DM2024 ISA5810 Lab2 Homework

Step 1: Initial Approach

- 方法：
  - 使用 RandomForestClassifier 作為初始模型。
  - 採用 TfidfVectorizer 提取特徵，包括 unigram 和 bigram，特徵數量限制為 10,000。
  - 訓練數據未處理類別不平衡問題。
- 結果：
  - 模型訓練時間長，且模型性能（準確率）不理想。
  - 沒有針對超參數進行調整，導致模型對多類別分類的表現不足。

Step 2: Switching to LinearSVC

- 方法：
  - 將模型切換為 LinearSVC（支持向量機），以提高訓練速度和性能。
  - 使用類別權重平衡（class_weight='balanced'）來處理數據不平衡。
  - 將 TfidfVectorizer 的特徵數降低至 3000，僅使用 unigram 和 bigram 特徵。
- 結果：
  - 訓練時間顯著減少，性能略有提升，但仍不足以滿足準確率需求。
  - 特徵維度降低後，數據的語義信息損失部分影響分類效果。

Step 3: XGBoost with Hyperparameter Tuning

- 方法：
  - 將模型改為基於梯度提升的 XGBClassifier。
  - 使用 RandomizedSearchCV 進行超參數調整，調整的參數包括：
    - learning_rate、n_estimators、max_depth、subsample 和 colsample_bytree。
  - 使用完整的 TfidfVectorizer 特徵提取，特徵數量限制為 5000，且使用 (1, 2) 的 n-grams。
- 結果：
  - 準確率明顯提高，但需要較長的訓練時間。

- 出現模型錯誤：XGBoost 模型要求數值類別標籤，導致最初的字符串標籤無法直接使用。

Step 4: Handling Label Encoding Issues

- 方法：
  - 使用 LabelEncoder 將字符串類別標籤（如 'anger', 'joy'）編碼為數字（如 0, 1, 2）。
  - 在生成提交文件時，使用 inverse_transform 將數值標籤還原為原始字符串標籤。
  - 確保數據處理後與 XGBoost 的輸入要求一致。
- 結果：
  - 成功解決了標籤格式問題，模型能夠正常訓練和預測。
  - 提交文件格式符合比賽要求。