國立臺北大學

110 學年度專題報告競賽

你是什麼意思🤣🙄😠？

以分類模型建立表情符號推薦器

系所班別：統計系三、四年級

姓名學號：林貝爾 (410678061)

陳姵蓁 (410878001)

易祐辰 (410878034)

黃亮臻 (410878050)

彭　琳 (410878044)

黃淑郁 (410878054)

報告日期：2022 / 06 / 17

專題競賽報告分工表

學號	姓名	工作分配
410678063	林貝爾	資料蒐集、資料預處理、特徵擷取、書面報告
410878001	陳姵蓁	資料蒐集、資料預處理、特徵擷取、分類模型
410878034	易祐辰	資料蒐集、資料預處理、特徵擷取、分類模型
410878050	黃亮臻	資料蒐集、資料預處理、特徵擷取、分類模型
410878044	彭　琳	資料蒐集、特徵擷取、書面報告、簡報製作
410878054	黃淑郁	資料蒐集、資料預處理、分類模型、簡報製作

摘要

我們觀察到人們使用通訊軟體時，經常會在語句中加上對應的表情符號，除了能增加對話的趣味性，也能幫助人們在無法面對面的交談中，更貼切的表達自己的情緒以及理解對方背後的情緒，使得訊息傳遞地更加精確。本次研究目的是為了探討現今社會人們在網路上使用一段句子時背後所包含的情緒，同時也會配合現今熱門的網路用語和縮寫的詞彙找出可能代表的情緒及適合搭配的表情符號，以更貼切的表達出該句子真正的含義。

關鍵字：文本探勘 (Text Mining)、情感分析 (Sentiment Analysis)、LASSO Regression、Word2Vec、TF-IDF、單純貝式分類器 (Naïve Bayes Classifier)、K-近鄰演算法 (KNN)、支持向量機 (SVM)

目錄

壹、研究動機與目的

過去，人們透過書信進行文字訊息傳遞，由於文字本身沒有情緒，為了能在文字中更加有效表達出自己真實的語意，故而使用各式各樣的文體、敬語等方式，使閱讀者更能了解文字中的意思。現代社會中，人們所使用的文字與訊息內容也越來越精簡，而精簡的口語化表達讓語意有更多解讀的空間。因此，表情符號就成為了文字訊息表示情感的新方式。

在語句中加入表情符號，除了能夠增加對話的風趣度與互動性，也能幫助人們在傳達訊息時，可以更貼切的表達文字背後的情緒意義，使接收者更正確的理解對方想傳達的涵義。透過觀察現今世代的對話方式，我們對於研究詞句與表情符號之間的關連性，有了強烈的好奇與深入研究的動機。

本次研究首要目的是探討現今社會人們使用文字訊息時，背後所表達的情緒，並找出最貼近這些文字情緒的表情符號；其次，我們也會為現今熱門的網路用語和縮詞，找出可能代表的情緒及適合搭配的表情符號，並製作表情符號分類器。

貳、文獻回顧

「情感」 常因主觀意識的差異，而有截然不同的解釋。情感分析 (Sentiment Analysis) 運用文字探勘 (Text Mining) 技術，以機器學習方法對文件資料進行情感的偵測、萃取及分析。人們常說： 「機器人沒有情感」，而情感分析卻能使機器學習人類的情緒。

我們參考了政治大學資訊管理研究所林育龍先生的碩士學位論文 《對使用者評論之情感分析研究—以 Google Play 市集為例》，其研究針對 Google Play 商店的手機應用程式 (APP) 評論進行文字探勘，並利用中文情感分析判斷留言的正負向情緒。我們簡單整理了這篇論文與我們的研究差異，主要有兩點不同，分別是研究資料的取材內容、研究方式與情緒分類：

1. 此篇論文為 2014 年的研究成果，同時我們的資料類型也不限於與手機 APP 相關，因此在不同時空背景下，我們認為不同族群的留言對於字詞的使用方法與情感表達的方式會有所差異，這會造成研究結果上的不同。
2. 我們使用表情符號取代過往的正負向情緒，使詞句的情緒區分的更加細緻，

情緒種類也會更豐富。

參、 資料收集與預處理

本次研究蒐集兩份資料，第一份資料用以建立表情符號分類模型，以網路爬蟲方法蒐集；第二份資料為網路問卷的填答結果，以此資料與電腦訓練之分類模型進行比較。

一、建立表情符號分類模型之資料

(一) 資料來源介紹

我們的資料主要透過網路爬蟲的方式獲取。因應不同社群軟體的介面，除了抓下當前頁面視窗內的文字外，我們亦將程序設計成當電腦偵測到「顯示更多」、「查看更多留言」、「查看更多回覆」 等，與留言相關的按鈕時，會自動進行點擊；若已取得視窗內所有的資訊，也會自動下拉滾輪或自動點擊進入下一則貼文，以取得更多資料。此階段我們總計收集 1,018,226 筆資料，如下所述：

1. Facebook 以繁體中文為主要語言的粉絲專頁、公開社團以及個人主頁，爬取其貼文及留言，取出具有表情符號的句子，在全部資料中佔比為 45.44 %。
2. Instagram 以繁體中文為主要語言的公開帳號其發布的貼文，爬取其貼文及留言，取出具有表情符號的句子，在全部資料中佔比為 52.51 %。
3. Line 聊天紀錄，使用 73 個不同的聊天室，取出其對話紀錄中具有表情符號的句子，在全部資料中佔比為 2.05 %。

上述提及之粉絲專頁、公開社團以及公開貼文，我們將其類型區分為 10 大類，內容涵蓋：網路紅人、歌手藝人、金融政治、新聞媒體、體育……等不同面向，並設法使各大類型的樣本數達到平均（Instagram 與 Facebook 爬取之網站類型佔比詳見附錄一）。我們了解不同網站下留言的族群不同，他們使用表情符號的機制可能也會不同，但網站類型包羅萬象，因此我們選取了成員們喜歡的網站、較知名的公眾人物以及台灣訂閱數前百大的 YouTuber 其粉絲專頁進行文字探勘。

(二) 文本預處理

經由網絡爬蟲獲取資料後，我們針對原始資料進行預處理。首先，雖然我們爬取的資料都是以使用繁體中文為主要語言的網站，但其中仍有簡體中文的使用者，因此我們先將資料轉換為繁體中文再進行斷詞。

為了有效進行後續的分析，我們先針對資料進行劃分，再依需求進行處理，下圖 1 為文本預處理之流程圖：

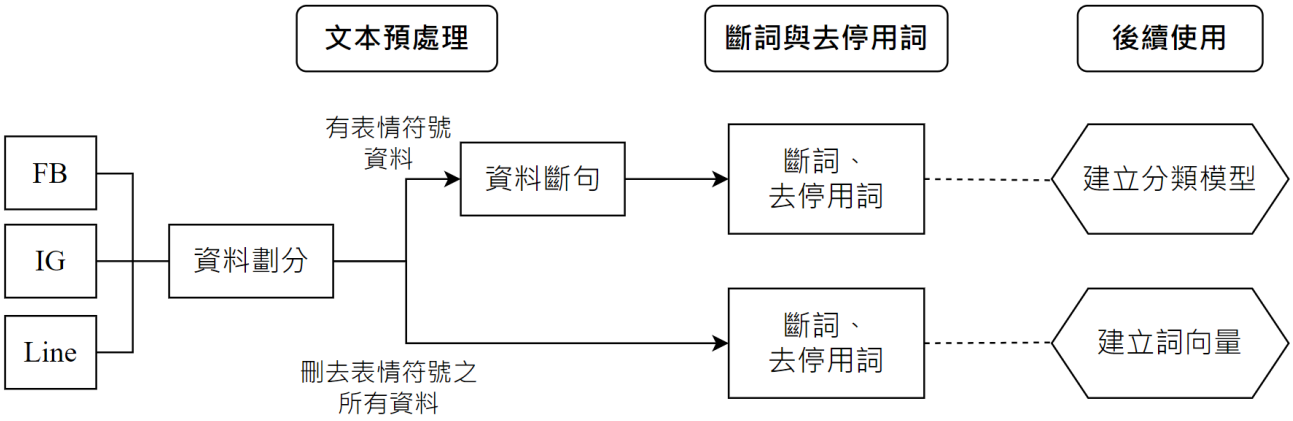


圖 1：文本資料預處理之流程圖。

1. 資料劃分

本次研究我們將原始資料整理成兩個目的不同的資料集：第一份資料集用以訓練 Word2Vec 模型，進行詞嵌入 (word embedding)；第二份資料集用以訓練表情符號分類模型。

(1) 訓練 Word2Vec 模型之資料集

訓練 Word2Vec 模型時，我們僅需要語句中的文字，並不需要表情符號。在原始資料中，資料的單位是一篇貼文、一則留言或是一則對話，此份資料不更動資料單位，只將原始資料中所有的表情符號刪除，留下文字。此份資料總計蒐集 828,848 筆。

(2) 訓練表情符號分類模型之資料集

根據大眾的使用習慣，我們認為表情符號會與前一句話的情緒最為相關，因此提取表情符號的前一句話作為分析單位。所以此份資料集需要將原始資料整理成「一個句子對應一個表情符號」的形式，而在原始資料中，資料的

單位是一篇貼文、一則留言或是一則對話，在此我們以 「逗號」、「句號」 或非繁體中文的字元作為切點，將其進行斷句，成為新的資料。我們認為表情符號應會與前一句話的情緒最為相關，因此提取表情符號的前一句話作為分析單位。斷句範例詳見附錄二。

2. 斷詞與去停用詞

我們所使用的斷詞工具是由中研院 CKIP Lab 中文詞知識庫小組所開發的繁體中文斷詞程式 CkipTagger。另外，我們也針對文本中的特殊詞、流行用語、公眾人物名字等等詞彙設定權重，增加斷詞的正確性。

「停用詞」 指的是經常出現在文本之中，但對於文意不會造成很大影響的詞彙。進行文字探勘時，為了能夠最大程度的凸顯文本特徵，通常會選擇過濾掉停用詞。使用之停用詞詳見附錄三。

(三) 建立表情符號類別

我們的資料包含 1,020 個表情符號，取出使用率較高、且具有情緒意義的 214 個表情符號，其餘表情符號則不納入本次研究的分析目標。我們根據大眾使用習慣與其代表的涵義，將具有相似情緒與相似使用時機的表情符號區分為 40 個類別（詳見附錄四），並以此 40 類別作為分類器的標籤 (label)。

(四) 資料平衡化

在我們的資料集中，40 類表情符號其分佈有嚴重的不平衡狀況（分佈圖詳見附錄五），最小的類別僅有 1,000 筆資料，而最大的類別有 111,024 筆資料。針對此狀況，我們將資料筆數大於 1,500 筆的類別抽樣 1,500 筆資料，其餘類別則不進行抽樣，建立了類別平衡之訓練集資料。

另一方面，雖然過度不平衡的資料會使模型難以分辨資料數較小的類別，傾向將資料分類至資料數較大的類別，但此 40 個類別有如此不平衡的分佈，表示大眾在使用表情符號時，的確較常使用前幾個類別的表情符號；又，資料筆數最少的類別有 1,000 筆，或許已足夠模型進行學習。因此，我們也使用此份不平衡資料訓練分類模型。與上段所述之平衡資料不同之處在於：使用平衡資料訓練可以使模型

對於各個組別都具有一定的鑑別度，而使用不平衡資料可以使模型針對大眾較常使用的類別有更好的辨別效果。

二、與模型分類表現比較之問卷調查

由於本研究缺乏以相同研究方法與對象研究之過往文獻，因此我們發放問卷，並以問卷填答結果與研究中建立的分類模型進行比較，希望可以看見對於表情符號的使用，人類判斷與電腦訓練之間有何差異。

我們從測試集中，自每一類表情符號隨機選兩題，總計抽出 80 個句子，隨機分成 A 卷至 D 卷共四份問卷，以線上問卷之形式發放，請填答者幫每個句子配適出一個最合適的表情符號，作為人們對表情符號的使用習慣依據。問卷回收份數為：A 卷 70 份、B 卷 62 份、C 卷 100 份、D 卷 66 份。

肆、研究方法

我們將研究流程分為兩個階段，將依序說明其研究方法。下圖 2 為研究流程圖：

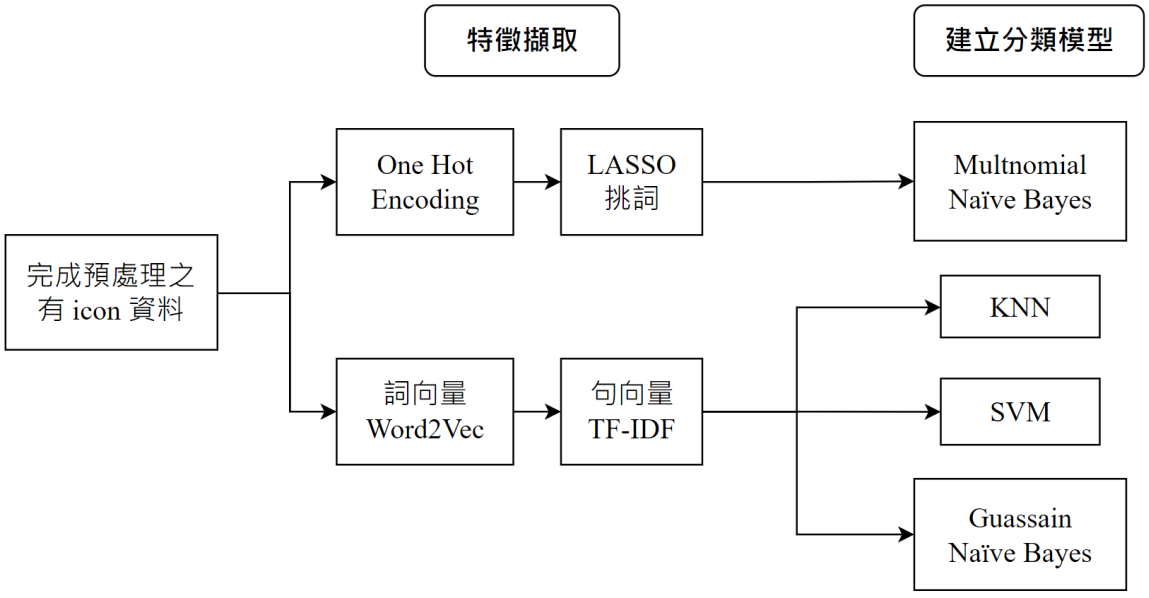


圖 2：研究流程圖。

一、特徵擷取

(一) Multinomial LASSO Regression

LASSO Regression (The Least Absolute Shrinkage and Selection Operator Regression，以下簡稱為 LASSO) 是一種正規化迴歸模型。其核心概念是在原配適標準上添加以迴歸係數絕對值大小表示的處罰函數，以此簡化模型。對於大規模數據變量的資料，LASSO 具有良好的變量選擇表現。

由於我們資料中表情符號為類別型變數，因此考慮以 multinomial variable 為反應變數的 Multinomial LASSO Regression，其估計式為，

$$\min_{\vec{\beta}} -\log(\text{likelihood}) + \lambda||\vec{\beta}||. \tag{1.1}$$

其中，$\vec{\beta}$ 代表自變數係數的向量，$||\vec{\beta}||$ 代表其 L1-norm。當懲罰係數 λ 設定越大時，越多自變項的係數會收縮至零。係數非零的自變項即可視為由模型選出的重要變數。

在我們蒐集的 82 萬筆資料中，涵蓋了 91,029 個詞彙，詞彙數量不但龐大且資訊十分稀疏。即使在刪除出現次數過少的詞彙後，仍剩下 9,808 個詞彙。因此，我們將每個詞彙以 One Hot Encoding 的方式代表其是否出現在一句子之中，將此視為自變數，便可使用 Multinomial LASSO Regression，在 9,808 個詞彙中，進一步篩選出更能有效連結表情符號的重要詞彙。

(二) 詞向量模型（Word2Vec）

Word2Vec 為一個淺層雙層的神經網路，透過學習後能夠將每個單詞都映射到一個特定的多維度空間。根據每個詞在空間中的座標，稱之為詞向量，我們得以看出詞與詞之間的關聯性與相似性。

我們使用的是：根據上下文的單詞來預測當前待預測單詞之機率的 CBOW (Continuous Bag-Of-Word) Word2Vec 模型。圖 3 為 CBOW Word2Vec 的架構圖。模型在進行訓練時，會經過幾次迭代，利用輸出層的損失回頭修正隱藏層的參數，而隱藏層的參數即為我們的詞向量。

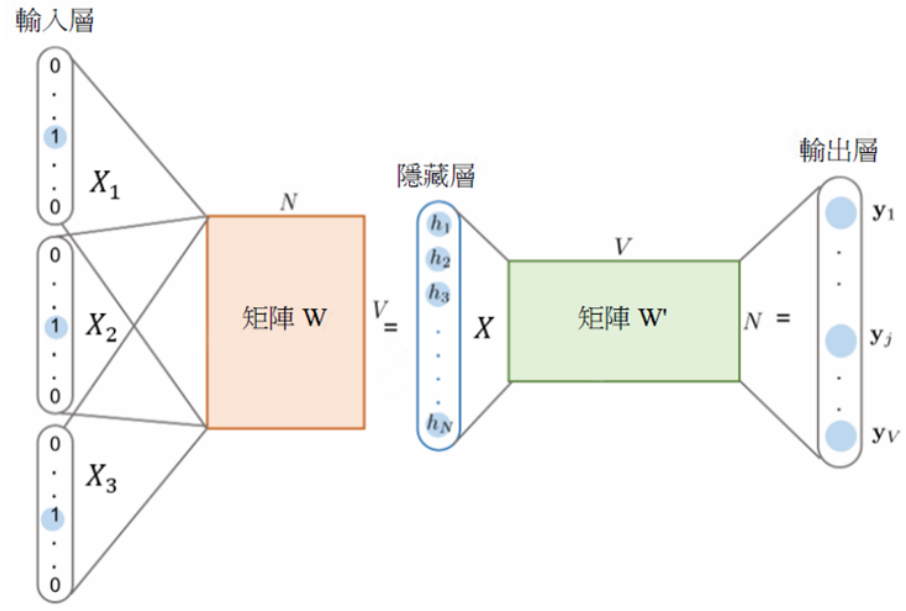


圖 3：Word2Vec CBOW 模型架構圖。

圖 3 主要參考資料：

Zhixing Lin, Lin Wang, Xiaoli Cui, Yongxiang Gu (2020) , Fast Sentiment Analysis Algorithm Based on Double Model Fusion.

訓練模型的過程中，我們著重的參數包含了：

1. 詞向量維度：維度太小無法有效表達詞與詞之間的關係，維度太大使得詞在空間中分布稀疏而難以找出規則。我們參考 Google 官方的相關研究，以及 Kaggle 競賽普遍使用的維度，選定 300 維為最後的詞向量維度
2. 模型的滑動窗口：代表形成詞組時，前後需採納多少個詞彙來預測目標詞。經由反覆測試，終決定將其設置為 9。
3. 最小詞頻訓練閾值：有助於過濾在文本預處理階段中被切錯的詞彙，也可過濾生僻的低頻詞，減少雜訊。根據文本預處理階段的經驗，我們將最小詞頻訓練閾值設置為 10，即：若詞彙出現次數少於等於 10 則予以剔除。
4. 負採樣：提供非上下文詞彙的樣本，稱為副樣本。Mikolov（2013）建議對於規模較大的訓練集選用 2-5 個負樣本，因此我們將負樣本參數設置為 5。

(三) 使用 TF-IDF 加權法轉換句向量

由於我們想要分析的對象為一個句子，因此得出每個 「詞」 的詞向量後，需要再一步將資料的詞向量轉為句向量。我們使用 TF-IDF 演算法對詞進行加權，將 Word2Vec 產生的詞向量轉為句向量。TF-IDF 法由 TF 及 IDF 兩個指標組合而成，用以衡量詞彙的重要性。

TF （Term Frequency） 計算某個詞在文本中出現的次數，即為該詞的詞頻。若詞在文本中出現的頻率越高，則表示其重要性越高。TF 值為將 「某詞彙 t 在文本 d 中出現的次數 （$n_{t,d}$）」 除以 「該文件中所有詞彙出現的次數總和（$\sum_{k=1}^{T} n_{k,d}$）」。其公式為下式 2.1：

$$TF_{t,d} = \frac{n_{t,d}}{\sum_{k=1}^{T} n_{k,d}} \tag{2.1}$$

IDF （Inverse Document Frequency） 計算包含這個詞的文本比例，若一個詞彙出現在不同文本的頻率很高，表示此詞彙對於該文本的代表性很低。IDF 將 「文本總篇數（D）」 除以 「這個詞彙出現過的文本篇數 （d_t）」。其公式為下式 2.2：

$$IDF_t = log\left(\frac{D}{d_t}\right) \tag{2.2}$$

將式 2.1 以及式 2.2 相乘可得該詞彙的 TF-IDF 值，即該詞彙的重要性。公式為下式 2.3：

$$TF - IDF_{t,d} = TF_{t,d} \times IDF_t \tag{2.3}$$

計算出 TF-IDF 值後，針對詞彙各自的重要性給予不同的權重，將 Word2Vec 的詞向量建構出句向量。而當詞向量轉為句向量的過程中，遇見 「未知詞」，也就是不存在 Word2Vec 詞袋中的詞彙時，我們以全零向量作為該字彙的詞向量，最後計算平均值以得到該句子的特徵向量。

二、建立分類模型

此次研究為機器學習中的監督式學習，分類句子的情緒，並將其與表情符號配對，我們使用 Naïve Bayes Classifier、KNN 和 SVM 此三個分類模型，並以 40 類表情符號為標籤 (label)，針對各個分類模型逐一測試，並比較各模型結果。

(一) Naïve Bayes Classifier

Naïve Bayes Classifier 模型假設所有的特徵都是獨立的，透過貝式定理的計算，我們就能得到在已知資料下哪個目標發生的機率最大，並由此去作出分類。在實際應用中，資料的特徵其實並不全然獨立，但因貝式分類器需要設定的模型參數較少，並且在大樣本數時，從樣本算出的機率不易有偏差，可以避免過度擬和的問題，故我們選擇貝式分類器作為本次研究的其中之一分類模型。下式 3.1 為貝式定理結合模型假設的 「所有特徵獨立」，其數學式：

$$P(y_i|x_1, x_2, \ldots, x_n) = \prod_j^n \frac{P(x_j|y_i)\,P(y_i)}{P(x_1, x_2, \ldots, x_n)} \tag{3.1}$$

當確定訓練集的同時，$P(x_1, x_2, \ldots, x_n)$ 即為已知的常數，故進一步將式 3.1 簡化成下式 3.2：

$$P(y_i|x_1, x_2, \ldots, x_n) \propto P(x_j|y_i)\,P(y_i)$$

$$\Rightarrow \hat{y} = \arg\left[\max_y P(y) \prod_j^n P(x_j|y_i)\right] \tag{3.2}$$

貝式分類器以最大後驗機率法 (Maximum a posterior，MAP) 估計 $P(y)$ 以及 $\prod_j^n P(x_j|y_i)$。而針對不同的 $P(x_j|y_i)$ 分佈狀況，貝式分類器再進一步地細分為多種模型，本次研究針對不同的資料形式選擇 Gaussian Naïve Bayes Classifier 以及 Multinomial Naïve Bayes Classifier 兩種模型進行訓練，其原理如下所述：

1. Gaussian Naïve Bayes Classifier（下稱：GaussianNB）

GaussianNB 主要使用於特徵為連續變數時，其假定特徵為常態分佈，以下式 3.3 計算 $P(x_j|y_i)$：

$$P(x_j|y_i) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(x_{ij}-\mu_{ij})^2}{2\sigma_{ij}^2}} \tag{3.3}$$

x_{ij} 為第 i 個類別的第 j 個特徵的資料，μ_{ij} 為其期望值，σ_{ij} 為其標準差，並以最大概似估計法估計 μ_{ij} 以及 σ_{ij}。

2. Multinomial Naïve Bayes Classifier（下稱：MultinomialNB）

MultinomialNB 主要應用於特徵為類別變數時，其以下式 3.4 計算 $P(x_j|y_i)$：

$$P(x_j|y_i) = \frac{N_{ij} + \alpha}{N_i + \alpha n} \tag{3.4}$$

n 類別個數，N_{ij} 為第 i 個類別其第 j 個特徵的資料總和，N_i 為 N_{ij} 之總和，即為第 i 個類別的資料筆數。針對 N_{ij} 為 0 的狀況，即第 i 個類別沒有出現具有第 j 個詞彙的資料，為避免後驗機率相乘為 0，進而忽略掉句子中其他詞對此類別的貢獻，因此我們設定 $\alpha = 1$，以避免機率相乘為 0。

(二) KNN (K Nearest Neighbor)

KNN 以多數決的方式，利用 K 個最鄰近的資料來判定新的資料是哪一群，在本次研究中，我們使用交叉驗證法，選出最佳的 K 值。公式為下式 4.1：

$$y = arg\left[max_{c_j} \sum_{x_i \in N_k(x)} I(y_i = c_j)\right],\ i = 1, 2, \dots, N \tag{4.1}$$

其中 x_i 為特徵向量，y_i 為類別，而 I 則為指示函數，當 $y_i = c_j$ 時，$I = 1$，否則 $I = 0$。

(三) SVM (Support Vector Machine)

SVM 通過核函式 (Kernel Function) 將低維空間不可分的樣本資料投射到高維度的特徵空間，通過訓練產生一系列超平面，利用超平面將不同的集合分類，並以間距最大化作為基準，找到資料的決策邊界 (Decision Boundary)。

此次我們選擇以 Gaussian Radial Basis Function kernel (RBF) 進行轉換，將資料投射到無限維。RBF kernel 公式為下式 5.1：

$$k(x, y) = e^{-\gamma || a-b ||^2} \tag{5.1}$$

其中 a, b 表示兩筆觀察值的位置，γ 表示兩點間 scale 的程度，這裡我們使用一般較常設定的 $\gamma = \frac{1}{2\sigma^2}$。由於 SVM 的預測結果為資料與決策邊界的距離 (margin)，輸出值沒有機率性質。而透過 Platt Scaling 可將不具備機率性質的 SVM 結果映射到 (0,1) 之間。Platt Scaling 公式為下式 5.2：

$$P(y_i = 1|f_i) = \frac{1}{1 + e^{Af_i+B}} \tag{5.2}$$

其中 f_i 為 SVM 的輸出值，A, B 為模型修正量。

在模型中 margin 大小的設定會影響到模型的精確度。soft margin 較能接受分類錯誤，傾向提供較寬鬆的決策邊界，而 hard margin 較不接受分類錯誤，傾向提供完全配適訓練資料的決策邊界。越大的 margin （即 soft margin） 較可以應對各種資料，但同時模型偏差會更大；越小的 margin （即 hard margin） 的模型偏差較小，但容易有過度擬合的問題。在此，我們會根據模型的表現，以交叉驗證決定 margin 的大小。

伍、研究分析與結果

在前章節中，我們介紹了 「資料平衡化」 以及 「特徵擷取」 的方法。我們結合了這些方法，訓練分類模型。此章節整理了不同分類模型搭配不同訓練集的分析結果。此外，我們的研究目的之一，是根據使用者輸入的句子推薦表情符號。由於每位使用者的偏好並不相同，因此我們的分類模型會推薦三或五類表情符號，讓使用者可以從中選擇。

這使得在訓練模型時，我們必須使用不同於一般評估模型學習表現時所使用的正確率 （accuracy）：若正確答案落在模型配適出的 3 類表情符號之中，我們則定義其預測正確，並以此定義計算正確率，以下簡稱其為 Acc3；依此類推，若正確答案落在模型配適出的 5 類表情符號之中，則計算其正確率為 Acc5。

一、LASSO 選詞搭配 Multinomial Naïve Bayes Classifier

我們將四十個表情符號分為十組，分別對於每一組做 4-class Multinomial LASSO Regression，在原資料 9,808 個詞彙中，篩選出重要的詞彙。受限於電腦記憶體容量，我們只能將資料超過 1,000 筆的表情符號類別，依其類別大小，隨機抽取 10%~100%，作為建模樣本。對於每一組的 LASSO Regression 模型，我們選取懲罰係數 λ，使得在模型配適後，非零係數的詞彙數量控制在約 500 上下。最後，將十組選出的詞彙取聯集，我們得到 2,587 個重要詞彙，資料中共有 735,048 筆資料 （句子） 使用了這些詞彙。

對於建立表情符號分類模型，根據是否以 LASSO Reression 做詞彙篩選以及是否採取資料平衡處理，我們建立了四個不同的訓練集 （如表 1 所示），並依序訓練 Multinomial Naïve Bayes Classifier。在測試集中隨機抽取 10,000 筆資料，模型表現如下表 2。

表 1：One Hot Encoding 形式訓練集。

訓練集編號	LASSO 選詞	平衡化	隨機抽樣	樣本數	變數個數
1.1	無	無	有	150,000	9,808
1.2	有	無	有	665,000	2,587
1.3	無	有	無	58,652	9,808
1.4	有	有	無	58,075	2,587

表 2：以不同訓練集訓練的 Multinomial Naïve Bayes Classifier 模型表現。

	使用不平衡訓練集		使用平衡訓練集	
LASSO 選詞	Acc3*	Acc5**	Acc3*	Acc5**
無	56.1%	66.6%	23.9%	38.6%
是	56.9%	68.9%	38.8%	48.7%

* Acc3：測試集中，正確答案落在模型預測出的 3 類表情符號中的正確率。

** Acc5：測試集中，正確答案落在模型預測出的 5 類表情符號中的正確率。

由表 2 可知，經由 LASSO Regression 降維後的資料，其正確率不論是 Acc3 或是 Acc5 均比不平衡資料高，但不平衡資料的 Acc3 以及 Acc5 上升幅度較小。

二、句向量資料搭配 KNN、SVM、Gaussian Naïve Bayes Classifier

透過訓練 Word2Vec 模型，我們得到各個詞彙的詞向量。下圖 4 選取了本次研究建立的部分詞向量，並使用主成分分析 （Principal Components Analysis），降至二維以視覺化。圖中呈現了 Wod2Vec 模型成功的將相近屬性的詞投影在空間中相近的位置。

接著，我們使用 TF-IDF 加權法，將詞向量轉換成句向量，得到共計 828,845 筆 300 維的句向量資料。我們將這份資料分割 10% 作為測試集，利用主成分分析將訓練資料由 300 維降至 20 維，再將測試資料降至同樣的 20 維主成分空間。

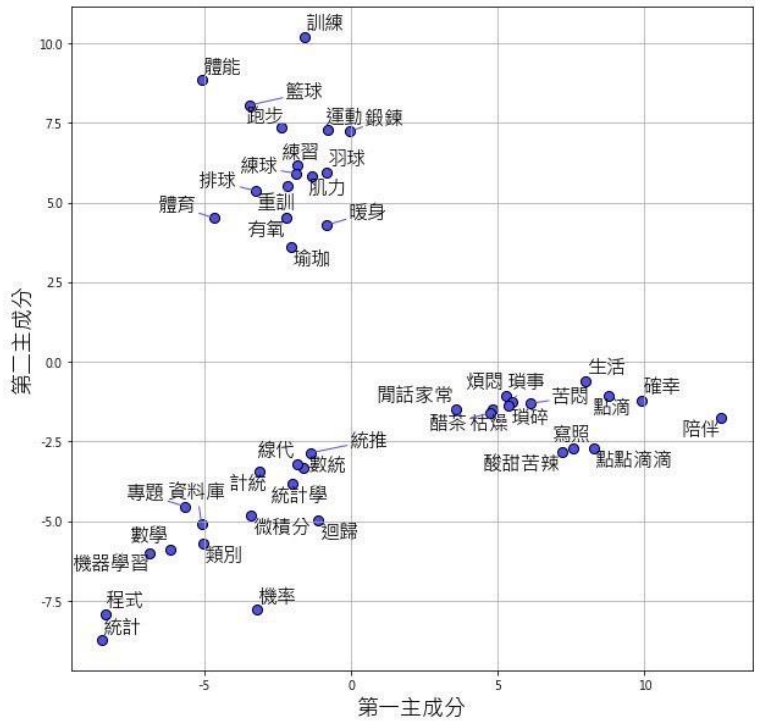


圖 4：詞向量視覺化圖。

根據是否對上述句向量資料作隨機抽樣以及是否採取資料平衡處理，我們建立了三個不同的訓練集（如表 3 所示），並依照合適的使用情境，分別訓練 SVM、KNN、以及 Gaussian Naïve Bayes Classifier （下稱：GaussianNB） 此三個分類模型。由於計算量過於龐大，我們只能隨機抽樣 10 萬筆資料進行 5-Fold 交叉驗證，選擇 SVM 以及 KNN 模型合適的參數； GaussianNB 因無需選擇模型參數，則不進行交叉驗證。模型表現如下表 4。

表 3：句向量形式訓練集。

訓練集編號	平衡化	隨機抽樣	樣本數	變數個數
2.1	無	有	300,000	20
2.2	無	有	745,961	20
2.3	有	無	58,291	20

表 4：以不同訓練集訓練的分類模型表現。

	使用不平衡訓練集		使用平衡訓練集	
分類器	Acc3*	Acc5**	Acc3*	Acc5**
SVM	53.3%	65.6%	36.4%	46.8%
KNN	55.0%	66.4%	34.8%	44.8%
GaussianNB[a]	45.8%	57.0%	28.6%	39.3%

* Acc3：測試集中，正確答案落在模型預測出的 3 類表情符號中的正確率。

** Acc5：測試集中，正確答案落在模型預測出的 5 類表情符號中的正確率。

a GaussianNB：Gaussian Naïve Bayes Classifier.

由表 4 可知，使用不平衡資料訓練的此三個分類模型，KNN 的表現最好；而使用平衡資料訓練的模型，SVM 的表現最好。GaussianNB 在兩種訓練資料中皆是表現最差的。本研究推論其可能的原因為 GaussianNB 對於特徵性質的要求，即特徵分佈為常態分佈的假設不成立。同時，我們將句子表示成連續變數，再將其離散化，這樣的過程可能導致資料的特徵變得較不明顯。

三、問卷調查結果

我們計算了問卷中 80 個題目的正確率，並與 SVM、KNN 以及 GaussianNB 模型在這 80 道題目上的正確率做比較，如表 5 所示。

表 5：問卷正確率與分類模型正確率之比較。

	Acc3*		Acc5**	
	使用之訓練集		使用之訓練集	
方法	不平衡	平衡	不平衡	平衡
SVM	35.0%	33.8%	40.0%	41.3%
KNN	20.0%	23.8%	25.0%	37.5%
GaussianNB[a]	20.0%	10.0%	30.0%	20.0%
問卷結果		20.0%		25.0%

* Acc3：測試集中，正確答案落在模型預測出的 3 類表情符號中的正確率。
** Acc5：測試集中，正確答案落在模型預測出的 5 類表情符號中的正確率。
a GaussianNB：Gaussian Naïve Bayes Classifier.

由表 5 可發現， SVM 的表現非常優秀，相較於人爲判別高出了 15 %。這可能表示我們的 SVM 模型藉由大量的資料，學習到了人們使用表情符號的習慣與偏好。而這些習慣與偏好恰好可能是我們自己都不了解或是從直覺上並沒有意識到的。

此外，我們從大眾填寫問卷的正確率與答題情況發現，有些句子只以表情符號的前一句話作為判斷依據，很難感受到其中的情緒。並且大家對每一個句子的理解各有不同，使用的表情符號也非常多元。因此僅抓取某個句子搭配的 「一個」 表情符號作為測試集唯一的標籤，可能不太合適。例如：「小貓好活潑啊」 此題在問卷結果中佔比最高的表情符號為 😍 （第 1 類），而我們的分類模型也預測其為第 1 類，但此題的正確答案卻為 😐 （第 20 類）。所以即使分類模型的預測結果與填答者的偏好相同，但因為與此題的資料來源者所使用的習慣不同，而在計算分類模型正確率時，被以分類錯誤計算。

陸、討論與後續研究建議

在這個章節中，我們針對前述分析研究不完善之處，提供改善方法以及後續研究建議。

一、資料不平衡狀況

原始資料的表情類別其分佈極度不均衡，我們認為這是影響正確率的重要原因，因此除了使用 「資料平衡化」 章節說明的以抽樣方法進行平衡處理以外，我們也嘗試透過其他欠採樣與過採樣的方式，以縮小多數類與少數類的數量差距，降低類別不平衡的影響。我們嘗試了以下三種方法處理句向量資料：

1. 使用 Tomek Link 刪除資料邊界處鑑別度不高的資料。
2. 使用 SMOTE （Synthetic Minority Oversampling Technique） 將低於資料筆數平均值的第 14 組至第 40 組，增加樣本至平均值。
3. 使用 SMOTE，初步增加資料數較小的組別其資料筆數，再使用 ENN （Edited Nearest Neighbor） 刪除資料邊界處的多數類樣本。

此三種處理方法的效果如表 6 所示：

表 6：不同平衡處理 KNN 模型表現。

	Acc3*	Acc5**
無	55.0%	66.4%
Tomek Links	54.9%	66.1%
SMOTE	44.1%	54.1%
SMOTE & ENN	14.6%	14.6%

* Acc3：測試集中，正確答案落在模型預測出的 3 類表情符號中的正確率。

** Acc5：測試集中，正確答案落在模型預測出的 5 類表情符號中的正確率。

由表 6 可知，將訓練集分別進行以上三種方法處理後，模型的表現皆比原始

不平衡資料差。此外，根據 SMOTE 以及 SMOTE 搭配 ENN 法的結果，我們推測同個類別的資料中可能存在著一定程度的分散，導致在合成少數類新樣本的過程中，產生了與多數類重疊的樣本，因而更難以被分類，正確率下降。

二、表情符號類別

表情符號類別數過多，可能也是影響模型表現的原因。因此，我們將欲分類的表情符號組數調降成 20 組，只取出資料量最多的前 20 類表情符號做為訓練集並重新進行分析。經此調降後的樣本資料類別仍存在不平衡的問題，但我們認為不平衡狀況已大幅改善，故不再針對此問題進行額外的處理。此 20 類的資料筆數共計 752,026 筆，隨機抽樣 10% 的資料做為測試集。使用此份訓練集訓練 SVM、KNN 以及 Gaussian Naïve Bayes Classifier，其模型表下如下表 7 所示：

表 7：使用 20 類表情符號訓練分類模型其表現。

分類模型	SVM	KNN	GaussianNB[i]
測試集 Acc3*	58.40%	**59.94%**	50.71%
測試集 Acc5**	71.50%	**72.51%**	64.00%

* Acc3：測試集中，正確答案落在模型預測出的 3 類表情符號中的正確率。

** Acc5：測試集中，正確答案落在模型預測出的 5 類表情符號中的正確率。

i GaussianNB：Gaussian Naïve Bayes Classifier。

由表 7 可知，由 20 個類別的資料訓練分類模型，雖然模型表現有提升，但整體正確率還是不甚理想，因此推測仍存在其他因素影響模型訓練結果。

三、句向量資訊不足

由各個模型的表現與上述處理資料不平衡的過程中，我們懷疑句向量資料比我們想像的還要分散，大眾使用表情符號的方式可能非常多元，於是我們希望將句向量資料降維進行視覺化，觀察句子間的關係。

t-SNE 演算法擁有將相像資料投射到更近的位置，不相像資料彼此投射到更遠的特性。我們認為此方法較能看出句向量之間的差異，因此這裡決定使用 t-SNE

演算法，將句向量投射到二維空間，並固定畫布，分別針對 214 個表情符號作圖。圖 5 為其中 4 個表情符號的視覺化圖：

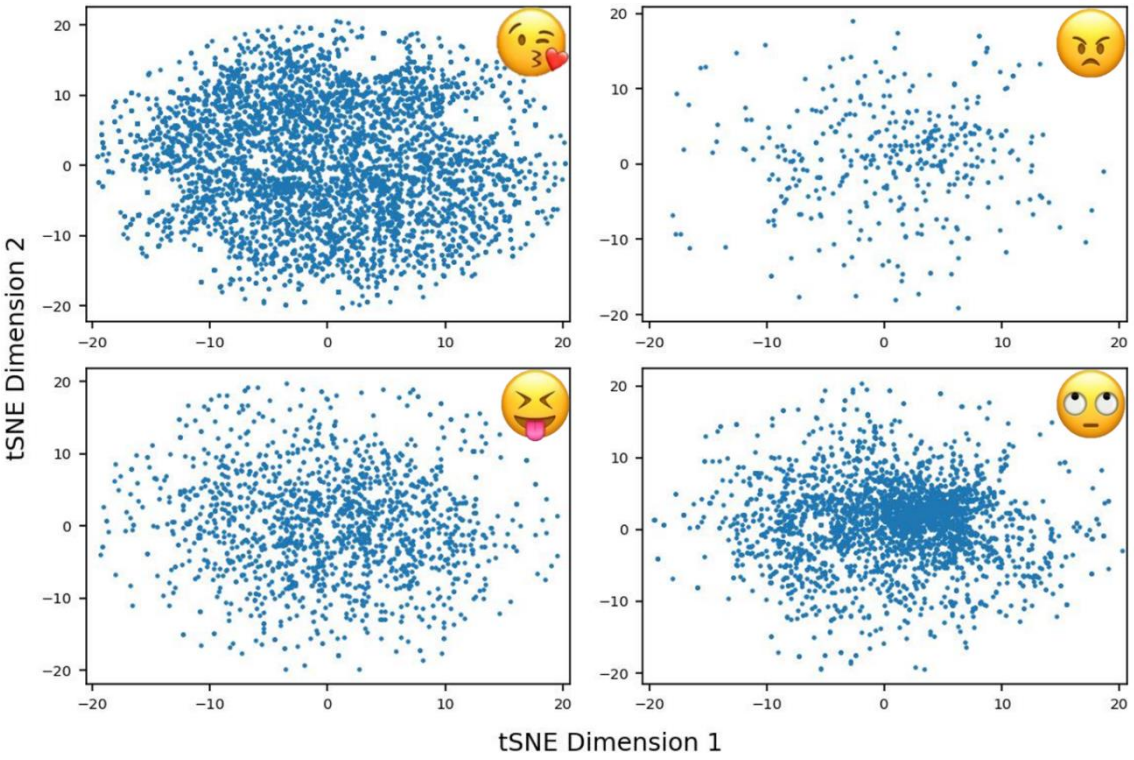


圖 5：依據不同表情符號的句向量資料，使用 t-SNE 演算法降維進行視覺化。

由圖 5 可知，在畫布中，資料並沒有各自於某個區塊成群的情況，我們推測相同表情符號的句子之間，可能沒有明顯的共同特徵，或是其實存在著共同的資訊，但在先前 Word2Vec 建構詞向量和建構句向量的過程中流失了這類資訊。

四、表情符號類別的建立

表情符號的類別，也會影響模型的表現。此處我們不採取主觀判定將 214 個表情符號分至 40 類的方式，而是使用資料本身的特徵做表情符號類別的判定。我們使用以 「經 LASSO Regression 特徵篩選的 One Hot Encoding 資料」，使用階層式分群法 （Hierarchical Clustering） 將此 214 個表情符號分為五大類，如下圖 6 所示。

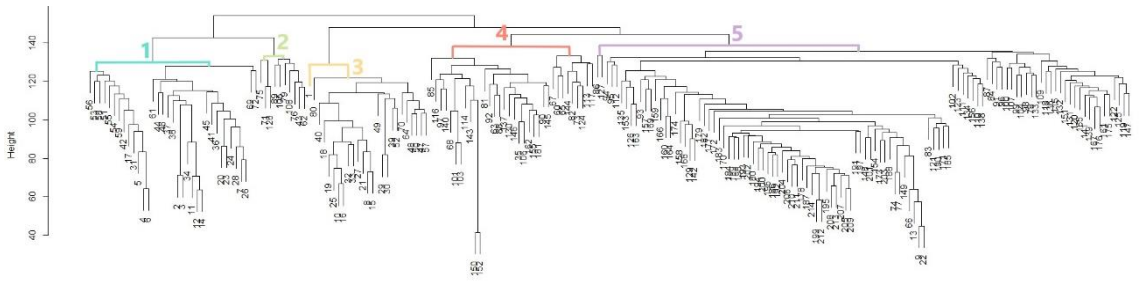


圖 6：使用分群法將 214 個表情符號分為五大類之樹狀結構圖

我們依比例抽取五大類資料，透過 LASSO Regression 篩選出重要詞彙，以 One Hot Encoding 的形式，使用 KNN 模型預測五大類表情符號。我們抽樣五大類資料共 9,000 筆，使用 KNN 模型搭配交叉驗證，其試行結果的準確率高達 99.9%（$K = 1$）。由此可知，分群法能捕捉表情符號類別間的差異，使得 KNN 能正確判斷資料屬於五大類的哪一類。

此外，由分群圖可以看出，資料中表情符號間的關係，與我們對表情符號的認知有些許不同。我們合併表情符號的方式，或許過於主觀，無法完全表現在實際資料上，有的甚至導致結果變差。

故我們認為後續研究可以結合以上的分群結果，結合研究者對表情符號情緒的認知，將大類的表情符號細分成更小類，並使用 Two-Stage 的方式進行預測，先預測資料為五大類的哪一個類別，再進一步預測其為此類之中的哪一個小類別。

五、透過詞彙代表性進行刪詞

進行 One Hot Encoding 時，我們認為 9,808 個詞彙量仍過於龐大，因此針對各個詞彙計算出「給定 40 類表情符號，出現該詞彙」的機率。一個詞彙共有 40 個對應到不同表情符號類別的機率值，我們計算了這 40 個機率值的變異數，若變異數很小則表示該詞彙對表情符號類別的代表性不足。因此我們將變異數最小的前 800 個詞刪去，降維至 9,008 個詞，並重新訓練 Multinomial Naïve Bayes Classifier，但卻發現正確率反而降低了。所以我們推測即使變異數較小，但該詞仍有提供模型資訊學習。

柒、總結

本次研究我們建立了多種分類模型，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。此外，爲了讓分類模型能提供多組適合的表情符號，我們改變了原始常見的分為一類的做法，改成算出各個表情符號類別的發生機率，讓使用者有多種選擇可以參考。對於本次研究不及之處，我們也有初步的推測，並以多種方法驗證，雖然無法提供一個完美的解決方法，但可作爲後續學者研究的方向。

捌、參考文獻

[1] 李朋軒。CkipTagger 開源中文處理工具。民國 109 年 9 月 10 日。取自：https://github.com/ckiplab/ckiptagger

[2] Tommy Huang。機器學習:如何在多類別分類問題上使用二元分類器進行分類 (Multiclass Strategy for Binary classifier) 。民國 107 年 3 月 日。取自：https://towardsdatascience.com/svm-and-kernel-svm-fed02bef1200

[3] Czako Zoltan (2018). SVM and Kernel SVM Tommy Huang. Retrieved November 13, 2018, form https://towardsdatascience.com/svm-and-kernel-svm-fed02bef1200

[4] Ajay Yadav (2018). SUPPORT VECTOR MACHINES (SVM). Retrieved October 20, 2018, from https://towardsdatascience.com/support-vector-machines-svm-c9ef22815589

[5] Alammar, J (2019). The Illustrated Word2Vec. Retrieved March 27, 2019 from https://jalammar.github.io/illustrated-word2vec/

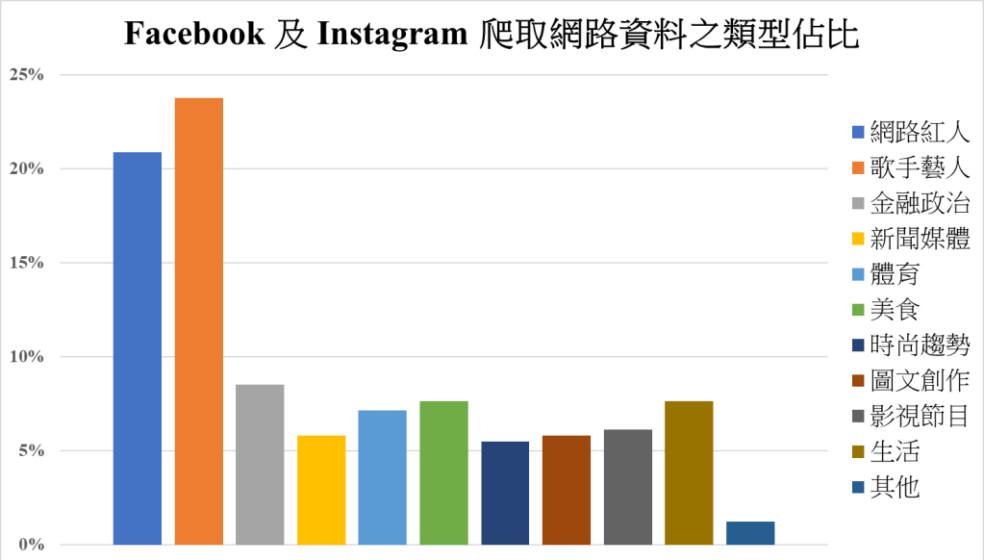
[6] Rana Singh (2019). Featurization of Text data. Retrieved September 12, 2019 from https://medium.com/analytics-vidhya/featurization-of-text-data-bow-tf-idf-avgw2v-tfidf-weighted-w2v-7a6c62e8b097

[7] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gre Corrado and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. *In Proceedings of the 26th international conference on Neural Information Processing Systems-Volume 2, pages 3111-3119.* (2013)

[8] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gre Corrado and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *In Proceedings of Workshop at ICLR. arXiv:1301.3781v1.* (2013)

玖、附錄

附錄一：Facebook 及 Instagram 爬取網路資料之類型佔比

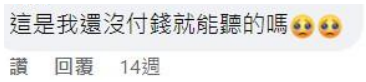


附錄二：斷句範例

➢ 範例一：以非中文字元的後一個字為開頭、表情符號為結尾，作為一筆資料

原始資料	魚丁糸專輯一定要是要買兩張以上的，聽一張收藏一張。這麼多年從來都是聽到無數人說沒買到你們的專輯有多遺憾的，從未聽過有買到會覺得後悔的😂 而且你們每次實體的精美程度是讓人捧到手中完全捨不得打開的那種，給人撲面而來的幸福感，每次都會驚歎于你們的用心！這次的復刻專輯是雙CD加魚丁秘密，真的沒有不買爆的道理啊，大家預購起來，一起買爆，不要給未來的自己留遺憾！ 讚 回覆 17週 2
斷句資料	從未聽過有買到會覺得後悔的 😂

- 範例二：一個句子後面接多個相同表情符號，則只取一個表情符號

原始資料	
斷句結果	這是我還沒付錢就能聽的嗎

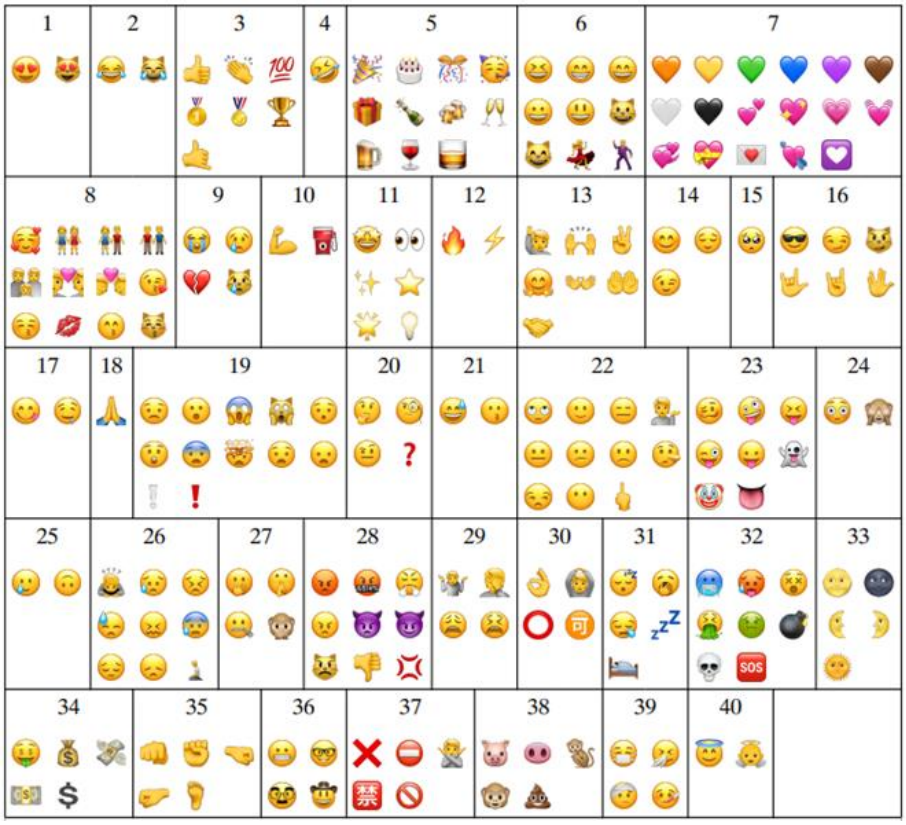
- 範例三：一個句子後面接多個不同的表情符號，則視為不同資料

原始資料	
斷句結果	感謝 感謝

附錄三：使用之停用詞清單

你	你們	妳	妳們	我	我們
他	他們	她	祂	它	牠
上	下	左	右	上面	下面
左邊	右邊	前面	後面	這	那
這個	那個	這裡	那裡	的	了
之	，	、			

附錄四：將 214 個表情符號歸納為 40 個類別



附錄五：原始資料集中 40 類表情符號之分佈