Sentiment Analysis on Uber Customer Reviews Dataset (2024) by Machine Learning and Deep Learning

一、動機

Uber 是一家全球知名的公司，其用戶評論成為研究用戶體驗和服務改進的寶貴資料。選擇的資料集為 2024 年最新的 Uber 客戶評論，能夠反映用戶對於服務的真實看法與需求。

該資料集涵蓋了來自不同地區、文化背景及語言的用戶回饋，這種多樣性為機器學習模型的訓練與測試提供了豐富的資料來源，有助於提升模型的泛化能力與適應性。再者，客戶評論大多以非結構化文字形式存在，語言表達方式靈活多變，包含隱喻、情緒化語句及俚語，這為自然語言處理（NLP）技術的應用與研究提供了挑戰和機會。利用該資料集，研究情感分析技術的應用，可深入探討 NLP 模型在處理多樣化語言場景中的性能表現及優化潛力。

最後，基於客戶評論進行情感分析，不僅能幫助企業深入了解用戶需求和痛點，還能為制定改進服務方案提供科學根據。

資料集敘述：（資料集來自以下連結）
https://www.kaggle.com/datasets/kanchana1990/uber-customer-reviews-dataset-2024

此資料集包含超過12,000條來自Google Play商店的Uber應用程式用戶評論，涵蓋評分、服務回饋和開發者回應等，用於了解用戶體驗。資料經過清理與匿名化，確保隱私合規與倫理使用。

以下是本次project有使用到的欄位

1. content: 評論
2. score: 評分(0-5分)
3. thumbsUpCount: 這則評論獲得了幾個讚
4. reviewCreatedVersion: 評論產生時的APP版本
5. at: 評論時間

二、分析工具：

1. 資料處理工具

- 資料清理與預處理：
    - `clean_text` 函數用於清理文字，移除非必要字符並標準化格式。

    - 缺失值處理：handle_missing_values 函數。
    - 離群值處理：handle_outliers 函數。
- 特徵工程：
    - 使用 TfidfVectorizer 將文本轉換為向量。
    - 將時間特徵（如日期、時間）進行分解，並創建新特徵（如 is_weekend）。

2. 資料可視化工具

- 資料分布分析：繪製分數（score）的分布圖、文字長度的分布圖。
- 詞頻分析：使用條形圖和文字雲視覺化高頻詞，並分析正面與負面評論中出現的詞彙。
- 混淆矩陣視覺化：用 seaborn 繪製混淆矩陣，展示分類結果。

3. 模型訓練與評估工具

- 機器學習模型：使用 sklearn 中的分類器（如 RandomForestClassifier、DecisionTreeClassifier）。
- 深度學習模型：使用 HuggingFace 上開源的預訓練模型(如 BERT、ALBERT、DistilBERT 和 XLNet)，嘗試做情緒分析的downstream task的訓練。
- 評估指標：使用 accuracy_score、precision_score、recall_score 等方法量化模型性能。

4. 實驗設計工具

- 資料集切分：利用 train_test_split 將資料分為訓練集與測試集。
- 樣本平衡：通過 SMOTE 方法解決類別不平衡問題。

三、實作與評估方法：

實作方法

1. 資料載入與清理：
    - 載入資料：從 CSV 檔案讀取 Uber 評論資料集。
    - 刪除多餘欄位：移除如 userName 等對分析無影響的欄位。
    - 處理缺失值與離群值：
        - 缺失值使用中位數或眾數填充。
        - 離群值通過標準差界限處理。
2. 特徵工程：
    - 文字特徵：
        - 清理文字內容（移除標點符號、轉小寫）。

        - 使用 TF-IDF 向量化處理文字特徵，設定最大特徵數為2000
    - 數值特徵：
        - 提取小時(hour)、星期幾(day_of_week)、年份(year)、月份(month)、是否為週末(is_weekend)
        - 計算文本長度並將其作為新特徵。
        - 標準化處理：thumbsUpCount, length, 以及所有時間特徵
        - 處理離群值：使用均值±3倍標準差的方法
        - 處理缺失值：數值型用中位數填充，類別型用眾數填充
    - 資料標準化：對數值特徵進行標準化處理，確保模型訓練的穩定性。
3. 資料分割與平衡：
    - 訓練/測試集劃分：將數據集按照 80/20 分為訓練集和測試集。
    - 樣本平衡：通過 SMOTE 方法生成合成資料，解決正負樣本不均問題。
4. 模型構建與訓練：
    - 使用多種模型（如隨機森林、邏輯迴歸、決策樹、XGBoost）進行分類任務。
    - 對模型進行參數調整（如隨機森林中的樹數與深度）。
5. 資料可視化：
    - 繪製詞頻圖與文字雲。
    - 展示分數分布與文本長度的影響。
    - 使用混淆矩陣評估模型預測的準確性。

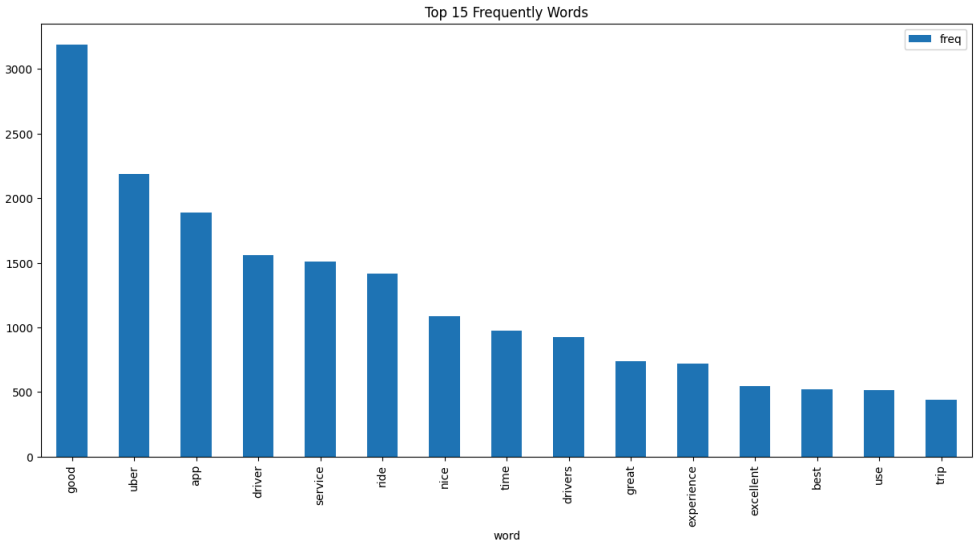
評估方法

1. 分類指標評估：
    - 準確率 (Accuracy)：測試集中正確預測的比例。
    - 精確率 (Precision)：正向預測中實際正向的比例。
    - 召回率 (Recall)：所有實際正向中模型檢測到的比例。
    - F1 分數 (F1 Score)：精確率和召回率的加權平均。
    - ROC-AUC：衡量模型對不同分類閾值的整體性能。
2. 混淆矩陣分析：
    - 分析模型在不同類別（正面/負面）上的預測準確性和錯誤率。
    - 用Heatmap可視化結果，便於理解模型偏好。
3. 模型比較與選擇：
    - 比較多種模型在測試集上的表現，挑選 F1 分數或 ROC-AUC 最優的模型。
    - 分析 SMOTE 前後模型性能的變化，確認樣本平衡對結果的影響。

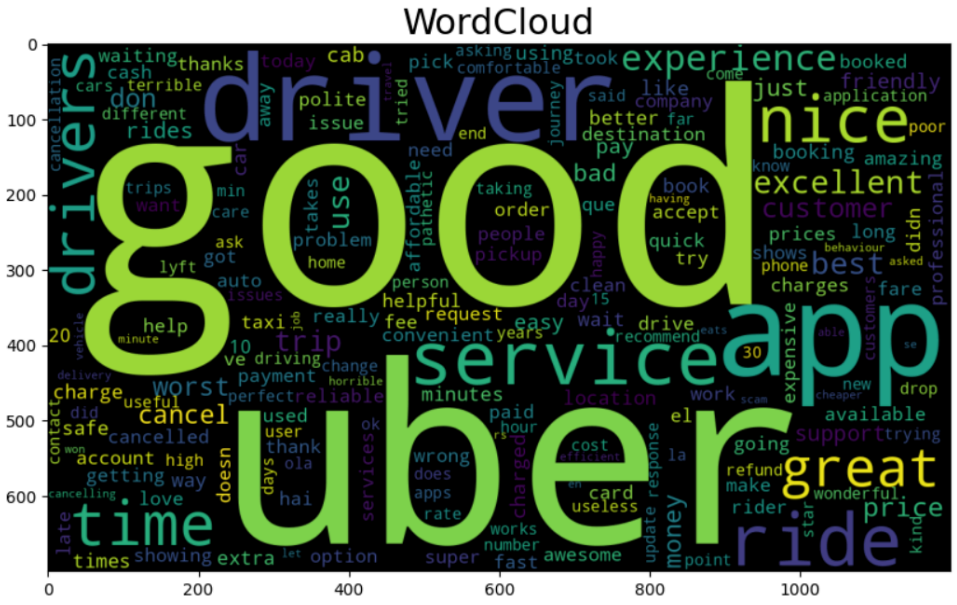
四、流程 (分析流程圖、結果截圖等)：

1. 資料視覺化
    a. content
        i. uber評論中的前15大高頻詞

圖一顯示uber評論的前15大高頻詞，顯示出詞語多為稱讚單字(good, great, best, excellent等)、有關服務的單字(driver, experience, use, trip等)這兩類。進一步以文字雲視覺化可以更顯見整體而言對uber的評價相當不錯，為了區分正面意見和負面意見的高頻詞，將進一步分別繪製正反面評價的文字雲。



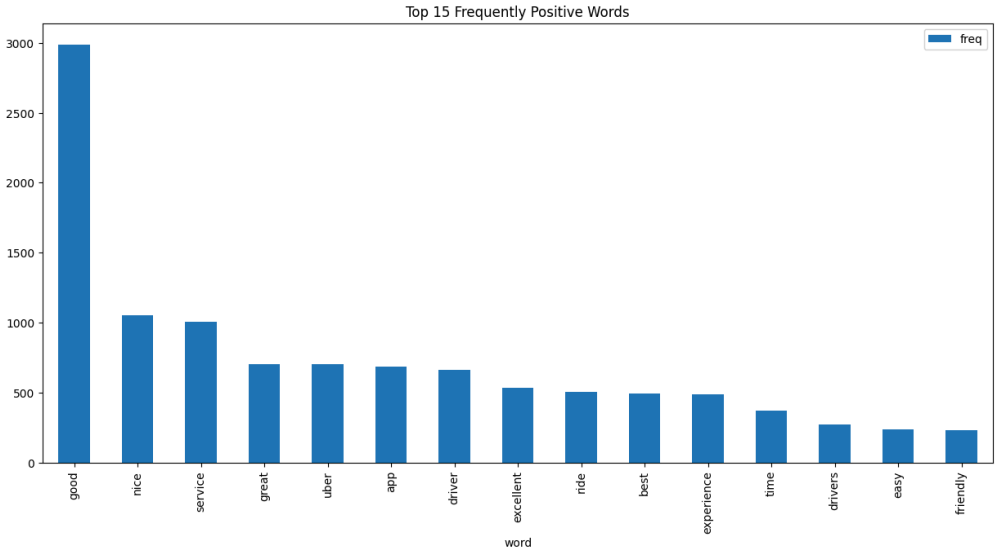
圖一、uber評論中的前15大高頻詞



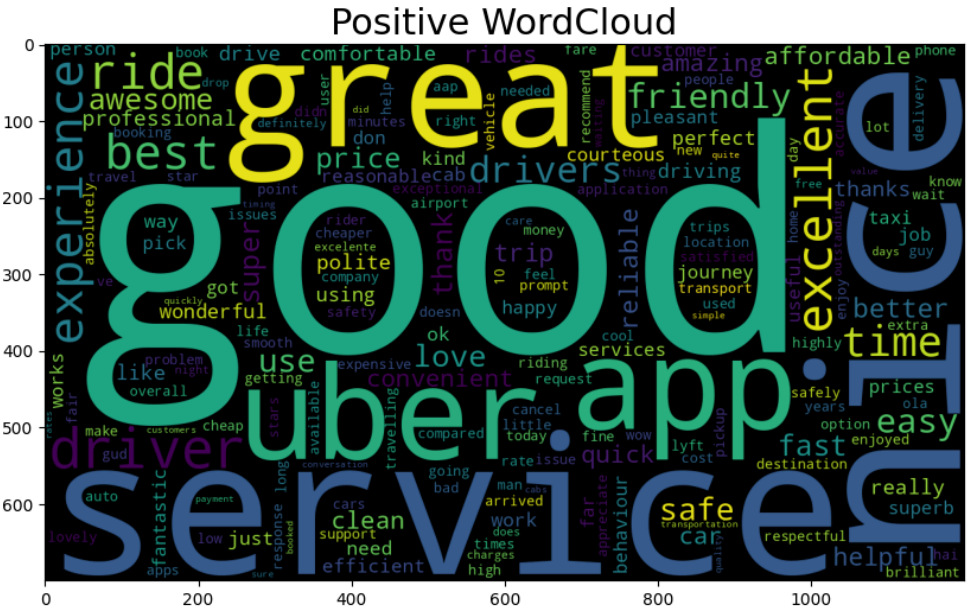
圖二、uber評論文字雲

ii. 正面評價的前15大高頻詞

由圖三、圖四可見正面評價者多稱讚uber是一個很好的APP，且優點在於提供良好的服務。



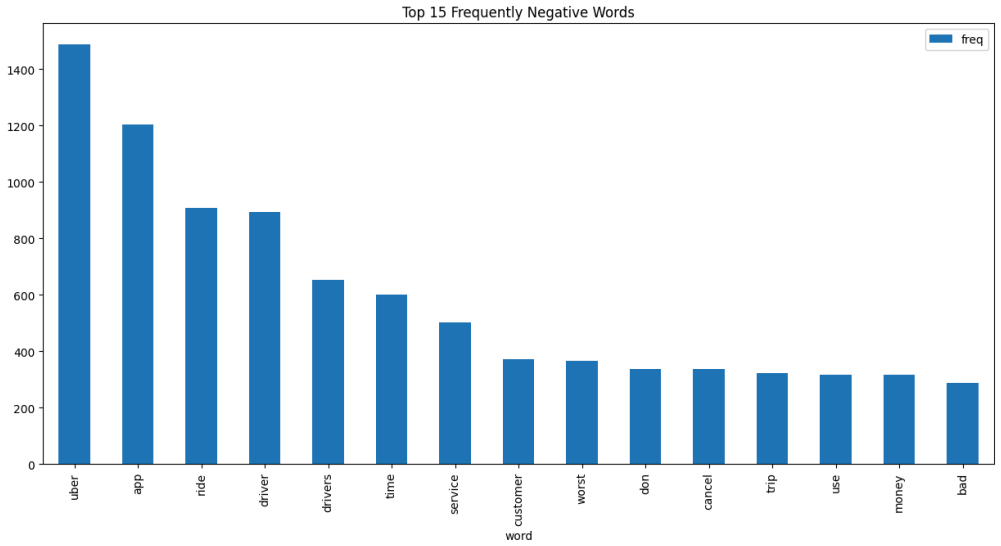
圖三、正面評價的前15大高頻詞



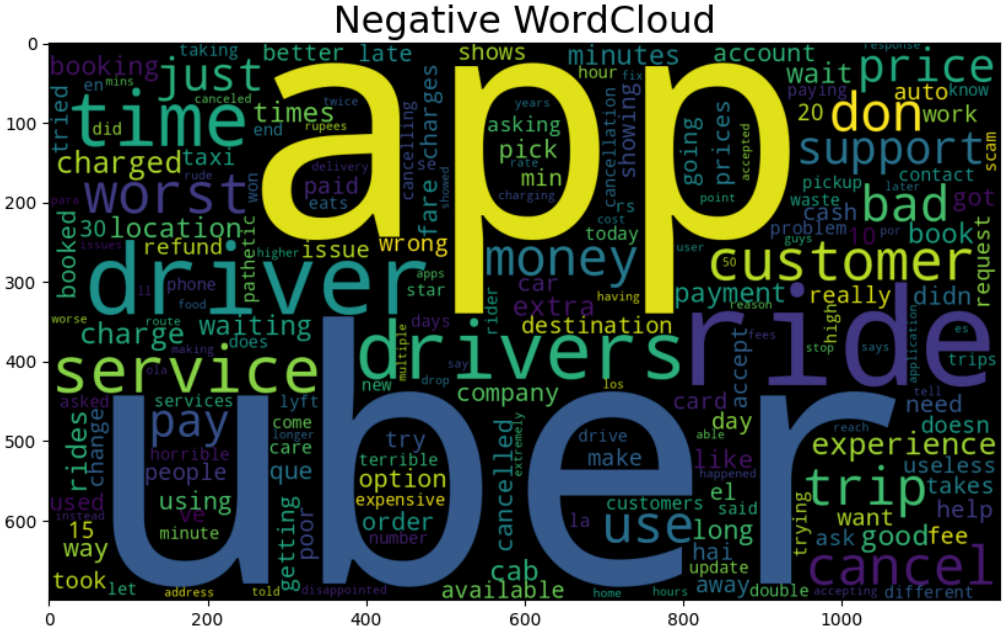
圖四、正面評價文字雲

iii. 負面評價的前15大高頻詞

由圖五、圖六可見負面評價者明確指出駕駛、app、時間的問題最多。其中，若將drivers、driver合併來看，將高居第一名，因此可見駕駛的問題最可能造成對uber的負面評價。



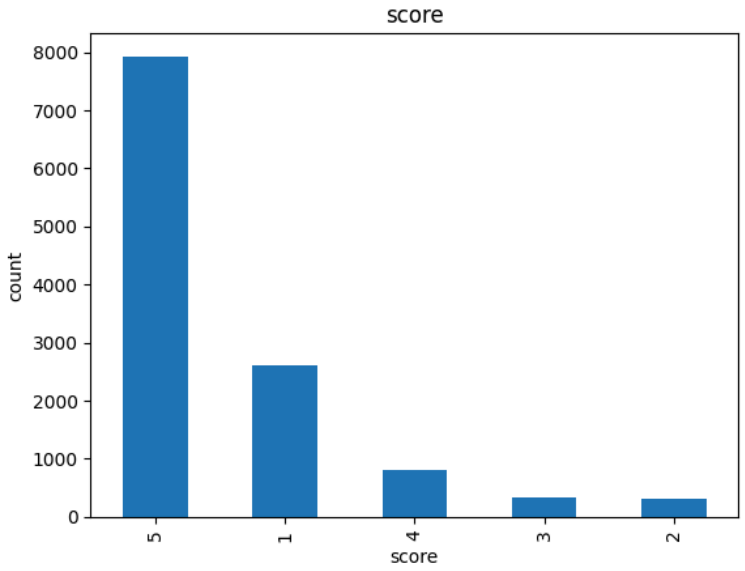
圖五、負面評價的前15大高頻詞



圖六、負面評價文字雲

b. score

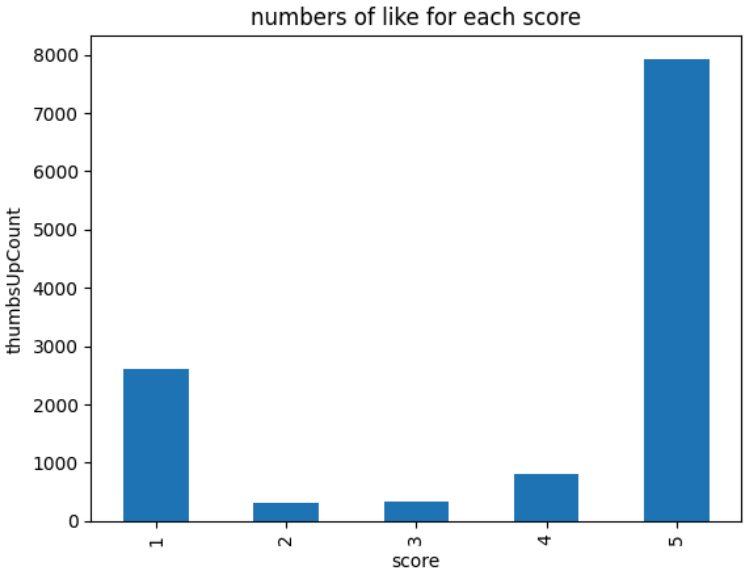
由圖七可知，約8000人給予五星好評、其次是約3000人給予一星評論，接著是四星、三星、二星。總體而言uber的評價非常高。



圖七、評分分布人數

c. thumbsUpCount

圖八顯示各評分的評語獲得了幾個讚，其分部與score相同。



圖八、各分數獲得讚數

d. at

圖九顯示多數評論者偏好在傍晚到晚上評分，圖十顯示假日評分和平日評分的人數比例接近1:3(即假日兩天會對上六個平日數)，表示評分的時間均勻分配在一周的每個星期但假日會稍多。

at	count
22	667
19	631
20	626
16	616
21	601
17	578
13	575

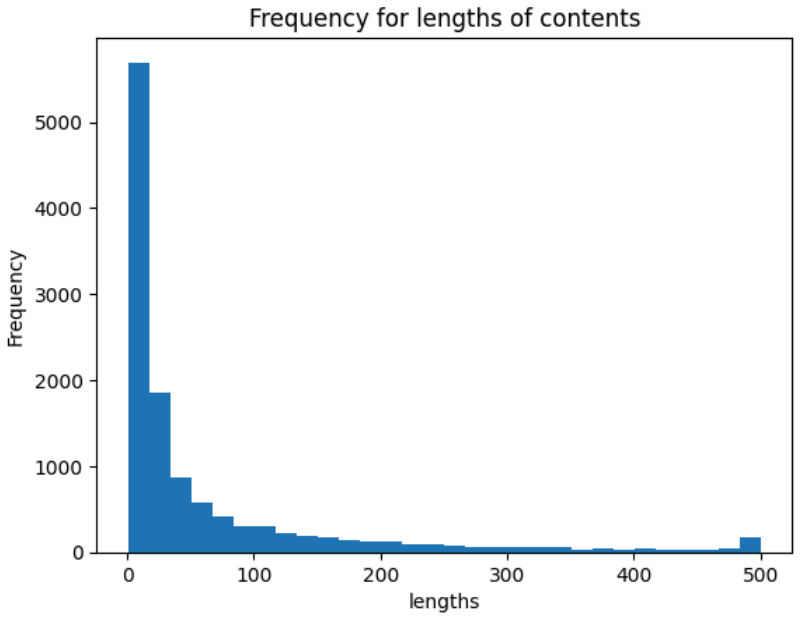
is_weekend	count
0	8804
1	3196

圖九、統計評論者在幾點撰寫　　圖十、統計評論者是否在假日撰寫

e. length

評論的長度也可能會關係到評分的高低，所以我們在資料集中會新增評論長度的欄位。

圖十一顯示多數評論者撰寫較短的評論。而根據程式，正面評論的平均長度為28.1，而負面評論的平均長度為161.8，可以推論，對uber不滿意的用戶通常會寫更長的評論。



圖十一、評論長度統計

2. 機器學習
    a. Random Forest

	原始模型	SMOTE模型
Accuracy	0.915	0.911
Precision	0.93	0.945
Recall	0.955	0.934
F1-score	0.943	0.939
Area under roc	0.879	0.891
Confusion Matrix	[509 125] [80 1686]	[538 96] [117 1649]

    b. Naive-Bayes

	原始模型	SMOTE模型
Accuracy	0.922	0.919
Precision	0.943	0.975
Recall	0.951	0.914
F1-score	0.947	0.943
Area under roc	0.896	0.924
Confusion Matrix	[533 101] [87 1679]	[592 42] [152 1614]

    c. Logistic Regression

	原始模型	SMOTE模型
Accuracy	0.933	0.921
Precision	0.944	0.954
Recall	0.967	0.939
F1-score	0.955	0.946

Area under roc	0.903	0.906
Confusion Matrix	[532 102] [58 1708]	[554 80] [109 1657]

d. Decision Tree

	原始模型	SMOTE模型
Accuracy	0.879	0.872
Precision	0.925	0.922
Recall	0.910	0.902
F1-score	0.917	0.912
Area under roc	0.852	0.845
Confusion Matrix	[503 131] [159 1607]	[500 134] [173 1593]

e. XGBoost

	原始模型	SMOTE模型
Accuracy	0.926	0.926
Precision	0.941	0.954
Recall	0.960	0.946
F1-score	0.950	0.950
Area under roc	0.895	0.909
Confusion Matrix	[527 107] [71 1695]	[533 81] [96 1670]

3. NLP
    a. BERT (epoch=5)

	原始模型	SMOTE模型
Accuracy	0.9571	0.9550
Precision	0.9716	0.9616
Recall	0.9700	0.9779
F1-score	0.9708	0.9697
Area under roc	0.9761	0.9582
Confusion Matrix	[584 50] [53 1713]	[565 69] [39 1727]

    b. ALBERT (epoch=5)

	原始模型	SMOTE模型
Accuracy	0.9429	0.8792
Precision	0.9727	0.9824
Recall	0.9490	0.8511
F1-score	0.9607	0.9697
Area under roc	0.9375	0.9120
Confusion Matrix	[587 47] [90 1676]	[607 27] [263 1503]

    c. DistilBERT (epoch=5)

	原始模型	SMOTE模型
Accuracy	0.9567	0.8138
Precision	0.9663	0.9832
Recall	0.9751	0.7599
F1-score	0.9707	0.8572
Area under roc	0.9402	0.8618

Confusion Matrix	[574 60] [44 1722]	[611 23] [424 1342]

d. XLNet (epoch=5)

	原始模型	SMOTE模型
Accuracy	0.9508	0.8550
Precision	0.9655	0.9823
Recall	0.9677	0.8177
F1-score	0.9666	0.8925
Area under roc	0.9358	0.8883
Confusion Matrix	[573 61] [57 1709]	[608 26] [322 1444]

五、分析結果與結論：

最佳表現模型：

本次情感分析研究針對Uber客戶評論資料集進行多種機器學習和自然語言處理模型的訓練與評估。經過比較，發現BERT模型在各項指標上均表現出色，尤其是在準確率、精確率、召回率和F1分數方面，均達到了最高的水準。BERT在原始模型的準確率為95.71%，而在使用SMOTE進行樣本平衡後，準確率輕微下降至95.50%。表示BERT模型在處理情感分析任務時，能夠有效捕捉文本中的情感特徵，並且在面對多樣化的用戶評論時，依然保持了穩定的表現。

在傳統機器學習模型中，XGBoost模型的表現同樣不容忽視，其在原始模型和SMOTE模型下的準確率均為92.6%。這顯示出XGBoost在處理結構化資料時的優勢，尤其是在特徵工程和模型調整方面的靈活性。其他模型如隨機森林、邏輯回歸和Naive-Bayes也展現了良好的表現，但相較之下，BERT和XGBoost則是眾模型當中表現最為突出的兩大模型。

SMOTE 的影響：

在本研究中，SMOTE（Synthetic Minority Over-sampling Technique）被用來解決類別不平衡問題，特別是在負面評價的樣本數量相對較少的情況下。透過生成合成樣本，SMOTE能夠有效提升模型在少數類別上的預測能力。實驗結果顯示，使用SMOTE後，許多模型的召回率有所提升，這意味著模型在識別負面評價方面的能力得到了增強。

例如，在隨機森林模型中，使用SMOTE後的召回率從95.5%下降至93.4%，而在Naive-Bayes模型中，召回率從95.1%下降至91.4%。這表明，雖然某些模型在使用SMOTE後的整體準確率略有下降，但在少數類別的識別能力卻得到了顯著改善。此外，SMOTE對於模型的F1分數也有正面影響，這是因為F1分數綜合考量了精確率和召回率，能夠更全面地反映模型的表現。

結論：
總體而言，BERT模型在本次情感分析中表現最佳，顯示出其在自然語言處理任務中的強大能力。XGBoost則在結構化數據處理上充分展現其優勢，成為一個可靠的替代選擇。而SMOTE的應用不僅改善了模型在少數類別上的預測能力，還促進了模型的穩定性和可靠性。實驗結果強調了在處理不平衡資料集時，採用適當的樣本平衡技術的重要性，並為未來的研究提供了有價值的參考。未來的研究可以進一步探索其他先進的模型和技術，以進一步提升情感分析的準確性和效率。