你是什麼意思🤣🙄😠？

以分類模型建立表情符號推薦器

一、 動機與目的

人們使用通訊軟體傳遞文字訊息時，常在語句中加上表情符號。我們對其中的關聯深感興趣，因而想探討文字背後的情緒意義，並找出最貼近文字情緒的表情符號，製作表情符號推薦器。

二、 資料蒐集

1. 網路爬蟲 – 用於模型的訓練資料

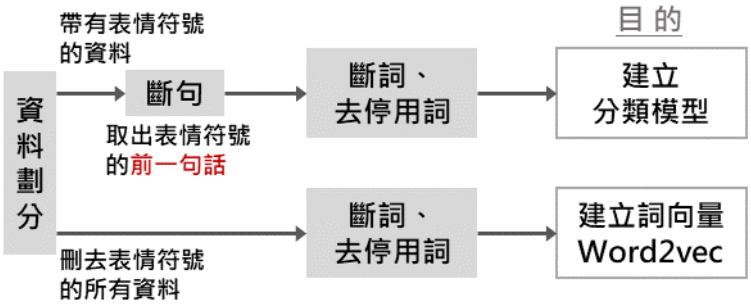
透過 Line 聊天紀錄以及爬蟲爬取 FB、IG 的貼文及留言，蒐集總計 1,018,226 筆資料。

2. 問卷調查 – 用於與模型分類表現比較

由測試集中隨機抽取 80 個句子，製作 4 份問卷，每卷各 20 句。填答者需填入心中覺得最適合之表情符號。回收樣本數為 70、62、100、66 份。

三、 資料預處理:

1. 文本預處理



2. 建立分類類別

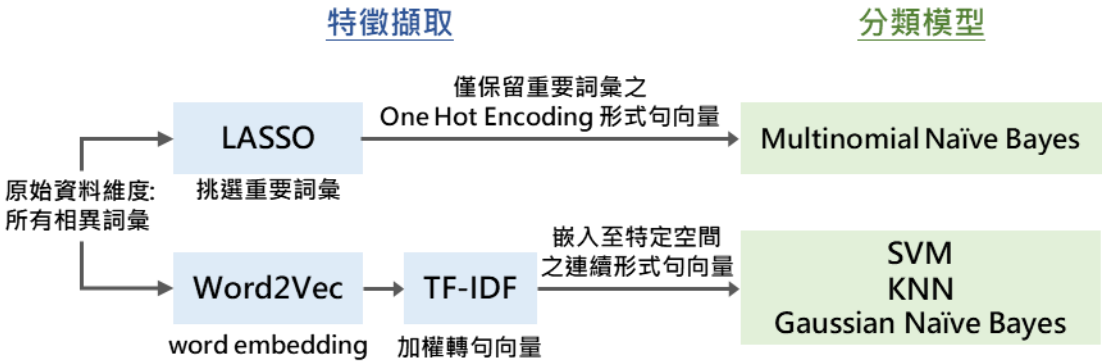
取出有情緒意義的 214 個表情符號，根據大眾使用習慣與其代表的涵義，依相似使用情境分為 40 個類別，此 40 類別即為分類器的 label。

3. 平衡處理與訓練集建立

使用以下兩種訓練集訓練分類模型:

- 原始不平衡訓練集: 模型針對大眾較常使用的類別有更好的辨別效果。
- 平衡訓練集 (將資料筆數較多的類別進行抽樣): 模型對於各個類別都具有一定的鑑別度。

四、 特徵擷取與分類模型



我們希望提供多種表情符號供使用者做選擇，於是改變了原始常見分為一類的做法，換成算出各個表情符號類別的發生機率。模型計算出機率前 3 或前 5 高的類別若包含正確答案，即當作回答正確，並以此評估模型的表現。

五、 研究結果與結論

本次研究建立多種分類模型，由不平衡資料訓練的模型為 KNN 的表現最好；而使用平衡資料訓練的模型為 SVM 的表現最好。

雖然 40 類表情符號中預測出 3 類中 1 類的正確率僅達 56.9%、5 類中 1 類的正確率僅達 68.9%，但不論是與問卷填答結果或是隨機猜測的正確率相比，都已經提升許多。

六、 討論與改善方向

1. 資料不平衡與分類組數過多

我們嘗試過以 Tomek Link、SMOTE、SMOTE+ENN 改善資料不平衡問題，也嘗試將組數調降為 20 組等方法，但模型表現仍不達預期。

2. 句向量資訊不足

以 t-SNE 演算法將句向量投射到二維空間，發現資料沒有依表情符號類別各自成群，推測相同表情符號的句子之間，可能沒有明顯的共同特徵，現有的特徵擷取方法還有進步的空間。另外，由問卷結果發現，僅以表情符號的前一句話判斷情緒略有不足，建立資料集時可考慮往前納入更多的句子。

3. 表情符號類別的建立

以階層式分群法將 214 個表情符號分為 5 大類，發現資料中表情符號間的關係與我們的認知有些不同，我們合併表情符號的方式，或許過於主觀。分組時可同時考量其於資料中展現的特性。