**應用機器學習於生成評論解析以OpenAI為例**

**資料集:現有資料集(GPT2所生成的生成資料集以及寫手資料集)**

# 第一章 緒論

## 研究背景與動機

隨著網路與科技的快速進步，人們只需在家動動手指便能透過網路快速獲取大量商品資訊；網路的便利性以及多樣性深受大眾喜愛，帶動電子商務平台的興起，使得「網路購物」成為現代人主要的消費習慣之一(Gowripeddi Hari Kumar et.al , 2022)。然而，網路購物的缺點在於消費者無法實際接觸商品，從而容易產生買賣雙方資訊不對稱的問題，進而影響到消費者在進行購買時所做的決策。

為了改善這個問題，現在各大電子商務平台，例如Amazon、eBay及蝦皮購物等平台，皆設有評論機制，此機制允許購買過商品的消費者，分享使用心得並給予評價；Schlosser在2005年的研究指出消費者會上網搜尋查找其他消費者的相關使用心得作為購物時的參考依據(Schlosser, 2005) ；當消費者想要在網路上選購商品時，除了查看產品特色以及技術規格以外，還能透過其他買家所分享的評論來了解產品的實際體驗心得與其性能表現。當評論反映出較正面的評價能提高消費者購買產品的信心，反之，當看到評論反映產品存在瑕疵或體驗不佳等問題，這類負面反饋可能會使消費者重新評估，考慮是否要購買產品。

隨著評論機制的普遍應用，近年來，各電子商務平台的商品評價中開始存在「灌水」的疑慮，一些不法賣家會僱用人類寫手發佈不實評論，這類評論不僅對於平台的信譽帶來了負面影響，也對消費者的購買決策帶來干擾，這些虛假評論對於平台的信譽帶來了負面影響。

為了遏制虛假評論帶來的負面影響，多個平台已經開始採取積極行動，致力於開發與實施各類檢測系統。例如，Trustpilot 在 2021 年檢測並移除了 270 萬條虛假評論，並通過不斷投資於欺詐與異常檢測軟件，顯著提升了虛假評論的識別和移除效率。同時，學術界也提出了多種檢測虛假評論的方法。例如，Chowdhary 和 Pandit在2018年的研究中應用了隨機森林演算法，結合使用者行為特徵工程來構建監督學習模型，有效提升了對虛假評論的識別能力（Chowdhary & Pandit, 2018）。Elmogy等人在2021年的研究中採用增強演算法，並將使用者評論頻率等特徵應用於邏輯回歸模型，成功提升虛假評論的檢測效率（Elmogy, Tariq, Ibrahim, & Mohammed, 2021）。。這些技術手段皆致力於實現更高的檢測精度，以應對虛假評論的蔓延問題。

儘管前者檢測虛假評論的技術正不斷進步，但自然語言處理技術同時也正快速發展；生成模型如GPT-3、LlaMa2等，在創建文本、圖像、語音及影片等方面，都取得了大幅度進步，生成式技術的革新在提升人們的生產力及創作力有著巨大潛力，但同時也帶來了 AI 生成內容被冒充為人類創作的風險，因為這些技術如今已能生成看起來和真實人類所撰寫的評論幾乎無異的內容。表1.1列出3則AI所生成的評論，可以發現這些評論與人類所撰寫的評論在寫作風格及內容上非常相似，這也使得大眾難以分辨評論是否可信，讓平台的公平和誠信再次面臨挑戰。

表1.1 AI生成評論示例

|  |  |
| --- | --- |
|  | AI生成的評論 |
| 1 | Sugarfish exceeded my expectations! The sushi was fresh, simple, and of good quality. The value is great, and the service was friendly and helpful. The sake we ordered was delicious, not too sweet. I would definitely return in the future. |
| 2 | I came to the Museum of Jewish Heritage for their show and was really pleased with the performance. I was also able to enjoy the museum, which is always a plus. However, when it came to brunch, the service was great, however the view was not as great. The food was amazing, though. The restaurant is open during museum hours and I would recommend it. The portions for French toast were a bit small, but I was able to get a large bagel and lox. Overall, I thought the experience was amazing and would definitely come back. |
| 3 | I had a great time at Vibes! The service was great and the cocktails were delicious. I especially loved the roasted octopus and the prawns, pepper croquettes and mushroom crepes. The music was at the perfect level so you could still have a conversation but still feel like you were in a lively spot. I will definitely be back here soon! |

雖然AI 生成評論與人類撰寫的評論非常相似，但還是有跡可循的，例如AI生成評論中經常包含誇張的情感表達；這些特徵將可以用來識別 AI 生成的內容，平台可以考慮利用這些傳統的文本特徵，作為識別和管理虛假評論的方法。

目前已有大量關於評論檢測的研究，但大多數文獻仍集中於檢測由人類寫手生成的虛假評論，對於檢測自動生成評論的研究相對較少。因此，本研究旨在填補這一空白，專注於檢測由生成模型(如GPT)生成的評論。我們將透過分析生成評論的特徵，並結合傳統人類評論的語言特徵，評估這些特徵對檢測性能的影響。此外，我們還將探討不同特徵組合的效果，以提高對生成評論的檢測準確性，從而增強消費者對網路購物的信任。

## 研究目的

當前評論發展現況中，生成評論對消費者購買決策的影響日益顯著，這些評論可能誤導消費者，使其對商品的真實評價產生偏差，進而影響他們的購買決策。因此，如何有效辨識生成評論，對於保護消費者權益及維護電子商務平台的誠信至關重要。

基於此研究背景與動機，本研究將收集包含生成評論和人類撰寫評論的資料集，並運用自然語言處理(Natural Language Process, NLP)技術進行文本分析。透過提取多種單一類型特徵，評估這些特徵對檢測模型準確率的影響，進而建構一個專門用於檢測生成評論的模型。此模型旨在協助電子商務平台更有效地識別並移除生成評論，從而提升消費者信任度，維護市場的公平性。

綜上所述，本研究的目的可歸納為以下兩點：

* 探討多種單一類型特徵在生成評論檢測中的效果，評估其對檢測模型準確率的影響，並透過特徵組合找出最佳的特徵組合策略。
* 建構生成評論檢測模型，使用可解釋性指標驗證該模型的有效性，提升其在實際應用中的信賴度和使用價值。

## 研究架構

本研究之研究架構分為五個章節，研究架構如圖1.1所示，各章節內容如下：

第一章 緒論

說明本研究之研究背景與動機、研究目的、研究架構以及研究限制。

第二章 文獻探討

本研究首先概述與統整生成模型的應用，特別是生成文本或評論中的應用情境。接著探討機器學習在評論辨識中的應用，總結相關技術在過去研究中的使用情況及其效果。最後深入探討影響生成評論辨識的各種因素並進行歸納。

第三章 研究方法

說明本研究的研究流程，包含資料蒐集、資料前處理、特徵提取、資料轉換、模型選擇以及模型評估及優化。

第四章 實驗設計與結果

根據本研究所提出之研究方法進行實驗設計，並依據實驗結果進行解釋與分析。

第五章 結論與未來展望

歸納本研究實驗結果，並提出研究限制與未來展望。

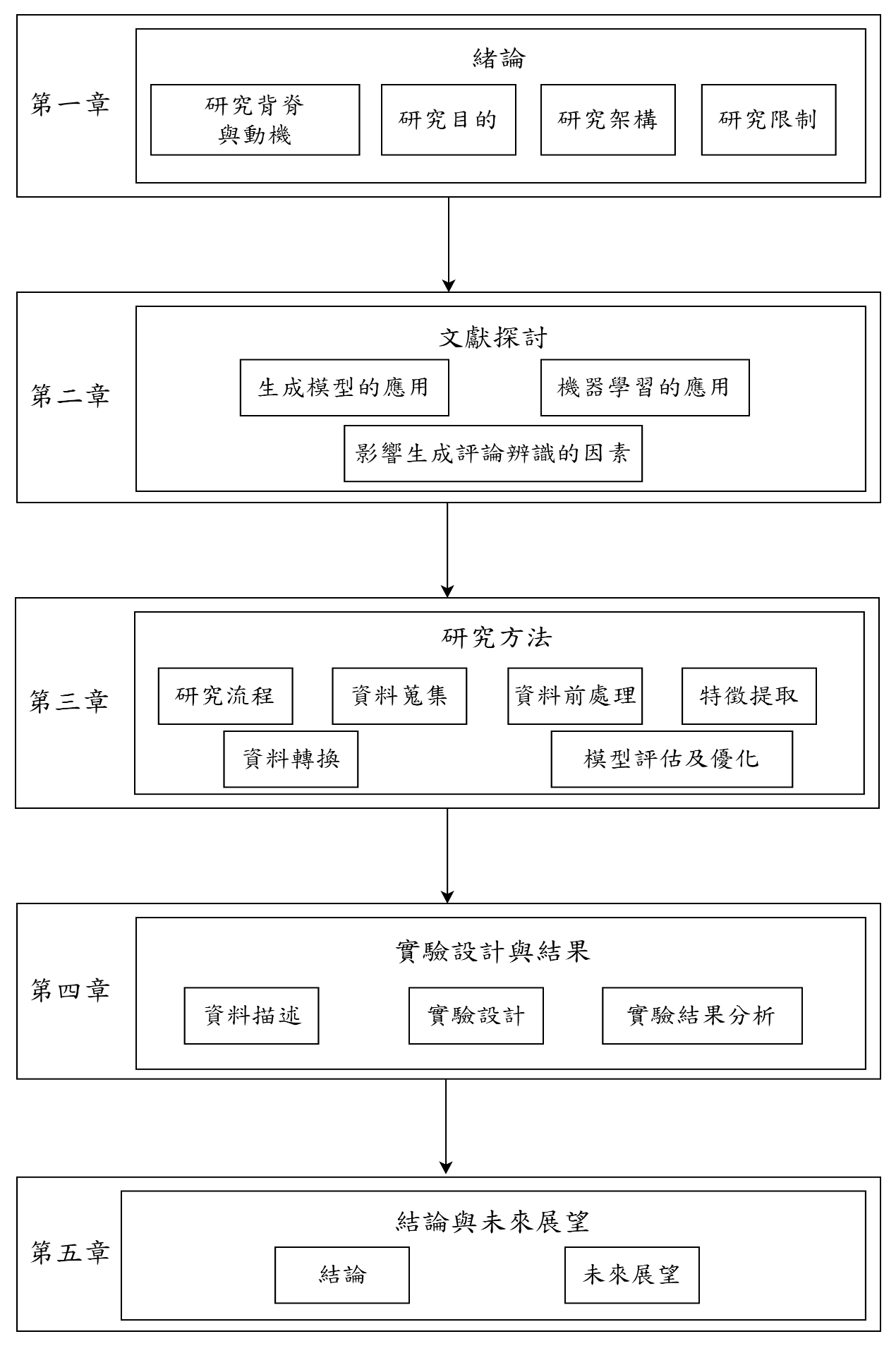


圖1.1 研究架構圖

## 研究限制

# 第二章 文獻探討

隨著資訊科技的快速發展，生成模型已被廣泛應用於各種文本生成任務，包括生成評論，尤其是在電子商務及社交平台上，這類評論可能對消費者的購買決策產生影響，因此本研究將應用文字探勘技術與機器學習模型進行特徵提取與組合，並訓練模型以預測生成的評論。本章節將探討過去文獻中關於生成模型的應用、機器學習在評論辨識中的應用，以及生成評論的特徵與辨識相關的研究。

## 生成模型的發展與應用

生成模型的概念最早是由Baum等人在1970年提出的隱藏馬爾可夫模型(Hidden Markov Model, HMM)，HMM通過對序列數據進行建模，推斷隱藏狀態的概率分佈，並依據這些隱藏狀態生成與預測未來的觀測數據。(Baum, Petrie, Soules, & Weiss, 1970)。這種依賴數據的概率分佈進行生成的特性，也被應用在早期的語音辨識上 (Rabiner, 1989)。

然而，HMM仍存在一些限制：HMM的假設是每個狀態之間的轉移是由馬爾科夫過程所控制，表示當前狀態僅依賴於前一個狀態，因此在處理更複雜的時間序列數據時可能過於簡單，無法捕捉長期依賴性，使得HMM在處理高維數據或是長序列數據的效果有限。

由於前述限制，學者們開始尋求生成模型的創新。隨著技術的發展， Kingma和Welling在2013年提出變分自編碼器(Variational Autoencoder, VAE)，將概率分佈引入自編碼器(Autoencoder)的框架中，使編碼器將輸入數據映射到隱含空間時不再只是單一點，而是表示一個概率分佈，因此解碼器在進行採樣時能夠生成多樣性的數據樣本(Kingma & Welling, 2013)；Goodfellow等人在2014年提出了生成對抗網路(Generative Adversarial Network, GAN)，此模型由兩個神經網路所組成：生成器及判別器。透過兩者的對抗學習，生成器不斷改進其生成策略，而判別器則不斷提升其的辨別能力，最終使模型能夠生成高度真實的數據(Goodfellow et al., 2014)。

隨著VAE及GAN的發展，生成模型在圖像生成(Radford, Metz, & Chintala, 2016)、語音合成(Hsu、Zhang, & Glass, 2017)、文本生成(Yu, Zhang, Wang, & Yu, 2017)以及風格轉換(Isola, Zhu, Zhou, & Efros, 2017)等任務中皆得到了廣泛應用，然而，生成模型的發展不止於此。生成預訓練模型(Generative Pre-trained Transformer, GPT)的出現是生成模型技術的另一重大突破。

OpenAI於2018年提出了GPT模型，該模型基於Transformer架構首先在大規模文本數據上進行預訓練，接著根據具體任務進行微調以優化模型，這種結合預訓練與微調的策略也使GPT在各個領域都有出色的表現(Radford, Narasimhan, Salimans,& Sutskever, 2018)。隨著GPT-2(Radford et al., 2019)、GPT-3(Brown et al., 2020)以及最新的GPT-4的推出，在文本(Brown et al., 2020)、語音(Jia et al., 2018)及圖像(Ramesh et al., 2021)等其他應用領域皆取得了顯著的進步。

除了OpenAI，其他科技公司如Google、Meta也提出類似的模型；Google的Gemini、Meta的LLaMA等，這些模型進一步推動生成技術的發展，不僅提升生成的品質，也顯著改善了模型在各類任務中的泛化能力，包括對話系統及個性化推薦等。

## 生成評論的特徵

在檢測生成評論的相關文獻中，發現生成評論的特徵可由兩大類組成：大型語言模型的生成模型特徵及傳統語言特徵(Guo, Zhang, Wang, & Jiang, 2023)。

生成模型特徵源自於大型語言模型（LLM）的生成機制，這些模型根據詞彙的預測概率選擇詞語，使生成的文本具備較高的可讀性，同時也帶來結構呈現模板化的缺點（Holtzman, Buys, Du, Forbes, & Choi, 2020），而人類所撰寫的文本則具備變化性。這類特徵可以通過衡量文本的隨機性與一致性來進行識別，包含困惑度（perplexity）和突發性（burstiness）等指標。Ippolito 等人在2020年的研究利用困惑度來評估AI生成文本的可讀性和一致性，並有效辨識AI生成的文本。Clark 等人在2021年的研究利用突發性評估文本的隨機性，從而區分人類撰寫的文本與AI生成的文本。

傳統語言特徵則專注於文本內容本身，涵蓋從語言的基本結構、詞彙選擇到情感表達等多方面，能夠有效地描繪文本的語義與語法結構。這些特徵過去經常被應用於檢測評論中；Waseem 和 Hovy(2016)利用語義特徵成功檢測來自Twitter的仇恨評論（Waseem & Hovy, 2016）；Ott 等人(2013)結合語義特徵及句法特徵來辨別虛假評論（Ott, Cardie, & Hancock, 2013）；Hu和 Liu(2004)利用情感特徵分析產品評論，總結消費者對產品的整體評價及反饋（Hu & Liu, 2004）。傳統語言特徵不僅有助於理解文本所傳達的內容，還能更精確地捕捉文本的細微差異。

## 機器學習與深度學習應用於評論檢測

 本節彙整了5篇關於檢測生成評論的相關文獻，如表X所示，針對各論文所使用到的比較模型、當中的最優模型、使用到的評估指標，以及論文貢獻進行簡要介紹，並將這些文獻中使用的特徵進行統整。Alamleh等人（2023）的研究使用TF-IDF向量化技術作為詞彙特徵，並透過隨機森林、支持向量機等分類模型進行AI生成評論偵測，結果顯示隨機森林模型表現最佳，且利用TF-IDF可以捕捉文本中的重要詞彙，從而提升模型對文本的區分能力(Alamleh et al., 2023)。

Cingillioglu等人(2023)使用N-gram模型進行詞彙特徵提取，並透過支持向量機進行分類，為了驗證分類結果的表現，作者將其與其他工具進行比較(如GPTZero)，結果顯示雖然支持向量基在總準確率的表現略遜於其他工具，但其達到最佳的召回率及F2分數。(Cingillioglu et al., 2023)。

Luo等人（2023）的研究提出了一個監督式混合概率方法來檢測AI生成評論。研究使用多種語言特徵，包含語義特徵、情緒特徵和可讀性特徵等，並引入大型語言模型的生成特徵，包含困惑度及突發性，這些特徵的結合能夠更全面地捕捉AI生成評論與人類撰寫評論之間的差異。在分類模型方面，研究採用隨機森林、支持向量機、AdaBoost等模型，結果證明這種多特徵的整合方法能夠有效提升分類模型在檢測AI生成評論方面的效能(Luo, Nan, Li, & Tan, 2023)。

表2.X機器學習與深度學習於AI生成評論檢測研究統整表

|  |  |  |  |
| --- | --- | --- | --- |
| **作者** | **使用特徵** | **使用模型** | **評估方式** |
| (Alamleh et al., 2023) | 詞彙特徵 | 邏輯斯回歸、  決策樹、  支持向量機、  神經網絡、  隨機森林、  BERT | 準確性、混淆矩陣、運算效率 |
| (Bader et al., 2023) | 詞彙特徵、情感特徵 | 決策樹、  隨機森林、  支持向量機 | 準確性、精確率、召回率、F1分數 |
| (Cingillioglu, 2023) | N-gram | 支持向量機(SVM) | 準確率、精確率、召回率、F2分數 |
| Jiwei Luo, Guofang Nan, Dahui Li, Yong Tan (2023) | 傳統語言特徵、新穎生成特徵 | K近鄰算法、邏輯斯回歸、 多層感知器、  隨機森林、支持向量機、AdaBoost | 準確性、精確率、召回率、F1分數、└AUC分數 |
| Gambetti及Han(2023) | 文本特徵、用戶行為特徵、寫作風格 | 朴素貝葉斯、隨機森林、XGBoost、羅吉斯回歸、BILSTM | 準確性、精確率、召回率、F1分數 |

由於針對生成評論的研究較為有限，本研究也參考了其他與任務相似的評論檢測文獻，並彙整如表X。

# 第三章 研究方法

本研究旨在透過分析生成評論，建構出一個能準確辨識生成評論與人類撰寫評論的檢測模型。此章節將分成研究流程、資料蒐集、資料前處理、特徵萃取、資料轉換、模型選擇與模型評估及優化七個部份，並進行詳細說明。

## 研究流程

## 資料蒐集

本研究所使用的資料集為combat-ai-restaurants[[1]](#footnote-2)及FakeReviews[[2]](#footnote-3)，一份作為本研究模型訓練的資料集，另一份則作為最終模型的泛化評估資料集。

combat-ai-restaurants資料集是由Gambetti及Han建立，並公開於Github平台。該資料集包含兩部分:一部分為來自Yelp.com的人類撰寫評論，時間範圍為2021年1月至2022年7月，另一部份為基於這些評論，利用GPT-3預訓練模型所生成的生成評論。該資料集共計23755筆。本研究將利用此資料集對模型進行訓練以及評估。

FakeReviews資料集是由Salminen等人建立，並公開於Github平台。該資料集包含兩部分:一部分為來自Amazon平台的人類撰寫評論，時間範圍為2018年1月至2018年12月，另一部份為基於這些評論，利用GPT-2預訓練模型所生成的生成評論。該資料集共計40000筆，本研究隨機抽取其中 4,000 筆，用於最終模型的泛化評估。

## 資料前處理

本章節將描述對資料集中的資料所進行的清理、斷詞斷句以及去除停用詞的步驟。下圖3.2為本研究資料處理之流程圖。

圖3.2 資料處理流程圖

* + 1. **資料清理**

本研究的資料清理過程中，我們將針對幾個部分進行處理，以確保資料品質及分析結果的準確性。

首先，由於本研究主要目的是分析英文文本，因此所有非英文的文本將被過濾掉，以避免語言混淆對模型訓練和後續分析的影響。其次，考慮到資料集可能存在重複的評論，為確保分析的準確性，過濾這些重複的內容是必要的。接著為確保不同字數區間的評論數量平衡，本研究對每個區間中評論數量明顯較多的類別進行隨機抽樣，以達到兩個類別在各字數區間擁有相同比例的評論數。最後，基於本研究僅使用評論內容（text）和標籤（label）兩個欄位，對FackReview資料集將所有非必要的欄位，如產品分類(Category)及產品評分(rating)進行清除，以確保資料的簡潔性和分析的有效性。

經過以上檢查並處理後，本研究所使用的兩份資料集combat-ai-restaurants及FakeReviews分別剩餘14854筆及7588筆資料。其中combat-ai-restaurants資料集包含7266筆生成評論及7588筆人類撰寫評論；FakeReviews 資料集則包含 2,000 筆生成評論與 2,000 筆人類撰寫的評論。

* + 1. **斷詞斷句及去除停用詞**

目前關於自然語言處理(NLP)的研究中，對於英文文本的處理，常使用工具如自然語言工具集(Natural Language Toolkit, NLTK)、SpaCy和Stanford NLP進行斷詞斷句和去除停用詞。這些工具提供了包括斷句(Sentence Segmentation)、斷詞(Tokenization)、詞性標註(POS Tagging)、命名實體識別(Named Entity Recognition, NER)以及停用詞過濾(Stopword Filtering)等功能。

而在英文文本的處理過程中，斷詞斷句是首先要進行的步驟，將連續的文本分解為獨立的單字，並移除標點符號。接著進行停用詞的過濾，停用詞是指在文本中頻繁出現但對語義貢獻較小的詞語，例如 "the"、"is"、"and"等，去除停用詞將有助於降低資料的維度，從而提升模型訓練的表現。

本研究使用NLTK進行文本的斷詞斷句，並根據NLTK預設的停用詞列表來刪除停用詞。表3.1顯示了斷詞斷句與去除停用詞的範例。

表3.1 文本斷詞斷句及去除停用詞範例

|  |  |
| --- | --- |
| 原文 | I tried the "famous" Mac and Cheese burger at this restaurant and it was a huge disappointment. |
| 斷詞斷句 | 'i', 'tried', 'the', "'famous'", 'mac', 'and', 'cheese', 'burger', 'at', 'this', 'restaurant', 'and', 'it', 'was', 'a', 'huge', 'disappointment', '.' |
| 去除停用詞 | 'tried', 'famous', 'mac', 'cheese', 'burger', 'restaurant', 'huge', 'disappointment' |

## 特徵提取

特徵提取是指為了萃取具有代表性的特徵，將原始數據從高維數據映射至一個低維的特徵空間的過程，維度的降低不僅幫助降低數據處理的複雜性，減少計算成本，也加深模型對數據的理解，從而提升模型的學習成效。

本研究旨在探討文本特徵對於檢測AI生成評論的影響，通過分析各種類型的特徵，確定何種特徵對於提高檢測準確性更為關鍵，本研究在文本特徵的分類上採用了Du等人(2019)及林維婕(2024)年所提出的研究，分別為語義特徵(Semantics)、情緒特徵(Sentiment)、可讀性特徵(Readability)、結構特徵(Structure)及句法特徵(Syntax)五大類，另外本研究針對生成評論，在五大類特徵的框架下進一步進行特徵集的擴展，圖3.X為本研究所採用的文本特徵類別及每一類別所包含的具體特徵值，本小節將對此進行詳細說明。

圖3.X 文本特徵類別及各類別所包含的具體特徵值

1. 語義特徵(Semantics)

語義特徵是指利用統計方法將文本中的詞彙轉換成能夠反映其語義和上下文資訊的數字向量，以從評論內容中提取單字和主題概念的意義。本研究的語義特徵類別包含以下七個特徵。

1. Unigram:

N-Gram模型是一種統計模型，它通過將文本按固定長度N切分，從而形成長度為N的字詞片段序列，每個序列稱為一個“gram”。這些字詞片段被用作計算條件機率的基礎，使模型能夠基於先前出現的字詞來預測接下來可能出現的字詞，以此來分析和預測文本內容的結構和連貫性。

Unigram是N-Gram模型中最基本的形式， N設定為1。在此模型中文本被處理成一系列獨立的單字，不涉及單字之間的相互依存性和語義關係，而Unigram也因其能夠提供有效的關鍵字訊息而被廣泛應用於文本分類和搜索引擎的關鍵詞搜索等領域。

本研究對每則評論文本進行分析，利用Unigram模型選出頻率最高的三個單字，隨後通過TF-IDF方法將其轉換成詞向量，作為本研究語義特徵之一。

1. Bigram:

Bigram是N-Gram模型中的另一種形式， N設定為2。在此模型中文本被處理成一系列相鄰的單字對，捕捉單字之間的直接依存性和語義關係，因此被廣泛應用於需要考慮詞序的場景，以提高文本分析的準確性。

本研究對每則評論文本進行分析，利用Bigram模型選出頻率最高的三個單字組合，隨後通過TF-IDF方法將其轉換成詞向量，作為本研究語義特徵之一。

1. 潛在語義分析(Latent Semantic Analysis, LSA):

LSA為Deerwester等人在1990年所提出的一種線性代數模型，其基本原理是首先建立一個文本與單字間的矩陣：矩陣的行代表單字，列代表文本，矩陣中的元素則反映詞彙在文本中的出現頻率或其他相關的加權值。接著，透過奇異值分解（Singular Value Decomposition，SVD）對該矩陣進行降維處理，LSA會捨棄較小的奇異值，因為這些奇異值通常反映較低的訊息量，從而降低數據的噪聲，更有效地找出文本的潛在語義關係。

Landauer和Dumais 在1997年的研究中展示了如何利用LSA進行主題建模，他們通過捕捉詞彙與文本之間的潛在關係來分析整個文本。在這個過程中，LSA 模型的每個維度或方向都被解釋為代表文本中的一個隱含主題。

本研究利用LSA對每則評論文本進行分析，主題數設為5，從而生成一個五維的向量空間。這五個維度各自代表一個隱含主題，隨後將這些維度轉換成向量，作為本研究語義特徵之一。

1. 潛在狄利克雷分配(Latent Dirichlet Allocation, LDA):

LDA為Blei等人在2003年提出的一種生成式概率模型。假設每篇文本是由數個主題組成，而每個主題又由數個詞彙組成，並引入狄利克雷分布(Dirichlet Distribution)作為文本的主題分布，使LDA能夠以非監督的方式從文本中提取出潛在的主題結構。

本研究利用LDA對每則評論文本進行分析，並將主題數設為5，生成一個五維的向量空間。這五個維度各自代表一個隱含主題，隨後將這些維度轉換成向量，作為本研究語義特徵之一。

1. GloVe:

GloVe，全名為Global Vectors for Word Representation，是Pennington等人在2014年所提出的一種詞向量模型。模型通過學習語料庫中的詞共現關係，將文本中的每個詞彙映射到固定維度的向量空間。在這個空間中，詞彙之間的語義相似性可以通過它們之間的距離來表示，越相近的詞彙距離越近。透過這個方法，模型能夠有效捕捉詞彙間的語義及共現關係並應用到後續的文本分析任務。

本研究採用來自GloVe官方所提供的詞向量語料庫，其中向量的維度選擇有50、100、200及300，為能夠捕捉更精確的語義關係，本研究採用300維度的語料庫，作為語義特徵之一。

1. 困惑度(Perplexity):

困惑度是一種用來評估語言模型好壞的指標，衡量模型對於測試資料的預測能力。具體而言，困惑度反映語言模型在給定序列中選擇下一個字的平均不確定性。困惑度的計算公式如式3.X所示。

|  |  |
| --- | --- |
|  | 公式3.X |

其中代表的是測試語料庫中所有詞彙的數量，模型代表的則是對第i個詞的預測機率。困惑度越低表示模型對詞語的預測越精確。

由於困惑度能反映模型對於語義結構的理解程度，從而幫助分析文本中的語義關係，因此本研究使用困惑度特徵作為語義特徵之一。

1. 突發性(Bursiness):

突發性指的是某些詞彙或語句在文本中突然出現的可能性，高突發性的文本一般伴隨著較大的語義變化，表示在短時間內或特定的上下文中，關鍵詞或語句會頻繁出現，從而引發顯著的語義變動。公式如式3.X所示。

|  |  |
| --- | --- |
|  | 公式3.X |

其中，代表的是某一詞彙在目前段落中的出現頻率，表示的則是在整個文本中，某一詞彙的平均出現頻率。由於突發性能夠有效捕捉文本中的語義變化及異常狀況，本研究將其作為語義特徵之一。

1. 情緒特徵(Sentiment)

情緒特徵是指從文本中提取出能夠反映情感或情緒相關的特徵，用來理解及分析文本的情感內容；快樂、悲傷、憤怒、驚訝等。而本研究採用英文文本中常用的情緒詞典及語意套件作為情緒特徵類別，共包含三個特徵。

1. SentiWordNet

SentiWordNet 是 Esuli 和 Sebastiani 在 2006 年提出的一個基於 WordNet 構建的情感分析詞典。該詞典首先通過手動標註部分詞義為正向、負向或中性，以此作為初步訓練數據。隨後，應用半監督學習方法來推斷其他未標註詞義的情感傾向。透過訓練分類器，SentiWordNet 能夠自動預測每個詞義的正向分數（PosScore）、負向分數（NegScore）及中性分數（ObjScore）。這些分數的範圍介於 0 到 1 之間，且三者總和為 1

本研究使用SentiWordNet對每則評論文本計算情感分數，作為情緒特徵之一。

1. General Inquire

General Inquirer是Stone等人在 1966 年提出的ㄧ個文本分析工具，首先將大量詞彙手動分配到不同的類別中，這些類別包括正向情感、負向情感、社會過程、認知過程等，並被用作分析文本的基礎。當研究者分析文本時，General Inquirer會計算文本中屬於各個類別的詞彙出現頻率，從而推斷該文本在情感和心理方面的傾向。

本研究使用General Inquirer對每則評論文本計算情感詞頻，作為情緒特徵之一。

1. LIWC

LIWC，全名為Linguistic Inquiry and Word Count，是Pennebaker等人在 2001年提出的文本分析工具。LIWC明確定義了各詞彙類別及其所包含的詞彙，其中有些類別之間具有層次結構；情感詞彙類別可細分為正向情緒詞(posimo)及負向情緒詞(negemo)，而負向情緒詞又可進一步細分為焦慮(anx)、生氣(anger)和悲傷(sad)三個子類別。另外每個詞彙可以同時屬於多個不同的類別，因此LIWC能夠精確地捕捉到文本中的情感細微差別。

本研究使用LIWC對每則評論文本計算情感詞頻，作為情緒特徵之一。

1. 可讀性特徵(Readability)

可讀性特徵是衡量文本是否易於讀者理解的指標，通常基於單詞長度、句子長度以及音節數等因素，從客觀的角度評估文本的閱讀難易程度。這些特徵反映了文本在語言結構、詞彙選用和句子組成上的複雜性。

本研究採用五種適用於英文文本的可讀性公式來構建可讀性特徵。

1. SMOG

SMOG，全名為Simple Measure of Gobbledygook，是McLaughlin在1969年提出的可讀性公式，用於估計文本的可讀性年齡，即讀者需要多少年的教育年齡才能夠理解該文本。此公式的特點在於其專注於複雜詞彙，特別是多音節詞，因為作者認為這些多音節詞是增加閱讀難度的主要因素。

SMOG公式透過計算文本中多音節詞的數量，隨後結合固定常數來估算出讀者所需的教育年齡。公式如式 3.X所示。SMOG 分數越高，表示文本可讀性越低，適合的讀者年齡越高。

|  |  |
| --- | --- |
|  | 公式3.X |

本研究對每則評論文本進行SMOG的計算，並將其作為可讀性特徵之一。

1. Flesch reading ease

Flesch reading ease是Flesch在1948年提出的可讀性公式，用於衡量文本的易讀程度。此公式依據句子長度及單字的音節數來評估文本難易度，公式如 式3.X所示。分數範圍從0到100，分數越高表示越容易閱讀。

|  |  |
| --- | --- |
|  | 公式3.X |

本研究對每則評論文本進行Flesch reading ease的計算，並將其作為可讀性特徵之一。

1. Flesch-Kincaid Grade Level

Flesch-Kincaid Grade Level是Kincaid等人在1975年提出的可讀性公式，這是Flesch reading ease的改良版本。公式如式3.X所示，同樣依據句子長度及單字的音節數來評估文本難易度，但不同之處在於，Flesch-Kincaid Grade Level透過調整權重，將分數轉換為對應的年級程度，以評估文本適合的讀者年齡。

|  |  |
| --- | --- |
|  | 公式3.X |

本研究對每則評論文本進行Flesch-Kincaid Grade Level的計算，並將其作為可讀性特徵之一。

1. Gunning FOG Index

Gunning FOG Index是Gunning在1952年提出的可讀性公式，用於衡量文本的易讀程度。此公式專注於句子長度及複雜詞彙的比例，並明確將複雜詞彙定義為多音節且非專業詞彙的單字。

公式如式3.X所示，FOX Index的值代表適讀年齡，值越高，表示文本可讀性越低。

|  |  |
| --- | --- |
|  | 公式3.X |

本研究對每則評論文本進行Gunning FOG Index的計算，並將其作為可讀性特徵之一。

1. 熵(Entropy)

熵是Shannon在1948年提出的一個概念，用於分析訊息的不確定性及複雜度。在文本分析中，熵反映詞彙出現的機率分佈，若詞彙分佈越均勻時，表示每個詞彙的出現頻率越接近，此時熵值也會越高。因此，文本的熵值越高，表示文本的複雜性及不確定性越高，閱讀難度也會相對複雜。

公式如式3.X所示，其中是文本中第i個詞的出現概率。

|  |  |
| --- | --- |
|  | 公式3.X |

本研究對每則評論文本進行熵的計算，並將其作為可讀性特徵之一。

1. 結構特徵(Structure)

結構特徵是指文本中與組織結構相關的特徵，如組織方式、句子結構、標點符號等。這些特徵有助於理解文本的內在結構與形式，本研究採用的結構特徵類別共包含以下七種特徵。

1. 字元數量

字元數量是指文本中所有字符的總數，包含字母、數字及標點符號。本研究計算每則評論文本的總字元數，作為結構特徵之一。

1. 單字數量

單字數量是指文本中單字的總數。本研究將每則評論文本中的文本去除標點符號等非相關文字後，利用len函數計算出單字數量作為結構特徵之一。

1. 句子數量

句子數量是指文本中句子的總數。本研究將每則評論文本依照標準的結束標點符號進行斷句後，再計算句子總數，並將其作為結構特徵之一。

1. 句子平均長度

句子長度是指每個句子的平均單字數，長句一般能表示更複雜的結構，短句則為較簡潔的表達。本研究首先將每則評論文本依照標準的結束標點符號進行斷句後，再計算出句子總數每一句的單字數，最後將所有句子的單字總和並除以句子總數，以得到句子平均長度，並將其做作為結構特徵之一。

1. 感嘆句比例

感嘆句一般用來表達強烈的情感或驚訝，因此文本中感嘆句所佔的比例可以反映出文本的情感強度。本研究計算每則評論文本中驚嘆號的出現次數，來確定感嘆句的比例，並將其作為結構特徵之一。

1. 疑問句比例

疑問句一般用來表達疑惑，而文本中疑問句所佔的比例可以反映文本的互動性或探討性。本研究計算每則評論文本中問號的出現次數，來確定疑問句的比例，並將其作為結構特徵之一。

1. 標點符號數量

標點符號數量是指文本中所有標點符號的總數，標點符號的使用頻率可以反映文本的節奏和語氣。本研究計算每則評論中的標點符號數量，並將其作為結構特徵之一。

1. 句法特徵(Syntax)

句法特徵是指文本中與語法結構相關的特徵，如詞彙的詞性、定義及複雜程度。這些特徵專注於句子內部的語法結構和語法關係，並有助於理解文本的語義和結構，本研究採用的句法特徵類別共包含以下三種。

1. POS標記(Part-of-Speech Tagging)

POS 標記是指將文本中的每個詞彙基於其在句子中的語法功能，標註為相應的詞性，如名詞、動詞、形容詞等。

本研究使用spaCy套件中的英文模型對每則評論文本進行詞性標註，接著統計各種詞性的出現次數，並計算其在文本中的比例，最後將結果作為句法特徵之一。

1. TTR

TTR，Type-Token Ratio，是一種衡量文本中詞彙豐富度的指標，透過比較文本中的不同詞彙數及總單字數來評估文本的詞彙多樣性，基於它能夠從詞彙多樣性角度提供對句法結構的洞察，本研究對每則評論進行TTR的計算，並將其作為句法特徵之一。

公式如式3.X所式，TTR 值越大，表示文本中使用的不同詞彙越多，詞彙豐富度越高；反之，則表示文本中的詞彙重複率較高，詞彙多樣性較低。

|  |  |
| --- | --- |
|  | 公式3.X |

1. 命名實體識別 (Named Entity Recognition)

命名實體識別又稱為NER實體識別，用於識別文本中的實體並將其分類到特定的類別，如人名、日期、事件等。特別是在句子結構因包含多個實體或專有名詞而變得複雜時，NER有助於更準確地分辨和組織句子中的訊息。

本研究使用spaCy套件取得所有可識別的實體類別，並對每則評論文本進行實體辨識及分類，並將其結果作為句法特徵之一。

## 特徵轉換

從文本提取出特徵以後，往往會遇到各個特徵有不同的單位及數值範圍，如果將特徵直接用於模型訓練，可能會影響模型的學習成效，尤其當某個特徵的數值範圍遠大於其他特徵時，模型可能因過度關注此特徵，而忽略其他具有重要資訊的特徵，這將導致模型產生偏差，並降低預測的準確性，因此需要對這些特徵進行適當的轉換。

資料正規化(Data Normalization)為過去機器學習數據預處理階段常使用的一種系統化方法，目的是將不同特徵的數值調整至統一的尺度範圍，避免模型訓練因數值差異過大，而產生不利的影響。

常見的正規化方法包括最小最大縮放（Min-Max Scaling）和Z分數標準化（Z-score Normalization），兩者的差異在於：最小最大縮放會將數據等比例縮放至[0,1]範圍內；而Z分數標準化則是通過減去均值並除以標準差，將數據轉換為均值為0、標準差為1的分佈。

考慮到需要保持特徵的數據原始結構和分佈，避免因失真而影響模型的學習效果，因此本研究最終選擇使用最小最大縮放作為特徵轉換的方法。公式如3.XX所示：

|  |  |
| --- | --- |
|  | (3.XX) |

## 模型建置

本研究採用多種模型進行比較及分析，包含長短期記憶模型(Long-Short Term Memory, LSTM)、隨機森林、XGBoost、AdaBoost以及BERT，這些模型在過去的文本分類任務中經常被採用，並展現令人滿意的表現。此外，為提升模型的穩健性，本研究採用十折交叉驗證(10-Folds Validation)，以確保結果的可靠性。

除了上述所述的模型外，本研究也引入集成學習技術以進一步提升模型的性能。集成學習透過結合多個基礎模型，期望能夠彌補單一模型的缺點，進而獲得更具穩健性和準確性的預測結果。

本小節將會介紹本研究所使用的模型以及交叉驗證方法。

* + 1. **LSTM**

LSTM由輸入門、遺忘門、輸出門和記憶單元組成，通過這些具有長期記憶能力的單元來有效控制訊息的保存與遺忘，提高序列數據的分類準確性。本研究對模型進行以下參數設置，如表3-X。

|  |  |  |
| --- | --- | --- |
| 參數 | 說明 | 參數設置 |
| units | 模型中每一層的神經元數量，直接影響模型的複雜性及記憶能力 | 128,256 |
| Dropout\_ratio | 訓練過程中每次迭代所隨機關閉的神經元比例，增加模型泛化能力。 | 0.2,0.5 |
| learning\_rate | 即學習率，控制模型學習的速度 | 0.001,0.005,0.01 |
| epoch | 即訓練週期，訓練集重複訓練的次數 | 50 |
| batch\_size | 一次訓練過程中，單批處理資料的數量 | 128,256 |

* + 1. **隨機森林**

隨機森林通過結合多棵獨立訓練且引入隨機性的決策樹來提高分類的準確性。本研究對模型進行以下參數設置，如表3-X。

|  |  |  |
| --- | --- | --- |
| 參數 | 說明 | 參數設置 |
| n\_ estimators | 即決策數的數量 | 50,100,200,300,400 |
| random\_state | 設定固定的隨機種子，確保模型結果的可重現性 | 42 |

* + 1. **XGBoost**

XGBoost 以梯度提升決策樹為基礎，結合了 Bagging 的特徵隨機抽樣和 Boosting 的逐步提升特性。在生成每棵樹的過程中，模型不使用所有特徵進行決策，而是隨機抽取部分特徵，這樣可以降低過度擬合的風險，此外，每個新分類器都會針對前一個模型的錯誤進行修正，最後將所有弱分類器通過加權整合成一個強分類器，從而提升模型的準確性和泛化能力。

本研究對模型進行以下參數設置，如表3-X。

|  |  |  |
| --- | --- | --- |
| 參數 | 說明 | 參數設置 |
| n\_ estimators | 即決策數的數量 | 50,100,200,300,400 |
| random\_state | 設定固定的隨機種子，確保模型結果的可重現性 | 42 |
| subsample | 模型在生成每顆樹時所使用的訓練資料比例 | 0.5,0.8 |
| max\_depth | 決策樹的最大深度，用來控制每棵樹的複雜度 | 5,6,8 |
| learning\_rate | 即學習率，控制模型學習的速度 | 0.001,0.01,0.05,0.1 |

* + 1. **AdaBoost**

AdaBoost則是完全基於Boosting技術，通過逐步建立弱分類器來提升性能。在每一輪訓練中，模型會根據上一輪的錯誤分類結果調整樣本的權重，最後將所有弱分類器通過加權的方式整合成一個強分類器，從而提升分類的準確性。

本研究對模型進行以下參數設置，如表3-X。

|  |  |  |
| --- | --- | --- |
| 參數 | 說明 | 參數設置 |
| n\_ estimators | 即決策數的數量 | 50,100,200,300,400 |
| random\_state | 設定固定的隨機種子，確保模型結果的可重現性 | 42 |
| learning\_rate | 即學習率，控制模型學習的速度 | 0.001,0.01,0.05, 0.1,0.3,0.5 |

* + 1. **BERT**

BERT 以 Transformer 架構為基礎，結合了雙向編碼機制，讓模型能夠同時考慮句子的前後文以準確地捕捉詞彙在上下文中的語義，並進行有效的預測。

BERT分為預訓練及微調兩個部分，在預訓練階段，模型通過遮蔽語言模型和下一句預測的任務來學習深層的語言結構和語義知識；在微調階段則根據具體所應用的下游任務進行參數調整，例如情感分析、文本分類等。

本研究對模型進行以下參數設置，如表3-X。

|  |  |  |
| --- | --- | --- |
| 參數 | 說明 | 參數設置 |
| learning\_rate | 即學習率，控制模型學習的速度 | 2e-5, 3e-5, 5e-5 |
| batch\_size | 一次訓練過程中，單批處理資料的數量 | 16,32 |
| epoch | 即訓練週期，訓練集重複訓練的次數 | 3,4 |

* + 1. **集成學習**
    2. **交叉驗證法**

## 模型評估

本研究在模型建置完成以後，將採用多種評估方法對模型表現進行全面的評估。首先，通過混淆矩陣對分類結果進行具體分析，表示方法如表3.7，並進一步計算衍生指標。接著，為深入探討特徵選擇與模型表現的關聯，本研究採用SHAP指標來剖析模型中不同特徵的重要性，確保模型的可解釋性並提升其性能。

1. 混淆矩陣

表3.7混淆矩陣

|  |  |  |
| --- | --- | --- |
|  | 實際為AI生成之評論 | 實際為人類撰寫之評論 |
| 預測為AI生成之評論 | True Positive（TP） | False Negative（FN） |
| 預測為人類撰寫之評論 | False Positive（FP） | True Negative（TN） |

True Positive（TP）表示實際為AI生成之評論且模型預測其為AI生成之評論，True Negative（TN）表示實際為人類撰寫之評論且模型預測其為人類撰寫之評論，False Positive（FP）表示實際為AI生成之評論而模型預測其為人類撰寫之評論，False Negative（FN）表示實際為人類撰寫之評論且模型預測其為AI生成之評論。

依據這四個類別可進一步衍生計算出準確率（Accuracy）、精確率（Precision）、召回率（Recall）、特異度(Specificity)及F1-score。

1. 準確率（Accuracy）

準確率用來衡量模型正確預測的能力，公式如式3.X所示。反映模型在所有預測中有多少比例是正確的，因此在不平衡資料的情況下，可能會受到目標欄位分布影響，導致Accuracy指標的代表性下降。

|  |  |
| --- | --- |
|  | (3.XX) |

1. 精確率（Precision）

精確率表示模型預測為正類的所有樣本中，實際為正類的比例，用來衡量型一誤差的判斷表現，公式如式3.X所示。

|  |  |
| --- | --- |
|  | (3.XX) |

1. 召回率（Recall）

召回率表示所有實際為正類的樣本中，被正確預測為正類的比例，用來衡量型二誤差的判斷表現，公式如式3.X所示。

|  |  |
| --- | --- |
|  | (3.XX) |

1. 特異度(Specificity)

特異度與召回率相反，表示所有實際為負類的樣本中，被正確預測為負類的比例，公式如式3.X所示。

|  |  |
| --- | --- |
|  | (3.XX) |

1. F1分數(F1-score)

F1分數為精確率與召回率的調和平均，綜合了兩個指標的整體表現，防止精確度過高但召回率過低，或者精確度過低但召回率過高的情況，公式如式3.X所示。

|  |  |
| --- | --- |
|  | (3.XX) |

1. SHAP指標

# 第四章 實驗設計與預期結果

## 資料描述與處理

本研究所使用的資料集為combat-ai-restaurants及FackReviews。combat-ai-restaurants資料集由Gambetti及Han建立，並公開於Github平台，資料集包含兩部分:一部分為來自Yelp.com的人類撰寫評論，時間範圍為2021年1月至2022年7月，另一部份為基於這些評論，利用GPT-3預訓練模型所生成的生成評論；FakeReviews資料集是由Salminen等人建立，並公開於Github平台。該資料集包含兩部分:一部分為來自Amazon平台的人類撰寫評論，時間範圍為2018年1月至2018年12月，另一部份為基於這些評論，利用GPT-2預訓練模型所生成的生成評論。

combat-ai-restaurant資料集的欄位包括文本內容（text）及類別（label），其中類別1為生成評論，類別0為人類撰寫的評論。資料集經資料清理後呈現為平衡資料集，包含7266筆生成評論及7588筆人類撰寫評論，共計14854筆資料。

FakeReviews資料集的欄位包括文本內容（text）、產品分類(Category) 、產品評分(rating)及類別（label），其中類別1為生成評論，類別0為人類撰寫的評論。資料集經資料清理後呈現為平衡資料集，包含2000筆生成評論及2000筆人類撰寫評論，共計4000筆資料，而剩餘欄位為文本內容及類別。

## 敘述統計

本小節將說明本研究所採用資料集的敘述統計。

如圖X及表X所示，在combat-ai-restaurant資料集中，生成評論的字數主要分佈在55字至119字的區間，平均字數為92.06字，最小字數為6字，最大字數為236字。人類撰寫的評論的字數主要分佈在60字至143字區間，平均字數為104.08字，最小字數為11字，最大字數為249字。

表X combat-ai-restaurant資料集字數統計

|  |  |  |
| --- | --- | --- |
| 類別 | 0(人類撰寫的評論) | 1(生成評論) |
| 總筆數 | 7588 | 7266 |
| 平均字數 | 104.08 | 92.06 |
| 最小字數 | 11 | 6 |
| 最大字數 | 249 | 236 |
| 中位數 | 90 | 79 |
| Q1(25%) | 60 | 55 |
| Q3(75%) | 143 | 119 |

圖X文本評論字數分佈圖

如圖X及表X所示，在FakeReviews資料集中，生成評論的字數主要分佈在66字至193字的區間，平均字數為131.24字，最小字數為10字，最大字數為259字。人類撰寫的評論的字數主要分佈在70字至198字區間，平均字數為131.85字，最小字數為10字，最大字數為259字。

表X FakeReviews資料集字數統計

|  |  |  |
| --- | --- | --- |
| 類別 | 0(人類撰寫的評論) | 1(生成評論) |
| 總筆數 | 2000 | 2000 |
| 平均字數 | 131.85 | 131.24 |
| 最小字數 | 10 | 10 |
| 最大字數 | 259 | 259 |
| 中位數 | 132 | 131 |
| Q1(25%) | 70 | 66 |
| Q3(75%) | 198 | 193 |

圖X文本評論字數分佈圖

## 實驗設計

**實驗劃分**

1. **實驗一：超參數搜尋**
   * **使用網格搜尋來確定最佳的超參數組合，專注於一個特定文本特徵。**
2. **實驗二：模型選擇**
   * **在確定的超參數下，選擇單一特徵進行不同模型的比較，找出表現最佳的模型。**
3. **實驗三：集成學習**
   * **如果多個模型的表現相似，進行集成學習的實驗，探討是否能通過結合多個模型來提高性能。**
4. **實驗四：特徵排名比較**
   * **在實驗三的基礎上，進一步比較不同特徵類型的模型性能，評估各個特徵對模型效果的影響，並進行排名。**
5. **實驗五：特徵組合與 BERT 比較**
   * **將選擇的特徵進行組合，並與 BERT 進行比較，評估不同模型和特徵組合的效果，從而確定最適合的特徵配置。**
6. **最終泛化評估**
   * **使用另一個資料集進行泛化評估，以驗證所選模型的通用性和穩定性，確保模型能夠適應未見數據的情況。**

1. https://github.com/iamalegambetti/combat-ai-restaurants/tree/main?tab=readme-ov-file [↑](#footnote-ref-2)
2. https://github.com/joolsa/FakeReviews?tab=readme-ov-file [↑](#footnote-ref-3)