國立政治大學數位內容碩士學位學程

Master Program in Digital Content and Technology

National Chengchi University

碩士學位論文

Master's Thesis

基於感知屬性的品味分析：以酒飲網站專業與消費者評論為例

Taste Analysis Based on Sensory Characteristics: a Case Study Using Expert and Consumer Reviews from an Alcoholic Beverage Website

指導教授：黃瀚萱 博士

陳宜秀 博士

研究生：姚璇亨 撰

中 華 民 國 一 零 九 年 七 月

July, 2020

DOI:10.6814/NCCU202001553

致謝

兩年的時光轉眼即逝，最後能研究自己感興趣的領域並順利在數位內容碩士學程準時畢業，首先要感謝黃瀚萱教授的指導；若缺少了黃老師的引領，我或許也沒有機會認識自然語言處理領域的趣味。再來要感謝我的副指導，知識淵博、對學生最用心、嚴厲亦溫暖的陳宜秀教授。若沒有陳老師，別說無法沐浴於設計思考與人機互動之春風，我可能根本無緣經歷這段旅程。

找到兩位老師的時間其實略晚，所以我真心感謝老師們包容我的進度時快時慢，更不吝百忙中抽空，在重要時刻給予我關鍵意見，幫助我順利完成研究。這邊還要感謝前任學程主任紀明德教授的引介，讓我有機會認識學程、透過修課與老師結緣。另外也感謝學程的林靜媚助教，一直以來給予重要的協助與提醒。

再來，我想感謝學程的諸位同學陪我走過這段學習歷程，彼此交流與互相協助，留下難忘的校園回憶，尤其是一同受教於黃老師與陳老師的鎮宇和硯筑，沒有你們的陪伴與協助，我應該很難這麼順利完成研究。這邊也感謝政大資科系語言與知識科技實驗室的夥伴們，雖然只相處了一年的時間，但留下很多難忘的一起奮鬥的回憶，你們是很重要的夥伴。

最後，我要深深感謝我的家人與身邊的朋友，以及旅程中曾經陪伴我的人們，謝謝你們認同我重返校園的決定，始終給予我最溫暖的支持與鼓勵。

國立政治大學 National Chengchi University

DOI:10.6814/NCCU202001553

目　次

第一章　緒論 5

第二章　文獻探討 10

2.1　感官分析研究 10

2.2　情感計算 11

2.3　以深度學習為基礎之情緒分析 13

2.4　多類別分類任務 18

第三章　研究方法 20

3.1　研究方法架構 20

3.2　資料集 23

3.3　多類別分類模型設計 27

3.4　多類別分類模型評估標準 29

3.5　感官分析實驗設計 31

3.6　感官分析實驗評估 34

3.7　Yelp 資料集測試實驗設計與評估 37

第四章　研究結果 40

4.1　多類別分類模型評估 40

國立政治大學 National Chengchi University

DOI:10.6814/NCCU202001553

4.2 電腦模型預測與一般消費者評測統計結果分析 43

4.3 電腦模型預測與一般消費者評測感知模式相關性分析 46

4.4 Yelp 資料集測試結果 48

第五章 結論 51

參考文獻 53

國立政治大學
National Chengchi University

DOI:10.6814/NCCU202001553

摘要

對食物、飲料的品味行為是人類文明發展與日常生活的重要元素。現今，基於機器學習與深度學習技術的發展與快速累積的多媒體資訊，透過輸入影像、聲音資料進行學習，便能為電腦模型建立對視覺、聽覺屬性一定程度的辨識與預測能力，相較之下，針對味覺與嗅覺屬性的相關研究則略顯缺乏。現今主流的味覺、嗅覺相關感知研究多來自食品科學界廣泛應用於消費品產業之感官分析方法，由受過專業訓練或業餘的測試者針對研究目標進行感官品評，以主觀感受給予評論、評分或排序，並建立對應之感官屬性資料以進行研究分析，然而，其研究成果易受限於評論者的主觀偏好與感知能力，在客觀性與通用性方面仍有改進空間。本研究以網際網路酒飲網站所提供專家感官品評資料與消費者評論文本為基礎，透過自然語言處理之情緒分析技術，建立兼具相對客觀參考價值與主觀感覺意義的分析途徑，並延伸應用與於公開資料集進行測試，期透過網際網路數位內容文本與資料探勘技術的整合運用，為品味相關研究領域帶來更具通用性與成本效益的研究方法。

關鍵字： 自然語言處理、意見探勘、文本情緒分析、深度學習、情感計算、多類別分類、感官分析、感官品評

國立政治大學 National Chengchi University

DOI:10.6814/NCCU202001553

第一章 緒論

品味（Taste）一字有兩種定義：一為人類口部與喉部和物質接觸後，透過感知（Perceive）所產生對味道（Flavor）的感覺（Sensation），即味覺。二為對於事物品質之美學標準辨別能力[3]。本文所探討之品味範疇乃屬於前者。

味覺是人類感覺系統（Sensory system）的一部分，其運作模式為透過位於人類的口腔內味蕾上的味覺受器與物體中的特定化學分子結合，與嗅覺、以及部分觸覺機制共同作用後，產生電流訊號刺激大腦皮質中負責處理味覺的區域而產生相對應的感覺[4]。味覺有五種基本的感知屬性，分別為：甜味（Sweet）、酸味（Sour）、苦味（Bitter）、鹹味（Salty）、鮮味（Umami），各由不同的味覺感受器來接受處理，透過彼此之間及與其他因素的組合與交互作用，便產生了各種不同的味覺屬性[5]。

嗅覺在味覺感知中扮演著舉足輕重的角色，嗅覺系統（Olfactory system）的原理和味覺類似，是由位於鼻腔內的氣味受器與特定氣體化學分子結合後，傳遞神經訊號給大腦處理而產生對應的感覺。不過，和相較於味覺的近距離接觸特性，嗅覺是屬於較長距離的一種感知屬性，其向度也複雜許多。一些品味相關研究亦發現，嗅覺系統中的鼻後嗅覺（Orthonasal olfactory），對於味覺感知有著極為顯著的影響。[6]

對於味覺與嗅覺的研究在感知屬性的分析上有兩個主要途徑，一派為以化學與物理學為基礎，透過專業儀器測量研究標的成分儀器量測法（Instrumental measurement）其取得成分數據後進行通稱為儀器分析（Instrumental chemistry）

DOI:10.6814/NCCU202001553

之量化研究；另外一派則是以社會科學與認知心理學為基礎的感官量測法（Sensory measurement），其通常由測試人員對研究標的進行感官品評（Sensory evaluation），即實際品嚐並依其感覺來撰寫文字評論或屬性評分，以感官品評產生之文本與數據資料結合質化與量化方法進行之研究通稱為感官分析（Sensory Analysis）。儀器研究與感官分析兩種主流味覺、嗅覺研究途徑各有其優缺點。儀器分析一般被認為是最具客觀性、能代表獨立事實的研究方法，除實驗執行成本相對低廉、所需時間較短，產出成果亦兼具高準確度與高精密度[1]之優點，能取得較有通用性的研究成果，其缺點是無法將物體客觀性質與個體主觀感知的互動納入考量，故研究結果常缺少實務應用價值；感官分析則能在一定程度上對實際的人類知覺反應進行測量，實驗數據亦具有不錯的準確性，且研究成果與實際應用高度相關，但其缺點為精準度普遍低落，實驗結果受測試者個人之主觀偏好、感知能力等特質差異影響極大，不同測試者即能造成實驗結果極大的變異，使研究成果較缺乏通用性[7]，且其實驗執行成本高，產出成果所需時間亦相對較長[8]。

人類透過知覺系統接受資訊、透過大腦再詮釋所產生的各種感覺，本質上難以完全反映客觀事實[9]，若從社會心理學的角度來看，個體的感覺透過所屬社會脈絡的文化意義之主觀詮釋，產生不同屬性的情緒（Sentiment），個體會再於基此一情緒給予感覺刺激不同程度正面、負面或中性之評價（Evaluation）[10]，承

1 準確（Accuracy）是每次獨立測量間平均值與已知數據真值間的差距；精密（Precision）則是實驗數據多次量度結果的一致程度。

DOI:10.6814/NCCU202001553

載此一評價的文本資料，即為感知分析研究之基礎，因此，如何提升感知分析研究之客觀性與精密度，以期加強研究成果的普遍性，亦為當代味覺研究領域亟需解決的問題之一。

近年來，拜網際網路發展之賜，文字、影像及聲音等多媒體資訊數量以前所未有的速度倍增，除帶動大數據（Big data）風潮的興起，亦推進了資訊科學機器學習（Machine learning）與深度學習（Deep learning）領域的快速發展，其相關技術已被廣泛應用於分析影像資訊之電腦視覺（Computer vision）、分析聲音資訊之音訊處理（Audio processing）及分析文字資訊的自然語言處理（Natural language processing）等相關領域之中，與多媒體資料和人類感覺互動有關的情緒分析[2]亦為熱門研究主題之一，新興之情感計算[3]領域也隨之快速發展。受惠於文字訊息所能承載之豐富情緒意涵，自然語言處理於情緒分析之研究發展已十分成熟，甚至能夠以其輔助其他領域之研究。音訊處理領域的相關研究指出，在音樂情緒分類[4]任務上，結合由歌詞文本情緒分析所提取之特徵（Features）進行之多模態機器學習[5]，可大幅提升分類模型的成效[11]，在一些橫跨多種音樂情緒類別的任務上，單獨使用歌詞文本特徵訓練模型的成效甚至凌駕於單獨使用音訊之

2 情緒分析（Sentiment Analysis）一般用來判斷資料的特點段落可能給人類感覺帶來的情緒極性（Polarity）程度，如：正面、負面或中性等之機器學習任務。

3 情感計算（Affective Computing）為結合資訊科學與認知科學領域，以發展能夠處理並模擬人類情感運作之系統的跨領域學科。

4 音樂情緒分類（Music Emotion Classification）為透過音訊處理分析來辨別特定音樂段落的情緒屬性，如：開心、悲傷、憤怒等之機器學習任務。

5 多模態機器學習（Multimodal Machine Leaning）於資訊科學領域指整合不同形式多媒體資訊進行的機器學習，如：文字加影像、影像加聲音等。

DOI:10.6814/NCCU202001553

上[12]。電腦視覺領域相關研究亦指出，結合文本情緒分析之多模態機器學習對於模型分析成效確有顯著之正面效果[13]，可見自然語言處理對文本資訊的情緒分析能力確實有助於對感知屬性資料的分析。味覺與嗅覺雖同為人類感知屬性，但機器學習與深度學習於味覺、嗅覺之相關應用，相較視覺、聽覺等其他感知屬性來說仍為少數，且現有應用研究大多集中在儀器研究之範圍，對需求極大的感官分析著墨甚少。然而值得注意的是，近年來隨著使用者生成內容(User-generated content）的流行與深度品味體驗的普及，從事感官品評之專業評論者與業餘評論者之間的能力差距正逐漸拉近[14, 15]，隨之大量產生的網路食評、酒評等評論文本也提供了更多自然語言處理技術的應用機會。自 Google 公司於 2017 年發表基於注意力機制（Attention）[6]的遷移學習（Transfer learning）[7]的”Transformer”模型[1]，並於隔年推出以其為基礎的 BERT[8]模型，為自然語言處理領域帶來革命性的創新之後，自然語言處理模型的辨識能力已有長足之進步，其具有對原始文字資料直接處理的能力，且無須事先標記詞性，對學習資料型態的限制大幅減少，應用範圍也隨之擴大不少。其後運用類似機制之改良模型如 XLNet[9] 等亦如雨後春筍般出現，使得 Transformer 模型幾乎已有一統自然語言處理領域之姿。因此，

6 深度學習模型機制之一種，參考人類從大量資訊中篩選重要資訊的注意力運作原理所設計。

7 機器學習的研究領域，專注於將用以解決現有問題的模型運用於其他相關的問題之上。

8 Bidirectional Encoder Representations from Transformers 之縮寫，為一自編碼預訓練（Autoencoding Pretraining）模型，改善了傳統深度學習模型無法充分運用上下文訊息，以及受長距離影響的缺點。

9 為 Google 公司 2019 年發佈之自回歸預訓練（Autoregressive Pretraining）模型，相較 BERT 對長文本有更好的學習能力。

DOI:10.6814/NCCU202001553

在可預見的未來裡, 擴展自然語言處理應用範疇的重要性將逐漸超過不同深度學習模型之間的效能比較。

本研究將嘗試整合以上技術並轉化現有問題, 將與品味相關之網路評論文本以自然語言處理之情緒分析方法處理, 把不同味覺感知屬性的品評視為多類別分類（Multiclass classification）任務[10]，以 Transformer 之 XLNet 模型進行訓練並比較預測成效。本研究預期之主要貢獻如下:

一、 透過對品味相關資料的研究, 進一步開拓機器學習與深度學習技術於人類感知屬性之研究範疇。

二、 訓練理解能品味意義的自然語言處理模型, 用以協助加強其他不同相關情緒分析任務, 如: 餐廳評論分析、食品類產品意見分析等分類模型之成效。

三、 為感官分析領域提供新的補充性研究方法，提高成本效益，並加強研究成果之客觀性與通用性。

[10] 即 Multiclass Classfication Task，為分類任務之一種，和二元分類（Binary Classification）與多標籤分類（Multilabel Classification）的差別在於，有多於一種的類別或標籤需要被分類，但在所有不同的類別（Class）當中，只能有一個是被標記正確的，可視為多標籤分類的一種特例，在此情形下則類別與標籤（Label）意義相同。

DOI:10.6814/NCCU202001553

第二章 文獻探討

2.1 感官分析研究

根據學者 Minim 等人的研究整理[16]，當今主流的感官分析研究方法有：風味圖譜法（Flavor profile）、全質構分析法（Texture profile）、定量描述分析法（Quantitative Descriptive Analysis）、光譜分析法（Spectrum）、自由圖譜法（Free profile）等。不同方法代表著不同的途徑，但實務上最常見的是根據實際需求混合不同方法的通用性描述分析法（Generic descriptive analysis）[17]。上述研究方法的流程內容大致相似，除自由圖譜法採用未經訓練評測者，任其自定義味覺、嗅覺屬性與專注在文字描述與排序外，在其他方法中，評測者皆會需要經過一定時間的事前訓練，並被要求具備相當程度的味覺敏銳度。在訓練的過程中，評測者會以質性方式以文字形容其感官知覺，並結合過往經驗以量化方式進行屬性強度的排序。受過良好訓練的評測者須能充分使用共通的質性語句並與量化參考指標以建立相對應的量表。近年來，許多研究[18]專注在如何改良感官分析研究方法在耗費時間與結果可靠性上的兩難。新近改良方法如 Flash profile[11]、Check-All-That-Apply 因而興起，其中 Check-All-That-Apply 方法雖發展較晚，但已快速成為研究消費者對產品知覺偏好的重要工具[18]，其採用類似問卷研究的調查方式，評測者不須經過事前專業訓練，亦無須進行量尺評分，僅須在理解問卷所給定感知屬性詞彙的意義後，從中勾選覺得符合

[11] 為 Free Profiling 衍生之方法，主要差異在於評測者須就所給予之量表作量化強度評分。

DOI:10.6814/NCCU202001553

之項目即可，有著容易執行的優點，然而也有結果相對缺乏精確性之缺點。一些結合化學儀器分析或電腦資料探勘技術[19]的研究途徑亦被提出，為味覺與嗅覺領域持續帶來新的發展。新興研究方法在減少實驗成本或加強客觀參考性方面皆各有其助益，但往往犧牲了研究針對問題的準確性，以及與實際消費者感受的相關性，如何將實驗所本之的質性敘述文字與量化屬性量表進行有意義且較不受個體差異影響的整合運用，仍是感官分析領域未來的挑戰之一。

由以上回顧可見，感官分析的量化研究途徑雖能運用統計學模型來對量表數值分析，但份量同等重要的質化文本資料因其非結構化的性質，似未被充分利用。然而，近年在資訊科學領域裡，隨著模擬人類大腦神經系統的深度學習技術逐漸流行，讓電腦模型對非結構化資料之分析能力大為提升。若能整合資訊科學領域中其他與人類感知屬性相關的研究，直接從消費者生成之大量文本資料中探勘更多感知訊息，以代替傳統的消費者調查量尺數值評分，預期將能替以味覺、嗅覺研究為主的感知分析領域帶來不少正面貢獻。

2.2 情感計算

情感計算是一門新近的跨領域發展學科，整合了資訊科學、認知科學與心理學之相關研究，起源可追溯至 19 世紀時心理學家與哲學家對於人造物擁有情感之可能性的相關研究。其目的在於探討如何賦予電腦近似於人類的心智能力，包括觀察、詮釋，甚至是情感反應，藉由提升人機互動的品質，讓電腦的人工智慧更接近人類的智慧，以創造更和諧的人機互動設計[20] 。

DOI:10.6814/NCCU202001553

在電腦視覺領域，與情感計算相關的研究一般是透過捕捉特定影像資料中人物面部表情的相關特徵，或是身體與手勢的運動模式，透過建立卷積神經網路[12]之深度學習模型學習特徵其與特定情感向度的正負關聯，並整合人類視覺的認知科學原理後，建立對各種影像資訊情感屬性的辨識能力。[21]與聽覺相關的音訊處理領域，與情感相關的研究大多集中在對具有情緒性質的話語（Emotional speech）或音樂等資料的處理，主要透過對特定音訊資料的聲學屬性，如：音調、響度等進行特徵捕捉，並同樣建立深度學習模型以進行和情緒向度關聯之學習與預測，除運用認知科學的聽覺原理外，亦有整合周邊語言學[13]、音樂學等相關領域之研究[22]。然而，如何正確地萃取出有助於情感辨識的感官屬性特徵，無論是在電腦視覺或是音訊處理領域，至今都被認為是很大的挑戰。

有鑑於此，相關領域的研究視角開始轉向透過多模態學習的模型設計，整合不同感知屬性的資料來加強電腦對情感的辨識能力[2]。其中，文字形式的資料由於本身承載了豐富的情緒意義，隨著 Youtube 等影音平台的流行，大量描述影像及聲音內容的文字資料更隨之同步出現，加上處理非結構文本資訊的自然語言處理技術發展十分成熟。因此，如何整合將文字與不同感官屬性之資料整合運用，已成為情感計算領域的新顯學。本研究也透過對文本情緒分析技術

[12] 即 Convolutional neural network，縮寫為 CNN。為一能運用二元資料結構，由一個或多個卷積層和頂端的全連通層組成之神經網路，在影像與語音辨識上被廣為應用。

[13] 周邊語言學（Paralinguistic）為研究語音之聲學屬性與在溝通訊息傳遞上所代表意義的語言學分支學科。

DOI:10.6814/NCCU202001553

的進一步探討，探索更多結合文字資料與味覺、嗅覺相關的感官屬性研究之可能性。

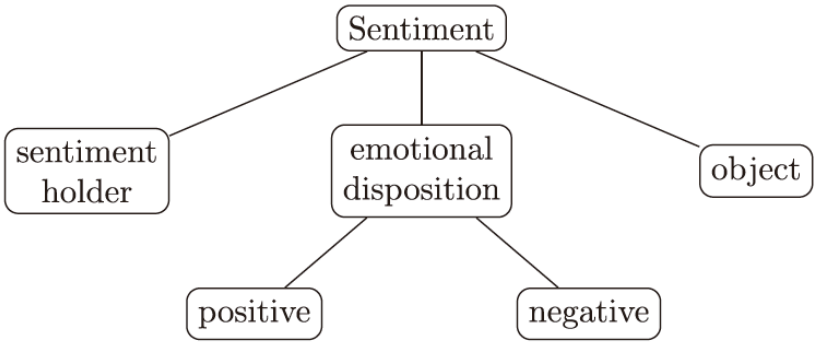
2.3 以深度學習為基礎之情緒分析

情緒分析為透過探勘多媒體資料的探勘，來挖掘個人對於特定實體（Entity），如人、事、物甚至議題等的感受。在全球資訊網（World Wide Web）與社群媒體（Social media）發展的推波助瀾下，一般個人得以透過文字、影像聲音等資料形式，大量生產表達自我感受的內容並廣為傳播，多元且大量資料為資料探勘技術提供相當豐富的應用發展前景。[23]目前情緒分析的應用範疇相當多元，舉凡金融業之股票市場趨勢分析、出版、音樂、電影等內容產業、電子產品與民生消費品產業的使用者意見研究[24]，甚至是政治領域針對選舉或公共政策的民意調查等，均能看見其涉足之蹤跡[25]。

現今電腦資訊科學在機器學習與深度學習技術的發展與運用下，對不同形式資料進行人類情緒分析的相關研究已有不少成果，而自然語言處理領域對文本資料進行深度學習之情緒分析，可說是目前與情感計算相關科目中發展最為成熟、完整的領域。

DOI:10.6814/NCCU202001553

圖 2.3-1 傳統情緒分析中情緒（Sentiment）之基本架構[2]

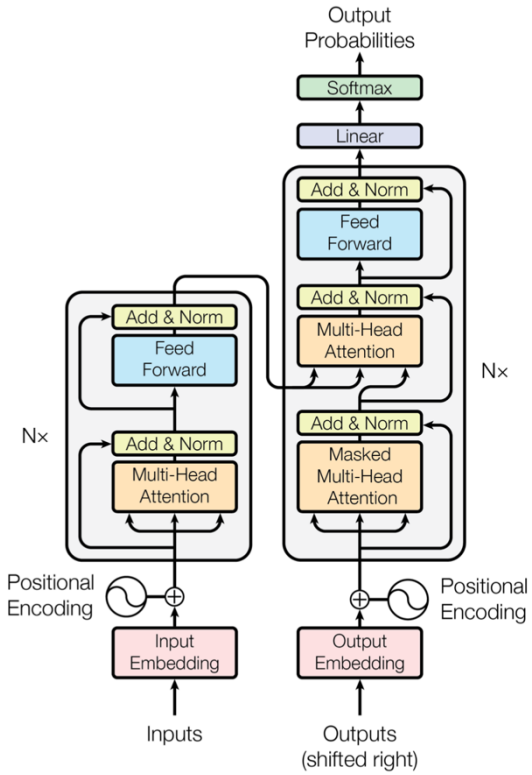


對文本的分析傳統上以針對文句的句法(syntactic)與語意(semantic)為主[26]。目前大多數主流的研究途徑是將事先標註情緒屬性標籤文本資料輸入深度學習模型，進行監督式學習（Supervised learning）。然而，在現實世界有著大量文字資料是沒有經過事先標註的，為了充分運用這些資料已解決真實問題，一些新的研究方法類型，如透過現成之辭典（Lexicon）或專家以其專業知識協助事先定義詞彙，以幫助提升模型成效之半監督學習（Semi-supervised learning）方法亦有長足的發展[27]。在 2017 年，Google 的 Vaswani 等學者發表了為深度學習領域帶來了重大影響的論文《Attention Is All You Need》[1]，其中提出了運用注意力機制的 Transformer 模型。Transformer 模型之架構以是一以編碼器-解碼器（Encoder-Decoder）結構為主，透過輸入一連串的資料序列，在每一個階段輸入皆會以自迴歸（Auto-regressive）方式運用前一個資料的資訊學習彼此之間的關聯。其運用了同為 Google 所提出之文字嵌入（Word embedding）技術[28]，將文字映射到特定向量空間，以向量形式來分析彼此的

DOI:10.6814/NCCU202001553

關聯及相似度，大幅減少了文字形態差異與歧異性對傳統自然語言處理所造成的障礙。

圖 2.3-2 Transformer 模型架構圖[1]

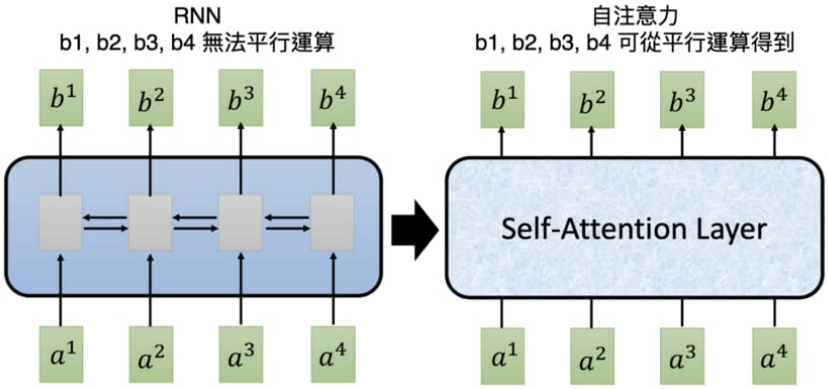


Transformer 模型採用的的注意力機制最早為電腦視覺領域所開發，參考人類大腦讓視覺透過聚焦在重點訊息，忽略其他不重要部分以節省認知資源的注意力原理進行設計，改良了傳統進行序列形式資料深度學習之遞迴神經網路模型（Recurrent Neural Network, RNN）只處理前一個序列之詞向量的缺點，注意力機制將編碼器所生成之所有詞向量全部提供給解碼器，讓解碼器根據不同任務特性決定「注意力」的焦點，以此賦予輸入或輸出的字詞給予不同的權重高低，建立長距離文本字詞之間的關聯並進行輸出，使深度學習模型能藉此強化

DOI:10.6814/NCCU202001553

學習效率，並解決傳統遞迴神經網路模型的不足之處。此一機制首先被延伸應用於機器翻譯領域之文本分析[29]上，開啟了自然語言處理相關應用之先河。Transformer 模型更進一步改良了注意力機制，提出新的「自注意力機制（Self-attention）」，讓編碼器在處理輸入及解碼器處理輸出之時，皆同時關注自己序列中的所有資料元素，以獲得更為完整的上下文資訊。

圖 2.3-3 自注意力機制圖解[30]



Google 在 2018 於 Transformer 模型的基礎上推出專門用於自然語言處理的 BERT 模型[31]，其透過雙向學習的設計以便同時有效地運用上下文的資訊，並以預訓練（Pre-training）結合精調（Fine-tuning）的方式，運用維基百科（Wikipedia）等大型公開語料庫事先訓練出具有一定程度效能之模型， BERT 已被證實在包括情緒分析在內等多個文本分類任務上有壓倒性的優秀表現，已有儼然成為自然語言處理領域新一代標準模型之姿[32]。

2019 年 6 月，Google 再度推出基於進一步改良之 Transformer 模型架構，效能比 BERT 更為強大的預訓練自然語言處理模型 XLNet[33]。研究者 Yang

DOI:10.6814/NCCU202001553

等人注意到 BERT 機制的一個缺陷，即其在預訓練階段時，會隨機對句子中 15%的詞彙標記（Token）[14]用隱藏的遮罩（Mask）來替代,等到預測時再嘗試還原成原本的字，以便藉此充分運用上下文資訊，但遮罩在精調階段時並不存在，這會造成兩個階段之間資料不對稱（Input noise）的問題。有鑑於此，XLNet 提出了稱為 Permutation language modeling (PLM) 的創新模型機制，避免在預訓練時使用遮罩，以解決了此一問題，且其亦能充分運用文本的上下文資訊，藉此讓 XLNet 在情緒分析等文字分類任務上，達到超越 BERT 的更好表現[34]。

自然語言處理之情緒分析受 Transformer 模型之賜，學習與預測能力雖有大幅提升。然而，傳統專注於處理單一屬性的二元分類模型，若要用於處理具有多個向度的嗅覺、味覺屬性進行辨識，架構似已不足適用之。

近年來在監督式學習的任務上，研究焦點多聚焦於如何讓模型從多個彼此不同的屬性中做出正確的分類判斷。此一發展對於情感計算及自然語言處理的文本情緒分析任務亦有很大的幫助，提升了相關技術應用於味覺與嗅覺感知屬性研究之可行性。本研究亦將就此做進一步之回顧探討。

[14] 在自然語言處理領域中，指對句子序列切割後，所得到具有意義的最需小詞彙單元，

DOI:10.6814/NCCU202001553

2.4 多類別分類任務

在傳統機器學習與深度學習領域中，對各種形式資料之監督式學習分類問題，大多都是透過對帶有單一標籤（Single-label）進行二元（Binary），即正向或負向標記的類別型（Categorical）依變數資料來進行訓練與分類。然而，現實生活中的資料集與分類問題往往更加複雜，針對特定實體之多個不同屬性中進行分類判斷的任務需求亦不在少數，能夠進行此一任務的多類別分類任務模型便為此而設計。無論是分辨不同種類、或區分不同程度等分類任務皆屬於多類別分類的範疇[35]；一般常見的多類型分類途徑，多為將其解構成二元分類的問題，以便直接應用二元分類模型來處理，最常見的有：一對全部（One-Against-All，OVA）以及一對一（One-Against-One，OAO）；前者為在 k 個類別中，每次針對一個類別以二元分類器進行分類判斷，將該類別與其他 k-1 個類別區分出來，逐次將所有類別分類完成，其優點在於需要的二元分類器數量與類別數目一致，缺點則是非常消耗記憶體資源，在大規模資料集的分類中可能會造成電腦硬體方面的問題，另外，在訓練集中，若各類別訓練樣本數均等，則每個類別與其他類別的訓練樣本為 1 : (k-1)，這會在訓練過程中帶來樣本數量不均衡的問題 。至於一對一，顧名思義則是每個類別每次都只跟另一個單一類別一同進行分類，在進行測試時，所有分類器會透過投票機制，根據票數的多寡來決定陌生資料的類別歸屬。此途徑解決了消耗記憶體資源及樣本數量不均的問題，但缺點則是需要產生大量的二元分類器來完成任務，如有 k 個類

DOI:10.6814/NCCU202001553

別，便需要 k (k-1) /2 個分類器來完成任務。普遍來說，一對一途徑的模型具有較一對全部為佳的分類成效，故受較為廣泛之採用。

另一方面，在生物、化學與社會科學等領域的研究中，針對數值型（Numerical）的依變數資料，運用線性迴歸（Linear regression）統計方法進行預測的方法則十分常見。近幾年來，在機器學習與深度學習領域，亦有研究嘗試將迴歸問題轉化為多類別的分類問題來解決，如學者 Salman 與 Kecman 在 2012 年的研究[36]，即嘗試將數值資料轉化為區間類別，用機器學習之分類模型取得了較傳統迴歸模型更佳的預測成效。回顧感官研究分析方法所處理之感官品評風味屬性資料，雖多以連續數值形式呈現，但究其本質實較接近問卷調查之量表評分，多為個人主觀感受之相對差異的表示，而非精確測量的絕對數值，如能參考將連續數值轉化為區間類別的研究途徑，以機器學習與深度學習之分類模型來進行學習與預測，亦，預期可有助於取得更佳之預測成效。

透過以上文獻回顧可以觀察到，多類別分類任務模型具有處理具備較多元複雜屬性資料的能力，頗有應用於分析與味覺或嗅覺相關之感知屬性資訊的可能性。本研究將進一步探索如何透過修改現行以 Transformer 模型為基礎之 XLNet 自然語言處理情緒分析模型，透過其處理多類別文本分類任務之優秀能力，有效運用具豐富感知屬性資訊的評論文本資料。藉此，本研究期盼能協助感官分析研究領域得以更便利地透過消費者研究獲得具有一定客觀參考性的研究成果外，亦同時希望能在情感計算領域開拓更多味覺及嗅覺相關應用研究的可能性。

DOI:10.6814/NCCU202001553

第三章 研究方法

3.1 研究方法架構

本研究採實驗法，透過三個部分的實驗設計，評估本研究基於感知屬性所建立深度學習品味分析模型之準確性與通用性。

首先，本研究將運用現有線上酒精飲料網站 Distiller 所提供，由專業獨立評測者以感官研究方法之針對 Whiskey 烈酒類別品項所紀錄之品評筆記（Tasting notes）文本與風味圖譜（Flavour profile）量化數值指標，，設計一以 XLNet 為基礎之自然語言處理模型，再由 Distiller 網站篩選之資料集進行多類別分類任務之深度學習訓練，接著再以另外選出之訓練品項資料對兩個模型其進行測試，再針對以此方法所建立之品味分析模型進行效能評估。

第二部分則是設計發展自 Check-All-That-Apply 感官分析實驗的 Rate-All-That-Apply 實驗，參考 Distiller 網站 Whiskey 烈酒分類對應之 14 個感官圖譜風味屬性設計問卷，再從現有專家評論資料中按便利抽樣選出六款酒飲商品，將其從訓練資料中抽出，再邀請未受訓練之一般消費者進行品味測試，首先透過閱讀該五款商品之專家品評文本，搭配酒飲商品的實際盲飲測試（Blind test），隨後判定出覺得符合的風味感知屬性數值區間（0 到 5）。最後，再將該六款商品的消費者品味測試與透過品味分析模型預測的結果進行比較，風味屬性標籤之預測值與原始專家評分資料之相似度，以評估品味分析模型在基於

DOI:10.6814/NCCU202001553

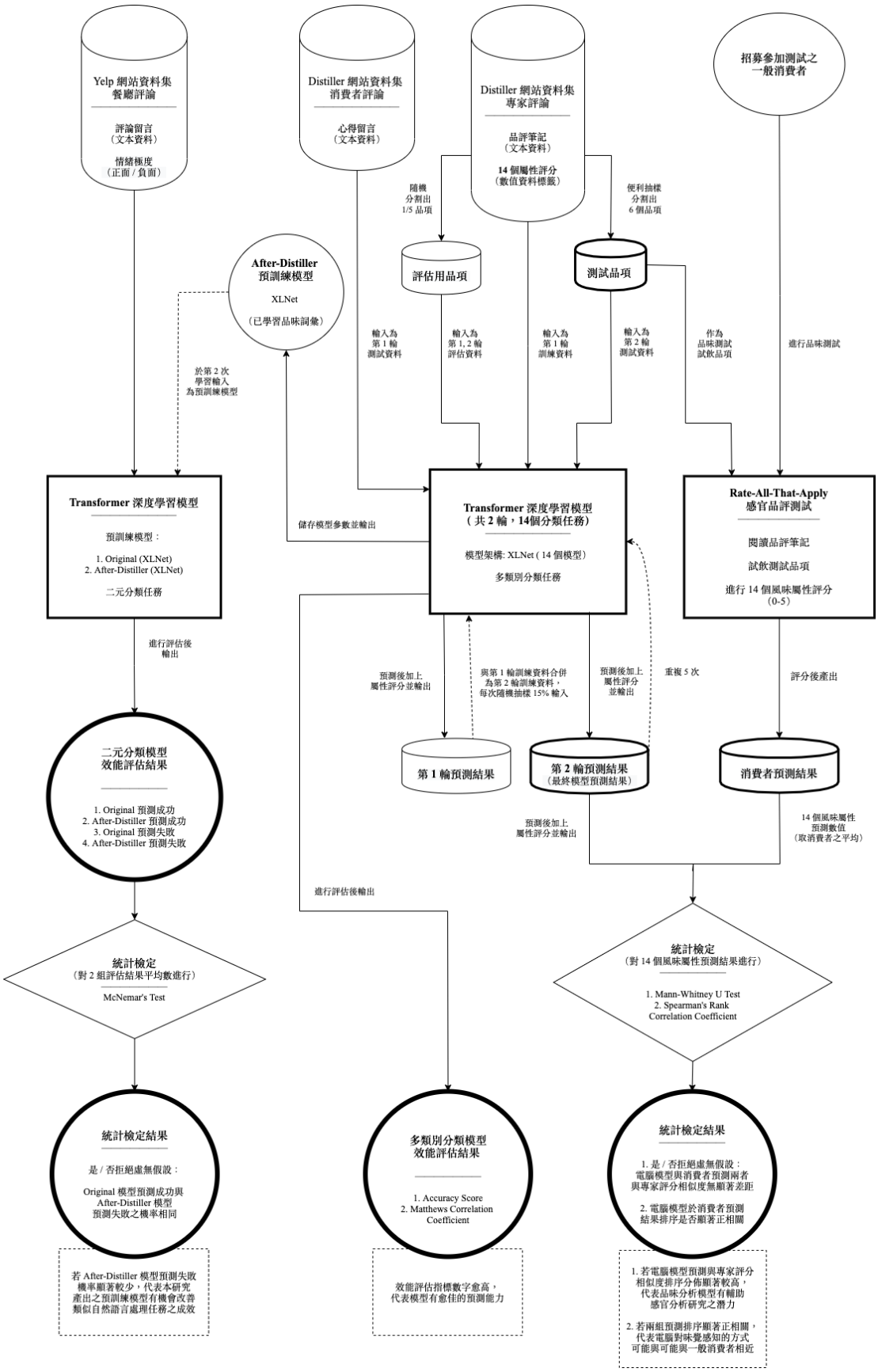
Rate-All-That-Apply 方法的感官分析測試上，其針對特定烈酒類別之風味感官屬性的辨識能力是否已與一般消費者感官評估之能力水平沒有顯著差異。

第三部分，則將針對餐廳評論網站 Yelp 公開資料集[37]進行之 XLNet 情緒分析實驗，將本研究品味分析訓練所產生之 XLNet 預訓練模型用以取代原始官方之預訓練模型予以重現。藉由比較不同預訓練模型之間預測成效差異進，來評估本研究所建立之模型是否對於其他與品味相關之自然語言處理情緒分析任務有所助益。本研究整體之研究架構整理如下圖。

國立政治大學
National Chengchi University

DOI:10.6814/NCCU202001553

圖 3.1-1：整體研究架構圖



DOI:10.6814/NCCU202001553

3.2 資料集

Distiller 為一以烈酒產品為主之酒飲資訊平台，本研究所指涉之專家評論，即為該網站上 8,266 筆由官方提供，由專業品評人員撰寫，包含產品資訊、品評筆記、風味圖譜之完整評論。專家評論主要特徵欄位整理如下表。

欄位	內容說明
品名	酒飲品項名稱
類別	酒飲品項之分類，以 Whiskey, Brandy, Vodka, Tequila, Gin, Rum, Liqueur 七大類為主
品評筆記	描述品味心得與酒款風味特性之文字
風味圖譜	14 種風味感知屬性之 0 到 100 評分

表 3.2-1 Distiller 網站 - 專家評論資料欄位

消費者評論部分，為使用該網站之消費者於網站各產品資訊頁內留言去所發表之評分與文字評論，其中有完整文字評論與評分內容者計有 151,387 筆 [38] 。消費者評論之特徵欄位整理如下表。

欄位	內容說明
品名	酒飲品項名稱
意見	描述心得與風味之文字

表 3.2-2 Distiller 網站 – 消費者評論資料欄位

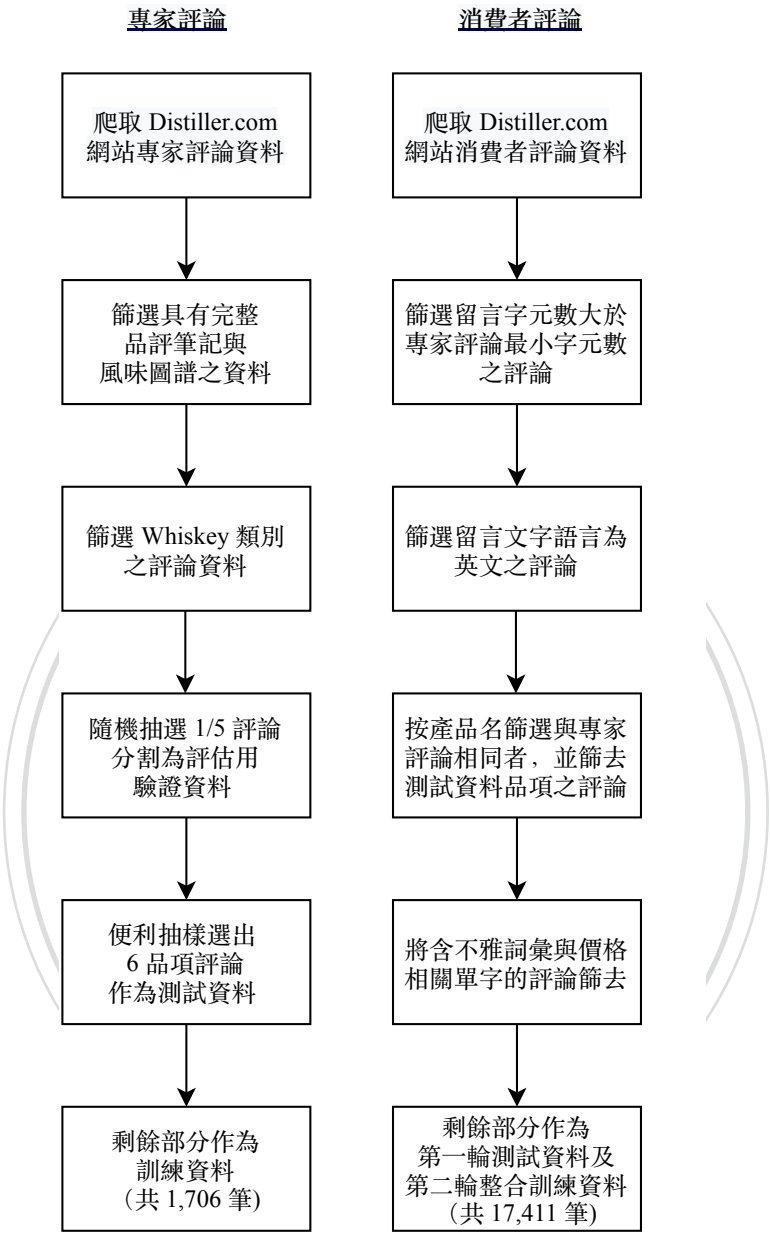
DOI:10.6814/NCCU202001553

本研究利用 Python 程式語言之 BeautifulSoup 套件撰寫爬蟲程式爬取網站本身建立之有效商品資料並結合 Pandas 套件進行資料梳洗，在篩選有含有完整提供評鑑資訊即品味筆記與風味圖譜兩者兼具之資料後，建立初步可用之訓練資料集。接著根據烈酒分類，篩選出占資料總數過半數的 Whiskey 烈酒類別品項資料，先隨機抽出 1/5 之資料作為驗證資料集[15]。再以便利抽樣法取出 6 個用於電腦模型測試與消費者品味測試所用之測試品項，將其評論從剩餘資料中移出，獨立作為測試用資料。最後剩下的 1,706 筆評論則作為專家部分的基本訓練資料，鑑於可用之專家評論資料量較小，本研究亦參考了半監督學習法中自我學習（Self-training）[39]的概念予以簡化運用，以便有效利用 Distiller 網站大量的消費者評論資料。然而，考量到消費者評論的長度、格式與文字形式、風格皆大不相同，為盡可能確保資料之有效性，對於消費者評論亦經過事先篩選之程序，處理方式如下：首先，統計出專家評論中最短評論之字元數，再從消費者評論中篩選出留言字元數大於此一數字者，接著，先將留言文字非英文的消費者評論篩去，然後再按照品項名稱選出同為 Whiskey 類別的評論。經由對消費者評論留言的觀察，發現留言內大部分提到品項價格相關訊息的評論，一般帶有較少關於品味的敘述，因此，最後再跟據本研究所整理之價格相關字詞表，並參考來自網路資源之禁用字清單[40]以篩去含不雅詞彙的評論後，將剩餘之 17,411 筆資料做為後續兩輪訓練之使用，資料處理流程示意如下圖。

[15] 驗證資料（Validation Data）為機器學習任務中用來評估（Evaluation）模型參數成效所使用的資料，通常與訓練資料及測試資料完全分隔，以確保模型的一般化（Generalization）能力。

DOI:10.6814/NCCU202001553

圖 3.2-1　Distiller 專家與消費者評論資料處理流程



最終處理後之 Distiller 資料集數量統計整理如下表。

DOI:10.6814/NCCU202001553

評論類別	訓練資料	驗證資料	測試資料	合計
專家評論	1, 706	428	6	2,140
使用者評論	17, 413	0	0	17, 413
總計	19, 119	428	6	19,553

表 3.2-4　Distiller 資料集數量統計

如前述，在經過清洗與梳理後，在 Distiller 資料集內，每一筆評論資料，都具有品評筆記或意見之文本資料，以及含有風味圖譜，即 14 個風味感知屬性介於 0 到 5 之間的數值資料。針對最終處理完成之專家評論與消費者評論文本資料長度部分，以字元數為單位，整理其敘述性統計指標於下表:

統計指標	專家評論（字元）	使用者評論（字元）
平均值	372	272
中位數	363	201
最大值	811	3, 462
最小值	119	119

表 3.2-3　篩選後之專家品評筆記與使用者評論字數長度統計

專家資料的風味圖譜部分，Whiskey 烈酒類別一共有 28 種具味覺、嗅覺或觸覺性質的感知屬性。由網站所屬之評專家評測者針對商品進行感官評價測試後，根據其感覺之強度大小給予不同屬性 0-100 範圍內的量化數值評分。本研究將 14 個標籤經參考相關風味詞彙定義[41-43]後，歸類整理如下表。

DOI:10.6814/NCCU202001553

風味感知屬性	中文翻譯	感知類別
Salty	鹹味	味覺
Sweet	甜味	味覺
Tart	酸味	味覺
Floral	花香味	嗅覺
Smoky	煙燻味	嗅覺
Full Bodied	酒體飽滿感	觸覺
Briny	海水味	味覺 / 嗅覺
Fruity	果香味	味覺 / 嗅覺
Herbal	草本味	味覺 / 嗅覺
Peaty	泥煤味	味覺 / 嗅覺
Rich	豐盈感	味覺 / 嗅覺
Spicy	辛香料味	味覺 / 嗅覺
Vanilla	香草味	味覺 / 嗅覺
Oily	油感	味覺 / 嗅覺 / 觸覺

表 3.2-1 風味圖譜感知屬性整理

Chengchi Univ

3.3 多類別分類模型設計

本研究基於 Distiller 資料集，為與後續 Rate-All-That-Apply 品味測試之問卷格式一致，將風味感知屬性之 0-100 分數，透過 Min-Max 歸一化處理[16]縮放至 0 到 5 之區間，轉換後，每個風味感知屬性各自變成具有 6 個級距類別標

[16] Min-Max 歸一化（Min-Max Normalization）為資料前處理的常見手法，透過將資料的數值按其特徵（Feature）的比例縮放至特定整數範圍（常見為 0 與 1 之間），以加強機器學習模型訓練的精度與收斂速度。

DOI:10.6814/NCCU202001553

籤的資料，在 6 個級距中，第 0 級代表數值為 0，其餘 1 到 100 之數字皆在歸一化並無條件進位後，歸類在 1 到 5 級當中。

多類別分類模型設計部分，以用於多類別分類任務[17]之現有 XLNet 自然語言處理模型為主。參考 Rajapakseyu[44]基於 Pytorch 深度學習框架與 Transformers 函式庫[45] 設計之 Simple Transformers 函式庫，以其多類別分類模型進行改寫，使其適用於 Distiller 網站資料集之多類別分類訓練與預測任務。接著，進行第一階段之自我學習訓練，先將專家評論資料輸入模型進行訓練，再讓模型對 Distiller 網站消費者評論之心得文字進行風味屬性評分預測，並將具有 14 個風味屬性的預測結果輸出為新的資料集，經與原始專家評論合併後，成為第二階段的新訓練資料集。

第二階段訓練的部分，則由最終整合專家評論與模型對網站消費者評論預測評分之新 Distiller 訓練資料集隨機抽樣出 15% 的數量，以 XLNet 架構之多類別分類情緒分析模型針對 14 個風味感知屬性分別進行 14 次訓練、評估與預測，獲得 14 組最終結果，以便協助區分出本研究之品味模型對於不同類型風味感知屬性的預測能力。待模型完成所有風味屬性的分類任務，並重複執行 5 次後，最後再將所有評估數值與預測結果整合求取平均，後續與消費者 Rate-All-That-Apply 品味測試之結果進行統計分析之比較。

在多分類模型訓練與預測任務完成後，亦將再把已學習到 Distiller 網站專家與消費者評論之品味相關詞彙意義的 XLNet 預訓練模型參數加以儲存，將其

DOI:10.6814/NCCU202001553

便轉用於 Rajapakseyu 設計之 Transformer 情緒分析模型[18]，以本研究學習 Distiller 資料集品味文本意義後所產生之 XLNet 預訓練模型（以下稱為 After-Distiller 模型）與採針對 Yelp 餐廳評論資料集進行分類任務，並與使用原始 XLNet 預訓練模型（以下成為 Original 模型）就同 Yelp 餐廳評論資料集進行分類任務的成效進行統計檢定比較，以確認 After-Distiller 模型成效相對於 Original 模型來說是否有所提升。

3.4 多類別分類模型評估標準

在 XLNet 多分類模型的部分，本研究主要採納兩個評估指標：準確度（Accuracy）以及馬修斯相關係數（Matthews correlation coefficient, MCC）。準確度的公式定義如下[46]：

$$\text{accuracy}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y}_i = y_i)$$

設 $\hat{y}_i$ 為第 i 個樣本的預測值，y_i 為其對應之真實值，則以上公式代表在 總數量為 n 的樣本中，正確預測的比例為多少，其中 $1(\hat{y}_i = y_i)$ 為一指示函數（Indicator function），表示符合 $\hat{y}_i = y_i$ 條件元素之集合。準確度的數值介於 0 到 1 之間，準確度愈高，代表模型預測成功的樣本比例亦愈高。即便準確度

[18] 此亦為 Transfer Learning 之概念；由於 XLNet 模型在訓練完後，原始預訓練模型根據訓練資料之內容，會進行內部參數之調整，故若將調整後之預訓練模型取代其他任務之原始 XLNet 模型，即可視為本研究設計品味模型的延伸應用。

DOI:10.6814/NCCU202001553

十分易於評估模型的預測能力，但其亦有「準確度悖論（Accuracy paradox）」[47]的問題存在，若樣本的類別分佈高度不均衡，例如單一類別占樣本總數的99%，此時若只參考準確度，則一個總是給出該類別預測的模型，其預測準確度亦達 0.99 ，但真實預測能力其實很差，故在類別不均衡的情況下，用準確度評估的模型預測能力往往會被高估。因此，本研究引入另一個較為平衡的指標，即馬修斯相關係數，來進行綜合評估。馬修斯相關係數一般常用於二元分類任務的評估上，其原始定義公式如下:

$$MCC = \frac{tp \times tn - fp \times fn}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}}.$$

其中，tp 為真陽性（Positive）預測數，tn 為真陰性（Negative）預測數，fp 為偽陽性預測數，而 fn 則為偽陰性之預測數；。故馬修斯相關係數可視為真預測數與偽預測數之差對於四種真偽、陰陽對偶組合的幾何平均數，其數值範圍在 -1 到 1 之間；數值若大於 0，愈大表示模型預測能力比隨機愈好；數值若小於 0，則代表模型預測愈與真實趨勢相反；數值若為 0，則代表模型預測能力基本與隨機無異。由於同時將不同類別的真實分佈納入考量，故馬修斯相關係數可有效彌補準確性作為評估指標的不足之處，有助正確評估模型的預測能力。如在前述的準確性悖論範例當中，若以馬修斯相關係數評估，其數值為 0，代表預測能力實無異於隨機。

DOI:10.6814/NCCU202001553

在多類別分類任務的情境中，馬修斯相關係數又被稱作 R_k 統計值[48]，k 代表類別的數量，其本質不變，但定義略有調整以便配合多類別分類任務的特性，多類別任務的馬修斯相關係數之定義公式如下:

$$MCC = \frac{c \times s - \sum_k^K p_k \times t_k}{\sqrt{(s^2 - \sum_k^K p_k^2) \times (s^2 - \sum_k^K t_k^2)}}$$

其中， $t_k = \sum_i^K C_{ik}$ 代表類別 k 真實出現的次數; $p_k = \sum_i^K C_{ki}$ 代表類別 k 被正確預測的次數; $c = \sum_k^K C_{kk}$ 代表正確預測的樣本總數; $s = \sum_i^K \sum_j^K C_{ij}$ 則代表全部樣本的總數。

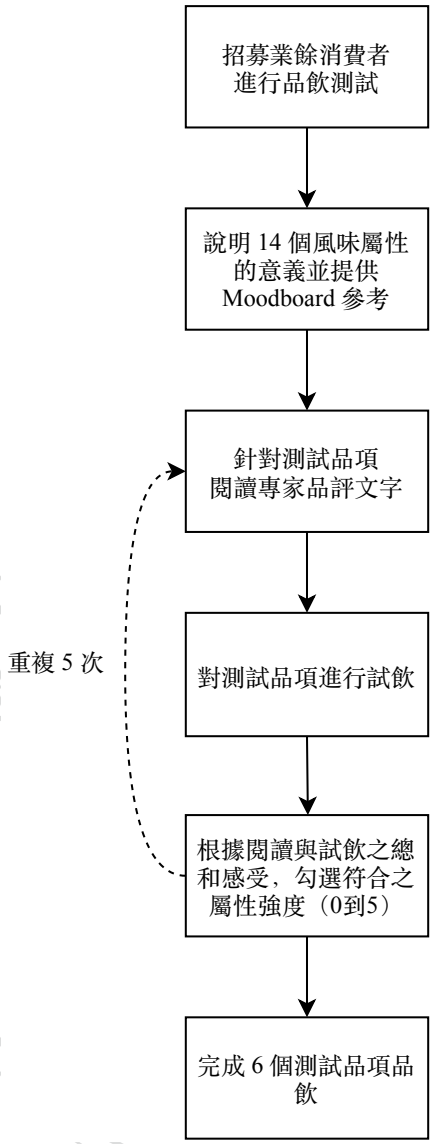
3.5 感官分析實驗設計

感官分析實驗部分，則參考由 Ares 等人改良 Check-All-That-Apply 研究法的 Rate-All-That-Apply 方法 [49] 進行設計。其執行流程如下圖。

國立政治大學 National Chengchi University

DOI:10.6814/NCCU202001553

圖 3.5-1　參考 Rate-All-That-Apply 設計之感官分析實驗流程



首先，本研究在 Distiller 資料集全部品項中先挑選出分別來自 5 個不同的 Whiskey 子類別及 5 個不同產地，以盡可能確保一定程度之風味差異性；經確認在國內有代理商進口，可循合法管道購買後，便確認其等作為電腦模型預測與消費者品味測試之研究標的。6 個測試品項之主要產品資訊列如下表。

DOI:10.6814/NCCU202001553

編號	名稱	子類別 / 產地
1	Suntory The Chita Single Grain	Single Grain / Japan
2	Jim Beam Pre-Prohibition Style Rye	Rye / Kentucky, USA
3	Bowmore 12 Year	Peated Single Malt / Islay, Scotland
4	Naked Grouse Blended Malt	Blended Malt / Scotland
5	Bulleit Bourbon	Bourbon / Kentucky, USA
6	Monkey Shoulder	Blended Malt / Speyside, Scotland

表 3.5-2 測試品項產品資訊[38]

接著，本研究透過社群網路發佈問卷調查，尋找有意願、可飲用 Whiskey 烈酒且時間許可之對象，最後抽選出 12 名未受專業訓練、其味覺與嗅覺感知能力正常的一般消費者，參加針對 6 款測試品項進行感官評估的品味測試。

參與測試之 12 名消費者，性別共計 7 男 5 女，年齡範圍介於 22 歲至 37 歲之間；品味測試之整體流程如下：首先，解釋整體測試之流程，接著針對本研究實驗之主要變項，即 Distiller 在 Whiskey 類別採用的 14 個風味感知屬性進行說明與解釋，並提供 14 個屬性之情緒板（Moodboard）[19]做為參考範例。

介紹完畢後，便開始進行正式測試。在每一個品項的測試中，皆依次完成以下步驟：首先，請參與測試之消費者閱讀來自 Distiller 網站由專家為測試品項撰寫之品飲筆記文字，除原英文文本外，亦提供本研究所撰寫之中文翻譯並列參考。接著，請消費者以鼻就杯，先行感受測試品項之香氣，接著再以口進

[19] 為透過圖像、文本拼貼而成的視覺呈現方式，通常用來描述對特定主題的整體感受，引起類似的情緒反應，為一歷史悠久並受廣泛採用的輔助研究工具。

DOI:10.6814/NCCU202001553

行正式品飲。進行品飲後，請消費者綜合文本閱讀與實際品飲之印象，就 14 個風味感知屬性分別予以 0-5 的評分。在每款品項測試的中間，提供白開水飲用，除補充水分外，亦沖淡前一測試品項留於口鼻之氣味。此一過程持續重複進行，直至所有測試品項皆評分完畢。

一般而言，如欲比較自然語言處理之電腦模型與一般消費者對評論文字的風味感知屬性預測能力，應該使電腦與消費者同樣皆以品評文字作為輸入資訊，但考量到一般消費者藉由文字體會風味，再把感受轉化為量尺分數的過程較為抽象且難以想像，故本研究之品味測試除提供風味感知屬性之說明與情緒板參考外，亦採文字閱讀加上實際試飲的方式來進行，以期稍微彌補一般消費者與專家之間風味感知屬性評分能力的差距，Ares 等人於 2017 年的研究[50]中亦指出，少量的訓練即能大幅提升一般消費者之感知評估能力，本研究亦盼能藉此一能力之提升，加強電腦模型預測作為比較對象的實用性。

本研究品飲測試之風味感知屬性評分採線上問卷形式進行，評分區間數量之選擇參考 Meyner 等人 2015 年研究[51]所歸納，於 Rate-All-That-Apply 感官分析使用之 6 點式量尺，以便更好對應代表不適用（Not applicable）之 0 數值的存在。

3.6 感官分析實驗評估

感官分析實驗結果，主要評估方法如下：首先將受試者針對 6 個研究標的評分的問卷進行統計。接著，將 6 個測試品項的 Distiller 網站專業品評筆記文本輸

DOI:10.6814/NCCU202001553

入本研究所建立之多類別分類品味模型，就其進行 14 個風味感知屬性級距之預測，隨後再將其預測結果與一般消費者評測之結果分別與 Distiller 網站之專家感官品評分數相減，求取兩組 6 個測試品項的 14 個風味屬性測試結果和專家品評結果之餘弦相似度（Cosine similarity）[52]。餘弦相似度為透過兩個向量之間夾角的餘弦數值，來估計兩者之間的相似性。在兩個向量夾角為 0 度角的狀況下，代表兩者有相同指向，此時餘弦值為 1；在兩個向量夾角為 90 度的狀況下，餘弦值為 0；在兩個向量夾角為 180 度的狀況下，代表兩者指向完全相反，此時最小值則為 -1。在其他角度，餘弦值會介於-1 到 1 之間，數值愈大代表相似度愈高，其結果與向量的長度無關，而僅與角度相關，其計算公式[53]如下，其中 x, y 分別代表進行比較的兩組列向量（Row vectors）：

$$k(x,y) = \frac{xy^{\top}}{\| x \| \| y \|}$$

餘弦相似度可被應用於任何維度的向量空間中，因此十分常見於文本挖掘任務的文件比較用途等高維度正空間中，其數值介於 0 到 1 之間。

接下來，在分別計算出電腦模型預測與一般消費者評測就測試品項的 14 個風味感知屬性與專家評測結果之餘弦相似度後，本研究將進一步運用 Python 語言之 pingouin 套件[54]，就兩組餘弦相似度進行單尾（one-tail）之曼-惠特尼 U 檢定（Mann-Whitney’s U Test）[55]，以檢驗電腦模型預測與專家品評之餘弦相似度排序分佈是否顯著高於於一般消費者評測結果與專家評分之餘弦相似

DOI:10.6814/NCCU202001553

度；基於本研究對電腦模型預測能力應不遜於一般消費者評測之假設，在此定義虛無假設（Null hypothesis，H_0）為電腦模型於六個品項之餘弦相似度排序與一般消費者品評餘弦相似度之排序彼此交錯或低於一般消費者，對立假設（Alternative hypothesis，H_1）則為電腦模型於六個品項之餘弦相似度排序高於一般消費者品評之餘弦相似度排序。

另外，除比較電腦模型預測與一般消費者評測能力的高低外，本研究亦欲探究電腦模型透過品評文字對風味屬性的感知能力和一般消費者透過試飲與品評文字的風味屬性感知能力是否有著相似的模式存在，故將再針對 6 個預測品項之 14 個風味屬性之電腦模型預測結果與一般消費者評測結果，透過複製至最大公倍數的方式對齊總樣本數量數量，再透過 Python 之 SciPy 套件[56]進行斯皮爾曼等級相關係數（Spearman's rank correlation coefficient）[57]之計算，藉以觀察兩組預測結果數值排序之分佈是否顯著相同，若確有顯著正相關的現象，則代表電腦對於品味文本的感知模式與一般消費者對品味文本與試飲樣本可能具有相近的感知模式；反之，則代表電腦對於品味文本的感知模式和一般消費者對品味文本與試飲樣本的感知模式可能有所不同。

本研究透過以上標準之檢驗，比較深度學習品評筆記之多類別分類模型與進行感官品評測試之一般消費者間的感知能力差距，若電腦模型之預測結果排序經統計檢定後確實顯著高於與消費者平均評測之結果，且電腦預測結果與一般消費者評測結果之排序有顯著之正相關，則我們可假設本研究所產出之模型作為感官分析研究替代一般消費者測試的補充方法，確實具有一定的可靠性。

DOI:10.6814/NCCU202001553

3.7 Yelp 資料集測試實驗設計與評估

參考 Rajapakse 針對 Yelp 餐廳評論公開資料集[37] 設計的情緒分析二元分類模型[44, 58]。首先以其原始設計模型，修改為使用 XLNet 構之預訓練模型，在此稱為 Original 模型，將該模型其經過微調後，再以 Yelp 的餐廳評論資料集的 560,000 筆訓練資料與 38,000 筆評估資料中作為輸入項，進行訓練與評估。接下來，再使用本研究經由 Distiller 資料集訓練，學習品味相關詞彙意義後儲存參數並輸出所產生之 XLNet 預訓練模型，在此稱為 After-Distiller 模型，以遷移學習的概念來進行同樣的情緒分析任務。

在 Yelp 資料集之二元分類模型評估部分，則採用麥內瑪檢定（McNemar's Test）[59] 作為模型分類效能的主要評估指標。麥內瑪檢定為一應用於成對二元類別型資料之 2 x 2 列聯表（Contingency Table）的統計檢驗方法，其可用來決定列聯表中各個行與列的邊際頻率是否相等，即進行邊際同質性（Marginal homogeneity）之檢定，其格式如下表。

	測試 2 ： 正結果	測試 2: 負結果	列總計
測試 1 ： 正結果	a	b	$a+b$
測試 1 ： 負結果	c	d	$c+d$
行總計	$a+c$	$b+d$	n

表 3.7-1 麥內瑪檢定列聯表格式

DOI:10.6814/NCCU202001553

麥內瑪檢定之統計檢定值公式如下:

$$\chi^2 = \frac{(b-c)^2}{b+c}$$

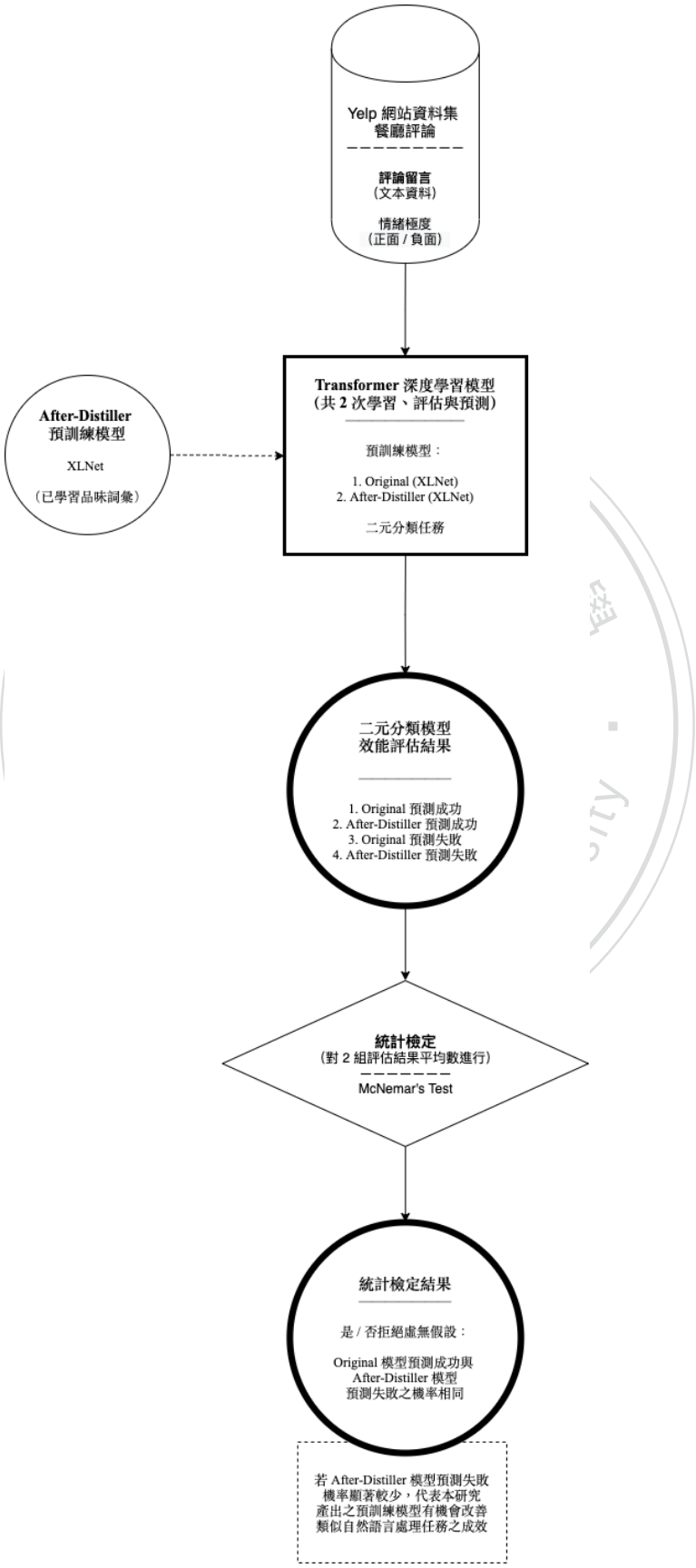
在邊際同質性呈現為兩組邊際機率相同的虛無假設下，只要列聯標內的 b 與 c 兩欄的不一致性（Discordant）達到夠大的水準，χ^2 便符合自由度為 1 的卡方分配（Chi-square distribution），若 χ^2 之結果為顯著，便有足夠的證據來拒絕虛無假設，此時對立假設成立，代表兩組邊際機率比例顯著不同。

本研究將進一步透過麥內瑪檢定來分別檢驗基於 XLNet 架構之 After-Distiller 預訓練模型預測能力是否較兩者的 Original 預訓練模型顯著為高。此檢定將針對 XLNet 的 Original 與 After-Distiller 模型預測結果產出檢定結果；其虛無假設皆為 After-Distiller 預訓練模型之預測失敗之頻率（即測試 2：負結果）等於 Original 預訓練模型預測失敗之頻率（即測試 1：負結果）；對立假設皆為 After-Distiller 預訓練模型預測失敗之頻率顯著少於 Original 預訓練模型預測失敗之頻率。

最後，將 Yelp 資料集測試實驗之整體流程整理如下圖。

DOI:10.6814/NCCU202001553

圖 3.7-1　Yelp 資料集測試流程



DOI:10.6814/NCCU202001553

第四章 研究結果

4.1 多類別分類模型評估

本研究之 XLNet 多類別分類情緒分析模型對 Distiller 資料集進行訓練所採用之預訓練模型為 xlnet-base-cased[60]，其主要模型超參數（Hyperparameter）設定如下表。

Hyperparameters	Value
Training batch size	4
Eval batch size	4
Gradient accumulation steps	4
Learning rate	2e-5
Train epoch	4
Max sequence length	256
Adam epsilon	1e-8

表 4.1-1 XLNet 多類別分類模型主要超參數設定

第一階段 XLNet 多類別分類情緒分析模型於專家品評資料，在進行 14 次自我學習訓練後，其模型評估成效如下表。

DOI:10.6814/NCCU202001553

風味感知屬性	Accuracy	MCC
Salty	0.523364	0.264225
Sweet	0.471963	0.219597
Tart	0.345794	0.107707
Floral	0.411215	0.123758
Smoky	0.509346	0.290316
Full Bodied	0.500000	0.260697
Briny	0.595794	0.290664
Fruity	0.469626	0.284198
Herbal	0.476636	0.207911
Peaty	0.771028	0.427353
Rich	0.483645	0.244250
Spicy	0.408879	0.179724
Vanilla	0.408879	0.159587
Oily	0.467290	0.214856
風味屬性平均	**0.480590**	**0.229552**

表 4.1-2　XLNet 多類別分類模型自我學習訓練成效

在第二階段訓練，經過共 6 次完整訓練過程後，XLNet 多類別分類模型之評估指標，在按測試次數取平均，且四捨五入到小數點後六位後，將結果整理如下表。

DOI:10.6814/NCCU202001553

風味感知屬性	Accuracy Score	MCC
Salty	0.478037	0.206476
Sweet	0.437383	0.165635
Tart	0.334112	0.093526
Floral	0.365888	0.100396
Smoky	0.492056	0.267740
Full Bodied	0.465888	0.211268
Briny	0.599533	0.302398
Fruity	0.428505	0.229355
Herbal	0.355140	0.075897
Peaty	0.774766	0.431429
Rich	0.459813	0.210057
Spicy	0.378505	0.129786
Vanilla	0.374299	0.108417
Oily	0.429439	0.145701
平均	**0.455240**	**0.191292**

表 4.1-3 XLNet 多類別分類模型訓練平均評估成效 (n=6)

由以上成效評估結果之整理，初步可大略觀察到幾個現象:

一、 本研究之品味模型針對以下訓練成效相對較佳：Peaty（泥煤味）、Briny（海水味）、Fruity（水果味）、Smoky（煙燻味）、 Full Bodied（酒體飽滿感）、Rich（豐盈感）、Salty（鹹味），這些風味屬性的準確度都達到 0.4 以上，馬修斯相關係數亦皆達 0.2 以上水平。其中針對

DOI:10.6814/NCCU202001553

部分 Whiskey 獨有的 Peaty 風味，模型的訓練成效水平更是特別顯著突出，準確度達 0.77 以上，且有 0.43 以上之馬修斯相關係數。

二、 除了 Peaty, Briny, Fruity, Rich 是綜合味覺與嗅覺的風味感知屬性外，成效較佳的其餘 Smoky, Full Bodied, Salty 三者分別屬於單純的嗅覺、觸覺與味覺。就以上現象之觀察，或可推論在 Whiskey 類別品項中較獨特、易辨認、且感知屬性單純的風味，或許較易被電腦模型所感知。

4.2 電腦模型預測與一般消費者評測統計結果分析

針對從 Distiller 資料集中篩選挑出的 6 個測試品項，首先整理 XLNet 多類別分類之品味分析模型之預測結果與專家品評風味圖譜分數之相似度後取絕對值，接著再統計 6 次的訓練與預測結果，將電腦訓練次數編號 Train1 到 Train6，並將測試品項編號為 Item1 到 Item6 後，列 14 個風味感知屬性之餘弦相似度結果為下表。

國立政治大學 National Chengchi University

DOI:10.6814/NCCU202001553

編號	Item1	Item2	Item3	Item4	Item5	Item6
Train1	0.917417	0.949177	0.951572	0.954864	0.913699	0.948122
Train2	0.911465	0.911007	0.955969	0.935484	0.881917	0.964890
Train3	0.881409	0.953373	0.955969	0.959381	0.911342	0.952194
Train4	0.894575	0.950542	0.943312	0.960898	0.913699	0.956098
Train5	0.931115	0.966518	0.952390	0.950947	0.858395	0.957641
Train6	0.908759	0.910038	0.963208	0.963208	0.882717	0.975610
平均	**0.907457**	**0.940109**	**0.950642**	**0.954130**	**0.893628**	**0.959092**

表 4.2-1 XLNet 多類別分類模型預測與專家品評結果餘弦相似度 (n=6)

另外，在消費者品味測試的預測結果部分，在統計共 12 位參與品味測試的消費者之平均預測結果，並將消費者編號為 C1 到 C12 後，以同前之格式將其與專家品評比較後的 14 個風味屬性的餘弦相似度列於下表：

編號	Item1	Item2	Item3	Item4	Item5	Item6
C1	0.883883	0.857907	0.746418	0.790323	0.937410	0.784672
C2	0.678971	0.863499	0.786391	0.851139	0.844283	0.907443
C3	0.753763	0.857321	0.821209	0.915417	0.861461	0.840855
C4	0.571662	0.887007	0.788160	0.806559	0.798099	0.789297
C5	0.897139	0.854779	0.728182	0.887713	0.866199	0.858581
C6	0.882442	0.589192	0.703427	0.851139	0.831068	0.730972
C7	0.662097	0.791667	0.893325	0.835399	0.883991	0.742006
C8	0.833950	0.831068	0.807428	0.840100	0.766424	0.861075
C9	0.564276	0.688606	0.882082	0.813257	0.811679	0.767146

DOI:10.6814/NCCU202001553

C10	0.645497	0.787726	0.744739	0.902230	0.860663	0.865411
C11	0.649404	0.743827	0.780939	0.749006	0.790666	0.851292
C12	0.648381	0.597319	0.788342	0.660738	0.857736	0.864480
平均	**0.722622**	**0.779160**	**0.789220**	**0.825252**	**0.842473**	**0.821936**

表 4.2-2　一般消費者評測與專家品評結果餘弦相似度 (n=12)

最後，將電腦模型預測與消費者評測各自與專家品評結果比較所計算出之兩組餘弦相似度進一步以曼-惠特尼 U 檢定做預測結果排序分佈的單尾檢驗，將其統計檢驗結果四捨五入至小數點後 6 位後，整理如下表。

U-Value	**P-Value**	H_0 ($\alpha = 0.05$)	**共同語言效果量**[20]
2516	0.000000	拒絕	0.970679

表 4.2-3　電腦模型預測與一般消費者評測之曼-惠特尼統計檢定結果

根據以上的檢定結果可見，電腦模型在學習品評文本後，於風味感知屬性的預測結果其與專家預測結果的相似度普遍來說確實較一般消費者的評測結果有較高的排序，且故本研究之多類別分類情緒分析模型確實可能具有較一般消費者為佳的風味感知屬性預測能力。此一結果相當值得未來相關後續研究做更進一步的研究探討。考量現行品味模型之效能尚不突出，且消費者的樣本規模尚屬較小，若之後能再進一步修改模型以提升其預測能力，並擴大進行品味測

[20] 即 Common Language Effect Size (CLES)，用來評估檢定結果的解釋能力。其值在曼-惠特尼檢定中的代表的意義為，從兩組樣本中各取其一進行比較後，其排序結果和檢驗方向一致之機率。

DOI:10.6814/NCCU202001553

試的一般消費者數量，可望能在後續的研究中取得更能支持此一現象的研究成果。

4.3 電腦模型預測與一般消費者評測感知模式相關性分析

最後，為了探究電腦品味分析模型與一般消費者品味測試之間的感知模式是否存在相似性，我們進一步將多類別分類電腦模型針 6 個測試品項的 6 次預測結果複製 1 次，使其樣本數與 12 名一般消費者針對測試品項的品味測試評測結果一致，即使兩者皆具有 72 個樣本。接著，再將兩組預測結果的樣本進行斯皮爾曼相關係數之統計檢定，其虛無假設為兩組樣本的排序彼此不相關；對立假設為兩組樣本的排序具有相關性。檢驗結果如下表。

國立政治大學
National Chengchi University

DOI:10.6814/NCCU202001553

風味感知屬性	相關係數	P-Value	H_0 ($\alpha = 0.05$)
Salty	0.019433	0.871289	不拒絕
Sweet	-0.009569	0.936412	不拒絕
Tart	0.060863	0.611544	不拒絕
Floral	0.138684	0.245315	不拒絕
Smoky	0.164085	0.168421	不拒絕
Full Bodied	-0.073918	0.537182	不拒絕
Briny	0.067777	0.571602	不拒絕
Fruity	**0.290490**	**0.013309**	**拒絕**
Herbal	0.013128	0.912846	不拒絕
Peaty	**0.343269**	**0.003157**	**拒絕**
Rich	0.151169	0.204954	不拒絕
Spicy	0.179589	0.131172	不拒絕
Vanilla	-0.050034	0.676400	不拒絕
Oily	0.178013	0.134645	不拒絕

表 4.3-1　電腦模型預測與一般消費者評測之斯皮爾曼相關係數檢定結果

由以上檢定結果可以發現，多類別分類情緒分析電腦模型與一般消費者的在大部分風味感知屬性的預測結果皆不具有顯著之相關性，其對於味覺與嗅覺的品味感知模式可能不盡相同，然而，比較特別的是在 Fruity 與 Peaty 兩個風味感知屬性的預測結果上，電腦模型預測結果與一般消費者的評測結果皆成功拒絕虛無假設，存在顯著的正相關性。其中 Fruity 的相關係數達 0.29，Peaty 更是達到 0.34 以上之水平，代表在這兩個風味感知屬性上，電腦模型與一般

DOI:10.6814/NCCU202001553

消費者可能具有相近的品味感知模式，而 Peaty 正好也是電腦模型預測成效最突出的一個屬性，或許代表著電腦模型透過品味文本來對此一特殊風味的感知能力與感知模式都相當接近於未受專業訓練的一般消費者，此一測試結果相當值得後續研究做進一步的探討。

4.4 Yelp 資料集測試結果

本研究修改自 Rajapakse 所設計用以進行 Yelp 餐廳評論情緒分析任務之 XLNet 二元分類情緒分析模型，其主要超參數設定如下表。

Hyperparameters	Value
Training batch size	4
Eval batch size	4
Gradient accumulation steps	1
Learning rate	4e-5
Train epoch	1
Max sequence length	128
Adam epsilon	1e-8

表 4.4-1 XLNet 二元分類模型主要超參數設定

在模型完成訓練與分類預測任務後，對使用原始 XLNet 預訓練模型，以及使用學習過 Disitller 資料集品味相關詞彙的 XLNet 預訓練模型兩者在 Yelp 餐廳評論資料集的二元分類任務之成效進行評估指標的比較列於下表。

DOI:10.6814/NCCU202001553

預訓練模型	Accuracy	MCC
Original	0.947368	0.894744
After-Distiller	0.949868	0.899740

表 4.4-2　XLNet 預訓練模型之 Original 與 After-Distiller 成效評估

接著，再將兩組 XLNet 模型針對 Yelp 餐廳資料集中抽出的 38,000 評估資料-所進行預測之全部結果統一整理為列聯表，得統整結果如下表:

	After-Distiller 成功	After-Distiller 失敗	列總計
Original 成功	35, 471	569	36, 073
Original 失敗	624	1, 336	1, 927
行總計	36, 095	1, 905	38, 000

表 4.4-3　XLNet 預訓練模型 Original 與 After-Distiller 結果列聯表

將以上列聯表之結果進行麥內瑪檢定後，其最終統計檢驗結果如下表:

χ^2	P-Value	H_0 (α = 0.05)
569	0.117919	不拒絕

表 4.4-4　列聯表麥內瑪檢定統計檢驗結果

由以上統計檢驗結果可見，使用 After-Distiller 預訓練模型，對於與品味亦有相關的 Yelp 餐廳評論情緒分析二元分類任務，和使用原始 XLNet 預訓練模型的成效相比，似乎在模型效能與預測結果上並沒有顯著的改善，由此可見，經本研究之多分類情緒分析模型訓練而學習品味相關詞彙之預訓練模型，

DOI:10.6814/NCCU202001553

如欲將其延伸應用於其他相關領域之分類任務，可能需要針對模型本身做更進一步的加強，或是運用更多與嗅覺或味覺有關的品味文本資料進行訓練。

國立政治大學
National Chengchi University

DOI:10.6814/NCCU202001553

第五章 結論

總結本研究三個部分的實驗結果，可以觀察到對於品味的感知的確具有高度主觀的特性，即便是透過專業的品評文字與風味圖譜進行訓練，電腦品味模型仍無法在每個不同風味屬性都達到良好的預測結果。然而，普遍來說，電腦模型預測結果與專家品評結果的相似度仍較一般消費者評測結果與專家品評結果之間明顯為高。對於一些較特殊的風味感知屬性，像是許多蘇格蘭威士忌獨有的 Peaty（泥煤味），電腦模型能透過學習專家品評筆記之品味文本與風味感知屬性評分而達到相對優秀的預測效能，甚至可能具有近似於人類一般消費者的感知模式，此一發現相當值得未來後續研究做進一步的探索。

至於一般消費者就品味測試的感官分析評測結果，普遍來說大致呼應 Ares 等人在 2017 年所做的相關研究結論[50]，一般消費者評測的準確度隨著個體的感知能力差異有著很大的變異性。

總和以上實驗之結果，可以認定當前電腦模型對於文字風味的品味能力，在特定領域訓練資料的輔助下，已有機會能凌駕於一般消費者之感知水平，此一成果亦可提供感官分析的研究領域作為產業應用未來方向之參考，如能透過電腦模型的輔助，替代一部分的消費者感官品評測試執行，除可在一定程度上加強研究結果的客觀性，更能節省大量的時間與金錢成本。

在資訊科學的情感設計領域，未來若能通過自然語言處理之情緒分析模型進一步理解人對於風味的文字解讀與表達能力，例如：消費者對於特殊、不討

DOI:10.6814/NCCU202001553

喜風味的發現、偵測能力較強，但對於日常頻繁接觸的風味之強度評估能力較弱，以及對於味覺的辨別力較嗅覺為佳等等命題，皆有機會透過更多的資料與實驗來驗證。透過這些頗具潛力的延伸研究，或有機會在未來開發出更接近真人感知水平的品味相關情感設計應用。

最後，在現有深度學習任務延伸應用的部分，目前透過學習 Distiller 酒飲網站資料之品味相關詞彙意義而產生的預訓練模型，目前對於現有 Yelp 之餐廳評論情緒分析任務並未帶來預期中的顯著成效提升，然而，考量到目前模型效能尚不突出，未來若能持續改進模型效能、擴充對其他品味相關資料集的學習訓練，甚至是進一步建立品味相關之專門術語字典作為輔助，預期將有機會對於加強與嗅覺、味覺相關之分類任務成效帶來一定程度的幫助。

最後，本研究衷心盼望，此次對於品味分析研究的初步嘗試，能夠發揮拋磚引玉的功效，能促成未來更多與嗅覺、味覺感知屬性相關，整合資訊科學與感官分析領域的品味分析研究，藉此對於跨領域知識科技的整合以及網際網路數位內容資料的運用形式等應用研究發展，貢獻微薄之心力。

DOI:10.6814/NCCU202001553

參考文獻

1. Vaswani, A., et al., *Attention is all you need*, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017, Curran Associates Inc.: Long Beach, California, USA. p. 6000–6010.
2. Soleymani, M., et al., *A survey of multimodal sentiment analysis.* Image and Vision Computing, 2017. **65**: p. 3-14.
3. Lehrer, K. and A. Lehrer, *The language of taste.* Inquiry, 2016. **59**(6): p. 752-765.
4. Trivedi, B.P., *Gustatory system: The finer points of taste.* 2012. **486**(7403): p. S2-S3.
5. Chiras, D.D., *Human Biology*. 2013: Jones & Bartlett Learning.
6. Piggott, J., *Alcoholic beverages: Sensory evaluation and consumer research*. 2011. 1-491.
7. Ross, C.F., *Sensory science at the human–machine interface.* Trends in Food Science & Technology, 2009. **20**(2): p. 63-72.
8. Ares, G., *Methodological challenges in sensory characterization.* Current Opinion in Food Science, 2015. **3**: p. 1-5.
9. Krantz, J., *Experiencing Sensation and Perception*. 2012: Pearson Education, Limited.
10. Stets, J.E., *Emotions and Sentiments*, in *Handbook of Social Psychology*, J. Delamater, Editor. 2006, Springer US: Boston, MA. p. 309-335.
11. Hu, X., K. Choi, and J.S. Downie, *A framework for evaluating multimodal music mood classification.* Journal of the Association for Information Science and Technology, 2017. **68**(2): p. 273-285.
12. Hu, X. and J. Downie, *When Lyrics Outperform Audio for Music Mood Classification: A Feature Analysis*. 2010. 619-624.
13. Baltrusaitis, T., C. Ahuja, and L.-P. Morency, *Multimodal Machine Learning: A Survey and Taxonomy.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019. **41**(2): p. 423-443.
14. Flanagin, A.J. and M.J. Metzger, *Trusting expert- versus user-generated ratings online: The role of information volume, valence, and consumer characteristics.* Computers in Human Behavior, 2013. **29**(4): p. 1626-1634.
15. Parikh, A.A., et al., *Comparative content analysis of professional, semi-professional, and user-generated restaurant reviews.* 2016: p. 1-15.
16. Minim, V.P.R., et al., *Optimized Descriptive Profile: A rapid methodology for sensory description.* Food Quality and Preference, 2012. **24**(1): p. 190-200.
17. Murray, J.M., C.M. Delahunty, and I.A. Baxter, *Descriptive sensory analysis: past, present and future.* Food Research International, 2001. **34**(6): p. 461-471.
18. Valentin, D., et al., *Quick and dirty but still pretty good: a review of new descriptive methods in food science.* International Journal of Food Science and Technology, 2012. **47**(8): p. 1563-1578.

DOI:10.6814/NCCU202001553

19. Granitto, P.M., et al., *Modern data mining tools in descriptive sensory analysis: A case study with a Random forest approach.* Food Quality and Preference, 2007. **18**(4): p. 681-689.
20. Tao, J. and T. Tan, *Affective Computing: A Review*. 2005, Springer Berlin Heidelberg. p. 981-995.
21. Cai, G. and B. Xia, *Convolutional Neural Networks for Multimedia Sentiment Analysis*. 2015, Springer International Publishing. p. 159-167.
22. Purwins, H., et al., *Deep Learning for Audio Signal Processing*. Vol. 13. 2019.
23. 楊子蔓, *應用深度學習架構於社群網路資料分析：以 Twitter 圖文資料為例*, in *資訊科學系*. 2018, 國立政治大學. p. 73.
24. 沈昱成, *基於社群媒體情感分析歸納產品屬性優缺點*, in *資訊工程學系*. 2016, 國立成功大學: 台南市. p. 44.
25. Zhang, L., S. Wang, and B. Liu, *Deep Learning for Sentiment Analysis : A Survey.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2018.
26. 陳禔多, *基於歌詞文本分析技術探討音樂情緒辨識之方法研究 Exploring Music Emotion Recognition via Textual Analysis on Song Lyrics*. 2017.
27. Ortigosa-Hernández, J., et al., *Approaching Sentiment Analysis by using semi-supervised learning of multi-dimensional classifiers.* Neurocomputing, 2012. **92**: p. 98-115.
28. Levy, O. and Y. Goldberg, *Neural word embedding as implicit matrix factorization*, in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. 2014, MIT Press: Montreal, Canada. p. 2177–2185.
29. Firat, O., K. Cho, and Y. Bengio. *Multi-Way, Multilingual Neural Machine Translation with a Shared Attention Mechanism*. in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2016. San Diego, California: Association for Computational Linguistics.
30. 李孟. *淺談神經機器翻譯 & 用 Transformer 與 TensorFlow 2 英翻中*. 2019 [cited 2020 June, 9]; Available from: https://leemeng.tw/neural-machine-translation-with-transformer-and-tensorflow2.html.
31. Devlin, J., et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019. Minneapolis, Minnesota: Association for Computational Linguistics.
32. 李孟. *進擊的 BERT：NLP 界的巨人之力與遷移學習*. 2019 [cited 2020 June, 9]; Available from: https://leemeng.tw/attack_on_bert_transfer_learning_in_nlp.html.
33. Yang, Z., et al., *XLNet: Generalized Autoregressive Pretraining for Language Understanding*. 2019.

DOI:10.6814/NCCU202001553

34. WenWei, K. *2019-NLP 最強模型: XLNet*. 2019 [cited 2020 June, 9]; Available from: https://medium.com/ai-academy-taiwan/2019-nlp%E6%9C%80%E5%BC%B7%E6%A8%A1%E5%9E%8B-xlnet-ac728b400de3.
35. Mehra, N.K. and S. Gupta. *Survey on Multiclass Classification Methods*. 2013.
36. Salman, R. and V. Kecman. *Regression as classification*. 2012. IEEE.
37. *The Yelp Restaurant Review*. [cited 2020 June, 9]; Available from: https://www.yelp.com/dataset/.
38. *Distiller*. [cited 2020 March, 30]; Available from: https://distiller.com/.
39. Yu, N. and S. Kubler. *Semi-supervised Learning for Opinion Detection*. 2010. IEEE.
40. *Banned Word List*. 2009 [cited 2020 July, 7]; Available from: http://www.bannedwordlist.com/.
41. Wine, W.F.a. *Describing Food*. [cited 2020 June, 9]; Available from: https://world-food-and-wine.com/describing-food.
42. Wishart, D., *The flavour of whisky*. 2009. **6**(1): p. 20-26.
43. *How to Write a Menu Describing Your Food*. 2020 Feb, 11 2020 [cited 2020 July, 7]; Available from: https://www.webstaurantstore.com/article/53/how-to-write-a-menu.html.
44. Rajapakse, T. *Simple Transformers — Multi-Class Text Classification with BERT, RoBERTa, XLNet, XLM, and DistilBERT*. 2019 [cited 2020 June, 9]; Available from: https://medium.com/swlh/simple-transformers-multi-class-text-classification-with-bert-roberta-xlnet-xlm-and-8b585000ce3a.
45. *Transformers*. 2020 [cited 2020 June, 9]; Available from: https://huggingface.co/transformers/.
46. *3.3. Metrics and scoring: quantifying the quality of predictions*. 2020 [cited 2020 2020, Aug 16]; Available from: https://scikit-learn.org/stable/modules/model_evaluation.html#label-ranking-average-precision.
47. Afonja, T. *Accuracy Paradox*. 2017 Dec, 8 [cited 2020 July, 9]; Available from: https://towardsdatascience.com/accuracy-paradox-897a69e2dd9b.
48. Gorodkin, J., *Comparing two K-category assignments by a K-category correlation coefficient.* Computational Biology and Chemistry, 2004. **28**(5): p. 367-374.
49. Ares, G., et al., *Evaluation of a rating-based variant of check-all-that-apply questions: Rate-all-that-apply (RATA).* Food Quality and Preference, 2014. **36**: p. 87-95.
50. Ares, G. and P. Varela, *Trained vs. consumer panels for analytical testing: Fueling a long lasting debate in the field.* Food Quality and Preference, 2017. **61**: p. 79-86.
51. Meyners, M., S. Jaeger, and G. Ares, *On the analysis of Rate-All-That-Apply (RATA) data.* Food Quality and Preference, 2015. **49**.
52. *Cosine similarity*. 2020 2020, Aug 10 [cited 2020 2020, Aug 14]; Available from: https://en.wikipedia.org/wiki/Cosine_similarity.

DOI:10.6814/NCCU202001553

53. *6.8. Pairwise metrics, Affinities and Kernels*. 2020 [cited 2020 2020, Aug 16]; Available from: https://scikit-learn.org/stable/modules/metrics.html#cosine-similarity.
54. *pingouin.mwu*. 2020 [cited 2020 July, 12]; Available from: https://pingouin-stats.org/generated/pingouin.mwu.html.
55. *Mann–Whitney U test*. 2020 [cited 2020 July, 8]; Available from: https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test.
56. *scipy.stats.spearmanr*. 2020 2020, July 23 [cited 2020 2020, Aug 15]; Available from: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html.
57. *Spearman's rank correlation coefficient*. 2020 2020, July 10 [cited 2020 2020, Aug 15]; Available from: https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient.
58. Rajapakse, T. *Simple Transformers — Introducing The Easiest Way To Use BERT, RoBERTa, XLNet, and XLM*. 2019 [cited 2020 June, 9]; Available from: https://towardsdatascience.com/simple-transformers-introducing-the-easiest-bert-roberta-xlnet-and-xlm-library-58bf8c59b2a3.
59. *McNemar's test*. 2020 2020, June 12 [cited 2020 2020, Aug 14]; Available from: https://en.wikipedia.org/wiki/McNemar%27s_test.
60. *Model: xlnet-base-cased*. 2019 [cited 2020 2020, Aug 17]; Available from: https://huggingface.co/xlnet-base-cased.

國立政治大學 National Chengchi University

DOI:10.6814/NCCU202001553