972*.

Varsha Reddy, Harika Abburi, Niyati Chhaya, Tamara Mitrovska, and Vasudeva Varma. 2022. 'you are big, s/he is small' detecting body shaming in online user content. In *Social Informatics: 13th International Conference, SocInfo 2022, Glasgow, UK, October 19–21, 2022, Proceedings*, pages 389–397. Springer.

Kyra Roodt. 2015. *(Re) constructing body shaming: Popular media representations of female identities as discursive identity construction*. Ph.D. thesis, Stellenbosch: Stellenbosch University.

William J Samarin. 1969. The art of gbeya insults. *International Journal of American Linguistics*, 35(4):323–329.

Shubhra Kanti Karmaker Santu and Dongji Feng. 2023. Teler: A general taxonomy of llm prompts for benchmarking complex tasks. *arXiv preprint arXiv:2305.11430*.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*.

Constanze Schlüter, Gerda Kraag, and Jennifer Schmidt. 2021. Body shaming: An exploratory study on its definition and classification. *International Journal of Bullying Prevention*, pages 1–12.

Hui-Po Su, Zhen-Jie Huang, Hao-Tsung Chang, and Chuan-Jie Lin. 2017. Rephrasing profanity in chinese text. In *Proceedings of the First Workshop on Abusive Language Online*, pages 18–24.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Han Wang, Ming Shan Hee, Md Rabiul Awal, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. 2023. Evaluating gpt-3 generated explanations for hateful content moderation. *arXiv preprint arXiv:2305.17680*.

Yau-Shian Wang and Yingshan Chang. 2022. Toxicity detection with generative prompt-based inference. *arXiv preprint arXiv:2205.12390*.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161*.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*.