

1. Dataset Overview:

這是一份從 Google map 評論區蒐集的文本資料集，用於 sentiment analysis 任務。

2. Data Type and Format:

Data type: text – star 文本對。text 是中文文本。star 是整數: 1 代表正面文本，0 代表負面文本。

File format: CSV file

3. Process of data collection:

使用 selenium 爬蟲程式到 Google map 的各個景點評論區中收集正反面評價的文本；以及使用 ChatGPT 生成部分 testing dataset。將收集到的文本以標點符號(e.g. 句點、驚嘆號、問號) 等將文本做進一步分割。

4. Dataset Size and Composition:

training dataset 約有 3200 筆資料，validation dataset 約有 350 筆資料，testing dataset 約有 470 筆資料。

5. Data Collection Conditions

評論的時間範圍: 2022 - 2025

地點: 位於台灣的各店家