臺東大學綠色科學學刊　第九卷第二期

2019 年，9（2），27 - 56

應用文字探勘於實用推薦文辨別之研究-以愛評網美食評論為例

黃怡蓁、林聖翔、楊建民、洪為璽

摘要

網路是世界上最有用的資訊查詢工具，隨著電子商務網站大幅度興起，消費者常於進行購買前閱讀網路相關產品與店家介紹文章，並於消費後上網進行經驗回饋分享，在這樣的相互作用之下，網路上的相關產品用戶生成內容越唾手可得，資訊與雜訊的分辨逐漸重要。 相關產品文章有效影響個人和組織的購買與產品走向預測之決策，本研究提出一監督式學習的迭代模型框架，探討非結構性之推薦文文章用詞對於潛在消費者是否實用之差別，以達到辨別評論為實用或非實用文的目的；採用 ipeen 愛評網之美食評論發表時間於 2008 年 1 月至 2018 年 12 月 內，共 1219 篇實用文與 478 篇非實用文作為檢測實驗資料，透過使用者與評論層級之雙層過濾，以 Support Vector Machine、Naive Bayes classifier、Random Forests 進行分類，藉由分析結果建立預測模型，並定期擴增詞庫以自適應地學習新實用文迭代模式，因應時代用詞變化，最佳模型之 準確度為 74.62%，精確度為 0.759，召回率為 0.773，F-score 則可達 0.754，後續可進一步拓展跨領域評論辨別。

關鍵詞：中文斷詞、資料探勘、網路爬蟲、 群集分析、用戶生成內容

黃怡蓁，國立政治大學資訊管理研究所。E-mail: 106356027@nccu.edu.tw

林聖翔，國立政治大學資訊管理研究所。E-mail: 106356022@nccu.edu.tw

楊建民，國立政治大學資訊管理研究所。E-mail: jmyang@nccu.edu.tw

洪為璽，國立政治大學資訊管理研究所。E-mail: fhung@nccu.edu.tw

Green Science & Technology Journal

2019 年，9（2），27 - 56

The Influence of Meal "Speaking of Things" on Consumers' Willingness to Buy——Discussion on LIS Viewpoint

Huang,I-Chen & Lin,Sheng-Hsiang & Jiann-Min Yang & Frank Hung

Abstract

The Internet is the most useful information query tool in the world. With the rise of e-commerce websites, consumers often read online related products and store introduction articles before purchasing, and share experience feedback after sharing. Under the interaction, the more user-generated content on related products on the Internet, the more important it is to distinguish between information and noise. Related product articles effectively influence the decision of individuals and organizations to purchase and predict product trends. This study proposes an iterative model framework for supervised learning to explore the differences between non-structural recommendations and usefulness for potential consumers. The purpose of discerning comments is practical or non-practical; the use of ipeen's review of food reviews was published from January 2008 to December 2018, with a total of 1219 practical articles and 478 non-utility texts as test data. User and comment level double-layer filtering, classified by Support Vector Machine, Naive Bayes classifier, Random Forests, build prediction model by analyzing results, and regularly augment the thesaurus to adaptively learn the new practical iteration pattern, in response to The word change in the era, the accuracy of the best model is 74.62%, the accuracy is 0.759, the recall rate is 0.773, and the F-score is 0.754. The follow-up can further expand the cross-domain commentary.

Keywords: Chinese word-breaking, data mining, web crawling, cluster analysis, user-generated content

Huang,I-Chen, Department of Management Information Systems, National Chengchi University. E-mail: 106356027@nccu.edu.tw

Lin,Sheng-Hsiang, Department of Management Information Systems, National Chengchi University. E-mail: 106356022@nccu.edu.tw

Jiann-Min Yang, Department of Management Information Systems, National Chengchi University. E-mail: jmyang@nccu.edu.tw

Frank Hung, Department of Management Information Systems, National Chengchi University. E-mail: fhung@nccu.edu.tw

壹、緒論

(一)、 研究動機與背景

隨著互聯網的發達與進步，使得消費者在網路上的互動漸趨頻繁，以用戶生成內容為主軸的評論網站日漸普及，用戶生成內容包含許多有關產品和服務之有價值的重要信息，消費者將藉由網路及報章雜誌尋求他人對特定產品的使用經驗、相關分享以降低購買風險和不確定，並主動將自身消費經驗回饋企業、分享給群眾；透過上述行為內化成其購買或前往店面時的參考依據。此時，線上評論的快速傳播特性與實用性與否即成為影響消費者購買決策的重要標的。

當使用者不再只是被動的接收訊息，而是成為主動的訊息傳播者。網路評論越趨繁多，企業決策者或是消費者能夠很容易地從網際網路上獲得它們所需要的資訊，其品質卻難以判斷與處理，在資訊超載（Informations Overload）的情況下，如何有效過濾雜訊的重要性便油然而生。

相較於過去消費者只能從電視與廣告中得到正向推播，或是透過親朋好友推薦以得到產品或商家的正反面評價，根據 Pang et al.於 2008 對 2000 多名美國成年人的兩項調查顯示，有 81%的網路使用者表示在購買產品或前往實體商家前至少會上網閱讀相關評論一次，73%至 87%的人表示相關評論對於其購買決策有顯著的影響，而有 32%會通過評分系統對產品，服務或發文者進行評分，研究結果顯示人們傾向於網際網路上得到專業評論家之使用評論或經驗分享回饋，再歸納為自身的購買參考。

資訊超載將造成使用者的認知負擔，消費者在閱讀時，須從大量網頁中逐一點選連結，再閱讀整篇內容以得到想要的店家或商品評價，文章中除了文字可能還會穿插圖片的非結構化型態，導致消費者在資訊的搜尋、整理與閱讀上會耗費許多時間。

綜上所述，消費者會依照相關網路評論發布者對商品或店家的推薦與否決定其購買決策，而網路資源雜亂無章，如何有效在資料量龐大的數據海裡，迅速並準確地過濾雜訊，找到對自身有用且具有一定可信任度的資料是一大關卡；以上問題可概括分為兩層，第一層為評論是否為使用者所需，消費者可經由自訂篩選條件，例如：地區、餐廳種類和價位等等，進行評論搜尋以得所需資訊，第二層則為評論內容對於消費者是否實用，經由消費者自訂之篩選條件搜尋後之相關評論，其評論內容尚需被判定是否對於消費者為實用資訊，即本研究中所提之實用評論或非實用評論，內容被需要度越高其為實用文的機率越高，反之則越低。

為了達到上述實用資訊之辨別，資料探勘的實踐就變得尤其重要，其為近年網際網路興起、數據大量產生後之新興技術，從未知的大量資料中，將數據有目的地進行蒐集、加工並分析，進而挖掘出隱含且有用的潛在資訊模式，最終提煉出有價值的結果，以預測未來走向的過程，主要針對結構化資料進行資料分析；而本研究中之實用資訊多以半結構化或非結構化之形式存在，使用者生成內容需以其延伸技術－文字探勘進行分類分析，觀察網頁結構，利用自製或現成之爬蟲應用程式捕捉文本，將非結構化資訊轉為結構化，進而計算字詞特徵值與隱藏之字詞關聯性以得到潛在資訊。

以用戶者生成內容為主之評論網站為數眾多，本研究選用之評論網站為愛評網，相較於其他評論網站不同之處在於，其為文章閱讀者提供一分類機制，依照發布文章之使用者點讚數量與觀看數量之比例，進行實用文與非實用文的分類，已經初步為平台使用者進行實用文推薦篩選，但其卻無法完全避免人為因素干擾推薦系統計算評分。

歸納主要原因有二，其一為使用者互刷好評，使用者可能為刷高帳號之實用率而與其他使用者發起互相按讚與互刷文章好評的行為，導致實用文分類不準確。其二為太新文章尚未被多數消費者閱讀，造成讚數過少無法被正確分類。

非實用評論因上述原因而錯誤分類為實用文的增加，將對用戶體驗以及用戶行為分析及推薦等許多面向皆帶來極大的影響；因此，該以何種改善機制歸納整理分散於網路上之各種評論，以正確識別實用文與否，是一個值得關注的議題。

(二)、 研究目的

為了解決上述的問題，我們希望以意見探勘技術進行自動化美食推薦文為實用或非實用文之分類，透過文字探勘與自然語言處理的技術去對文字內容作處理，進而達到辨別評論實用或非實用文的目的，儲存大量的非結構化評論資料與各篇文章是否為實用文的標記，透過中文自動分詞將各篇文章依序斷詞並透過詞頻-逆向文件頻率(Term Frequency /Inverse Document Frequency，TF-IDF)萃取關鍵字詞，進而分析會被消費者視為實用文之文章的用詞遣字特徵，得出實用文詞庫並進行分類。

在文章撰寫者發文的同時便預測文章分類，使發布者可進而更改內文以更貼近消費者的真正需求，使閱讀文章者能不被人為因素，例如：互刷好評，所造成的錯誤分類誤導，快速得到目標產品或店家的真實評價以及所需資訊。

歸納本研究之目的為：

- 運用中文斷詞及正規化等資料前處理，進行非結構性文章內容之用詞相關分析，

改善並應用 TF-IDF 演算法萃取實用評論文章及非實用評論文章之個別（綜合）關鍵字詞，得到訓練詞庫。

- 探討以主題性分析為主的消費者行為之應用，包含文章語義關聯之主題模型分類應用與單純以關鍵字詞頻為基礎的文章分類效能變化，以達到驗證先分群後分類能提升分類準確度之目的。
- 使用關鍵字詞庫建構一自動化偵測實用文之模型，以支持向量機(Support Vector Machine，SVM)、單純貝式分類器(Naive Bayes classifier，NB)與隨機森林樹（Random Forests，RF）三種分類分析演算法進行實用文分類，改善平台舊有利用點讚數區別之方式，並定期擴增關鍵字詞庫內容以分析評論隨著時間變化所指涉的關鍵字變化。
- 利用分類分析之結果建立一預測模型，預測尚未分類之評論文章標籤，以達到在文章發布的同時即可擁有分類標籤，使文章發布者可即時改善文章內文，並減少因人為因素干擾而導致分類錯誤的誤導行為。

(三)、 論文架構

在本文中，我們提出了一個監督式學習的方法來辨別實用評論的檢測模型，本文的其餘部份安排如下：我們在第二章中列出了一些與中文評論之遣詞用字相關分析技術的相關文獻，第三章則為資料前處理與詳細資料的收集設計與分類方法，於第四章介紹並討論實驗結果，第五章則是未來可探討之方向與結論，最後，第六章為參考文獻。

貳、文獻探討

(一)、 網路評論之於購買決策之研究

網路評論的口碑行銷對於消費者購買決策之影響的研究其來有自，相較於傳統行銷工具，例如電視廣告推播等，更容易加入自身經歷，真實反映產品性能或商家優劣，參與討論的消費者可以獲得更多具體的產品信息，以及產品或商家的正反面案例，增加產品購買量或前往商家之驅動力，以下分為兩點討論。

第一點為評論發布來源，Bickart(2001)指出網際網路使用者所生成之評論內容比營銷人員生成內容具有更高可信度，Eagly, Wood, & Chaiken(1978)提出觀眾對來源意圖的歸因是可信度感知的關鍵因素，抒發個人觀點與自身經驗的網路評論撰寫者較營銷人員所提供的消息更值得信賴，因為其撰寫者與該產品或商家並無利益關係，將減少撰寫不實內容與誇大其辭的意圖，而消費者也能更加感同身受，顯著增加行銷接受

度。

第二點則為負面評論以及品牌效應，李啟誠、李羽喬(2010)提出一探討網路負面口碑之於消費者之負面效應，針對網路上之負面口碑與消費者購買決策進行假設，得出若消費者對於該產品之基本知識認知不足將受到網路負面口碑顯著影響，導致潛在消費者流失；而江義平、溫演福(2012)研究特定產品與線上評論關係時，則得出網路口碑是累積眾人評價之後所構成的品牌效應，可經由評論文章幫助企業主了解顧客對產品的了解與喜好程度，進而改善未來之服務內容與產品製造方向。

綜上所述，在線評論、相關部落格文章和新聞等皆能顯著影響消費者的購買行為，因此了解消費者對文章閱讀的喜好，依據評論撰寫者對於產品或商家的遣詞用句特徵，進一步分析文章之情緒導向，以得知消費者的推薦與否，找出需改善或建議發展方向，為推動網路口碑極重要的一環。

(二)、 文字探勘於資訊擷取與評論分析之相關研究

1. 資訊擷取之技術探討

如何在大量資訊中提取隱含未知且有價值資訊之方法即為資料探勘，常見的資料類型可分為結構化、半結構化和非結構化三種，資料探勘為針對結構化資料進行資料分析，文字探勘則為其延伸技術，兩者有緊密的關聯，現今多數文件多為文字結合圖片或影片等半結構化或非結構化資料類型，文字內容隨趨勢發展，將形成許多不同的說法與特殊常用字，因此如何從非結構化的文字中，找到隱藏的規則與關聯結構，萃取並分析出有用的重要資訊或知識，進一步達到文字探勘的目標，為本研究之重點之一，目前以中央研究院的中文斷詞系統 CKIP（Chinese Knowledge Information Processing），以及 Python Based 的開源中文斷詞程式庫－jieba(結巴)為主要自動分詞之工具。

資料探勘的技術日益重要，並且廣泛的應用在各種領域上，在過去的文獻中可以顯示出網路評論的運用已經發展了許多面向。Garg et al.(2011)提到評論探勘之熱門演算法技術應用大略可分為三類，一為關聯規則(Association Rule)，在大量資料中發現隱藏之特定模式，找到資料間彼此相依的關聯性，例如著名的 WAL-MART 尿布與啤酒之購物籃分析，二為分類分析(Classification Analysis)，利用已知資料之分類標籤屬性建立預測模型，根據已知特徵預測未知數據之分類標籤，三則為群集分析(Cluster Analysis)，與分類分析相似，不同之處在於其在運算前數據特徵為未知，將類似的族群聚在一起，並在分群之後去解讀分群意義以得到隱含的組內相似性。

本研究將關聯規則視為消費者相關行為分析，主要探討文本集的主題性相關分析，後兩者則為網路評論類型分析，例如：虛假評論辨別、找出特定文章特徵以分類、預測相關議題走勢。以下將分別討論。資料探勘的技術日益重要，並且廣泛的應用在各種領域上，在過去的文獻中可以顯示出網路評論的運用已經發展了許多面向。

Garg et al.於 2011 提到評論探勘之熱門演算法技術應用大略可分為三類，一為關聯規則 (Association rule)，在大量資料中發現隱藏之特定模式，找到資料間彼此相依的關聯性，例如著名的 WAL-MART 尿布與啤酒之購物籃分析，二為分類分析 (Classification analysis)，利用已知資料之分類標籤屬性建立預測模型，根據已知特徵預測未知數據之分類標籤，三則為群集分析(Cluster analysis)，與分類分析相似，不同之處在於其在運算前數據特徵為未知，將類似的族群群聚在一起，並在分類後解讀分類意義以得到隱含的組內相似性。

本研究將關聯規則視為消費者相關行為分析，例如：情緒分析、口碑分析、相關產品推薦，後兩者則為網路評論類型分析，例如：虛假評論辨別、找出特定文章特徵以分類、預測相關議題走勢。以下將分別討論。

2. 消費者行為之主題性分析

以消費者情緒分析為主之應用舉不勝舉，依照過去的文獻回顧可大略推得情緒分析之處理過程為文本資料蒐集，尋找潛在主題與關鍵字，推敲評論撰寫者關注之重點，進行情緒分析以得知各篇評論之正負向情緒，最終達到減少文章閱讀、整理時間的目的，本章節主要列舉在巨量資料中，探勘資料間之相互關係與規則，找出民眾關注之潛在議題，並分析未來可能發展方向之文獻。

林名彥(2015)蒐集台大 PTT e-shopping 論壇上的客訴文章作為資料集，尋找並瞭解網友們經常抱怨的主題和關聯的字詞，歸納整理出 10 個主要客訴原因，進而提供商品管理和服務上的參考，以即時處理顧客的抱怨。

吳珮菁(2012)發表研究，認為舊有運用協同過濾來進行個人化的推薦有盲點，

因為購買過不一定是喜歡，因此須從消費者的回饋評價中，分析情緒用詞以了解消費者購買後的真實情緒感受，利用主題分析找出顧客重視之產品主題，得知該消費者對於該產品的滿意程度，進而改善並有效提升產品推薦的成效。

林國仲(2017)應用網路爬蟲於 PTT 汽車版蒐集文章，透過主題分析建立特徵

詞庫，利用辭典法進行情緒導向分析，標籤出各篇文章是正向或負向，減少消費者在逐一點擊文章，並完整閱讀文章後才能得知文章對產品的正負評價所花費的時

間，以提升效率。

綜上所述，意見探勘在消費者行為分析扮演著重要的角色，以擷取關鍵字特徵的方法，將完整的產品相關評論結構化，找出文本集中的隱性關聯性進行相關主題尋找或依照辭典法進行情緒分析，以得知文章之情緒導向，也可進一步執行文章類型的分類分析。

3. 網路評論類型分析

機器學習技術在意見探勘分類中越來越流行，評論類型的分類檢測方法可以分為三種型式分別為：監督式學習（需要標記所有數據集）、半監督式學習（標記少量數據集）、非監督式學習（無須標記任何數據集）。

監督式學習與非監督式學習各有優劣，監督式學習須於事先訓練模型以掌握目標資料特徵，使分類辨識結果之效能提高，但資料變動時將須重新學習資料特徵，較耗費時間。非監督式學習則不需事先訓練，但機器在學習時並不知道其分類結果是否正確，其訓練結果可能跟希望的結果完全不相干。(劉力華(2010),陳世榮(2015))。

任柏衛(2015)提出一監督式學習之美食推薦系統，利用文章分析找到評價好的餐廳推薦給使用者，爬取 PTT 食物版已標記為正向或負向之評論，以 TF-IDF 計算字詞頻率，並生成高頻率字詞詞庫，再以支撐向量機(Support Vector Machine, SVM)進行模型訓練與分類預測，藉由分類結果作為評分依據。

在非監督式學習的領域裡，由 Chen et al.(2017)提出一在 Twitter 上搜索發布惡意縮短網址與重複發文之機器人辨別，依照四步驟依序組成，以 crawler 爬取關鍵字，duplicate filter 過濾重複發送相同關鍵字之使用者帳戶，collector 收集有疑慮帳戶之歷史發文，最後由 bot detector 進行辨別，並與當時在 Twitter 上做 bot 檢測的兩種方法比較，一為監督式方法－BotOrNot(2016)，二為非監督式方法－DeBot (2016)，並證明自身比其二者皆準確，此方法目前局限於關鍵字為縮短之惡意網址的爬取，並無法直接進行無標記數據集的虛偽判別，是值得關注且有可發展空間的一環。

Tsur et al. (2010)與 Narayan et al.(2018)分別提出辨監督式學習之模型，前者以自定義之 Semi-supervised Algorithm for arcasm(SASI)演算法，蒐集句型特徵，並進行相似句法分群，人工處理小量標記的負向評論，將句子中含有[公司行號]/[產品名稱]/[特定型號]等之特徵，以[company], [product] and [number]等字詞取代，再以此 Pattern-based 進行自動提取相似句型模式，後者則使用 PU-learning 演算法，從非常少的標記數據集和大量未標記的數據集中去檢測並辨別虛偽評論，以 Ott et al.(2011)

創建之數據集作為標記數據集，從而加入大量未標記數據進行迭代訓練並分類，並比較 Decision tree、Naive Bayes、SVM、 k-NN、Random forest 和 Logistic regression 之分類準確性。皆可改善上述監督式學習之資料變動與非監督式學習之準確度衡量的問題，達到不須重新學習特徵且能定義準確度之目的。

Sedhai et al.(2018)發表一迭代更新資料集之監督式學習研究，將重點關注於 Twitter 上，提出 Semi-Supervised Spam Detection (S3D) 框架，以人工小量標記惡意發文並批次更新資料集的方法檢測 weet 是否為惡意發文，以達到動態適應並學習惡意發文之新寫作模式以提高準確性。

綜上所述，隨著網路評論撰寫者的寫作模式不斷更新，利用批次更新標記群集或半監督式方法去監督並抓取變更之評論特徵與句型結構模式，可以就小量資料執行特徵萃取後，預測並描述巨量資料的意義，極具應用潛力(陳世榮(2015))。

本研究使用批次更新資料集之監督式學習的方法，在訓練與測試資料皆有標記的狀況下建立模型與其對應關係，得到具有預測能力的資料輸出，並以批次增加訓練集的方式，觀察資料集增加之分類準確度的變化，得到訓練集的資料多寡 是否影響分類效能之結果，達到最完善之分類效果。

參、研究方法

本研究的分析對象是依據愛評網(ipeen)的美食頻道評論文章為主，其平台依據文章發布後之觀看者的按讚及好評數量，決定此篇文章是否可被歸類為實用文。

而本研究旨在找出實用文與非實用文之間的文章構句特徵差異，以達到發布同時便可自動化判定此未標記文章是否為實用文的目的，減少人為因素干擾而導致的分類錯誤，利用批次更新資料集的方式，提高關鍵字詞典之準確度，並以分類結果建立一預測模型。

(一)、 研究流程

本研究流程可分為兩個主要模塊，如圖 1 所示，第一部份為即時評論分類，從一小部份已標記樣本集開始，以 python 爬取愛評網(ipeen)美食頻道排名前 15 間推薦餐廳 (2008-2018) 的評論文章、評論發布來源以及其是否為實用文之標記；利用使用者層級過濾與評論層級分類的雙層過濾模式進行評論分類分辨，共分為正則表達式、中文斷詞、主題分群、閥值設定與特徵選取等五個步驟。依序處理非結構化文章，文本文件經前置處理後進行主題群集分析、特徵提取，得到以實用文為主體之關鍵字詞庫，再以關鍵字詞庫訓練模型進行實用文與非實用文之評論分類，檢驗詞庫優劣。

第二部份則為定時樣本集更新，批次進行愛評網(ipeen)之美食評論文章爬取，定期更新文本資料集以建立關鍵字詞庫，利用新的資料群集更新檢測模型，因應使用者生成內容隨時間產生的用字遣詞變化，增加準確度。以下將進行詳細介紹。

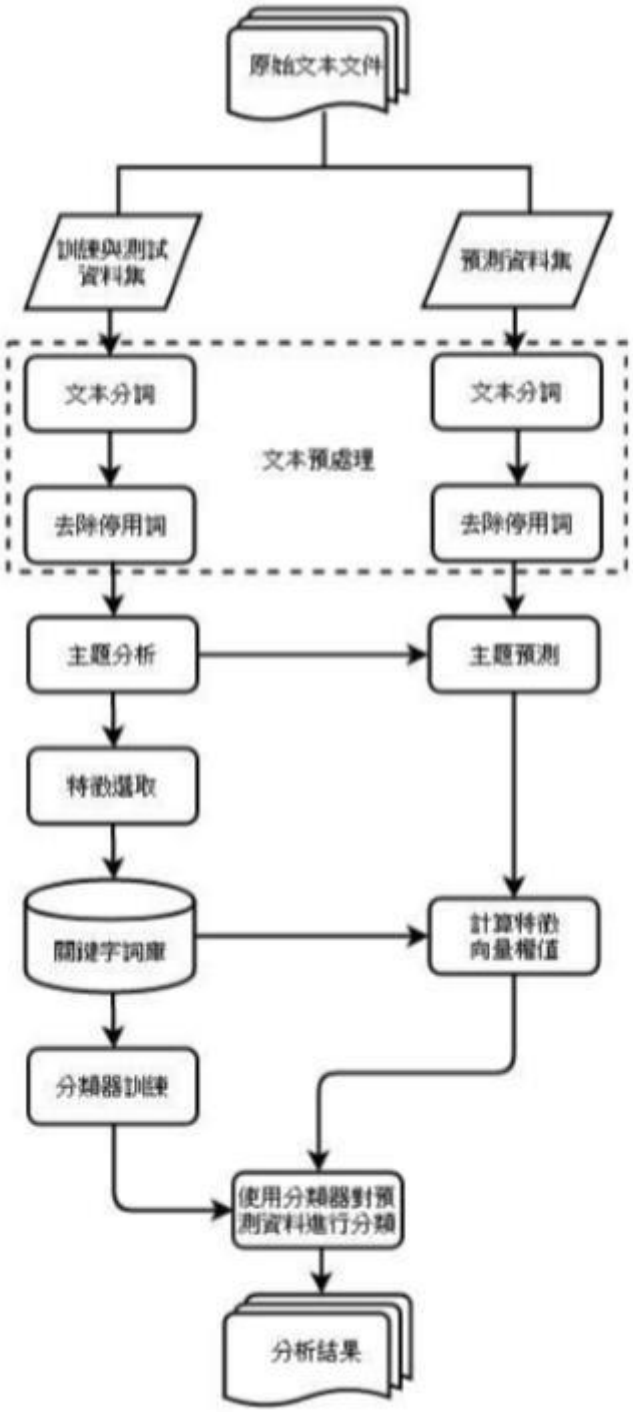

圖 1、本研究流程。

資料來源：本研究。

(二)、 資料前置處理

本研究使用 ipeen 愛評網(ipeen)的美食頻道排名前 15 間推薦餐廳 2008-2018 年之評論文章為分析對象，資料前處理流程可分為三個步驟，說明如下：

1. 數據集描述

台灣能瀏覽有關美食評論文章的線上論壇為數眾多，例如：痞客幫(pixnet)、愛食記(ifoodie)等，但上述網站尚無鑑定系統以界定此篇文章是否對消費者有實際的用處，因此本研究鎖定愛評網(ipeen)的美食頻道評論，其針對每一篇文章皆有提供實用文與否之檢測，實用文與否之定義為評論內容對消費者越有實際幫助者其為實用文的機率越高，反之則越低。

本研究以 python 撰寫爬蟲程式，收集愛評網(ipeen)的美食頻道之評論本文、實用文與否之標記以及發布者帳號之實用率，發表時間為 2008 年 1 月至 2018 年 12 月內，共 1,219 篇實用文評論與 478 篇非實用文評論作為資料來源。

表 1、美食評論文章數量整理表。

	美食評論文章數量整理				
年份	2008-2012	2012-2015	2016-2018	總計	年份
文章數量	實用文	470	420	329	1219
	非實用文	142	147	189	478

資料來源：本研究整理。

2. 正則表達式

原始文本文件內包含半形與全形標點符號、數字與英文字母，將影響後續之關鍵字詞頻計算，於此階段使用正則表達式去除。

3. 中文斷詞

自動分詞為自然語言處理之基底技術，字詞為文本之最小且具意義的單位，在進行相關語言處理、文本分析或計算字詞頻率等研究前，須將原始文本內容結構化，分辨文本中的字詞，轉換成特徵向量，方能進行分類程序。

而中文的原始文本呈現方式不同於歐美語系，歐美語系之英文單字可用空格或符號進行分隔，而中文字詞的切分基本上是依照辭典所收錄之字詞進行比對，字詞與字詞之間並沒有空格或符號將其間隔開，因此會有產生歧義詞的問題，處理上將較英文文本複雜。

中文自動分詞目前有兩個套件可以使用，其一為中央研究院的中文斷詞系統

CKIP（Chinese Knowledge Information Processing），其二為 Python Based 的開源中文斷詞程式庫－jieba(結巴)，兩者皆可協助使用者進行中文斷詞與詞性標註等功能，不同之處則在於前者 CKIP 是非開源軟體，使用者因而無法自行調整所需功能，且其伺服器並不穩定等較不利於研究等因素，本研究採用基於 python 之 jieba(結巴)套件進行中文自動分詞的處理。

jieba(結巴)支持三種分詞模式，分別為默認精確模式(cut)、全模式(cut_all)與搜尋引擎模式(cut_for_search)，而由王力弘(2015)之研究提出自動分詞的分詞成效將影響模型之分類準確度，而在文本分析中注重的是關鍵字的品質而非數量，過多的字詞將導致效率降低以及雜訊過多的現象，得出以精準模式進行斷詞後，其結果最適合文本分析之結論。

因此，本研究亦採用精確模式進行自動分詞，分詞示意圖如圖 2 所示，D 表資料集內評論文件數，T 則表各篇文章斷詞後之相異字詞數，分別對實用評論文章集及非實用評論文章集進行 jieba(結巴)斷詞後，再依各文件取其相異的字詞聯集，所得到之字詞將形成初步詞庫用於下一階段之字詞頻率計算。

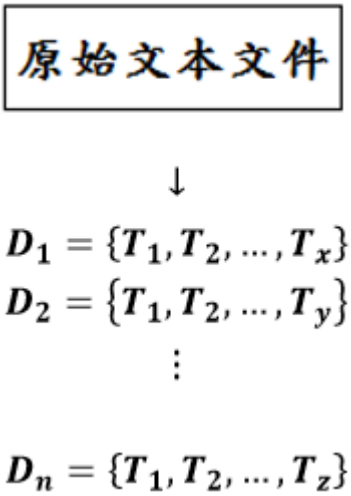

圖 2、jieba(結巴)自動分詞後結果示意圖

資料來源：本研究整理

4. 訓練與測試資料集切分

本研究將經過使用者層級過濾後之資料，以隨機給予每篇評論文章一個介於 0-1 的數值的方式做資料集區分，一為訓練資料，二為測試資料。小於 0.2 的歸類為測試資料集，大於的則為訓練資料集，資料量比例約為 80%：20%。

將數據集分為訓練資料與測試資料兩個子集，訓練子集可再區分為 D1:useful

train, D2:useless train, 測試集則為 D3:useless and useful test，將 D1,D2 依照上述步驟計算出潛在關鍵字詞詞庫，進行分類模型訓練，再將 D3 的評論文章投入進行測試，評估測試之分類結果。

(三)、 用戶級別檢測

有別於先前研究只專注於評論階級的檢測，參照 Sedhai et al.(2018)發表之研究增加第二層使用者信用判斷，以加強去除非實用評論被誤判為實用評論的可能性。簡單的來說，愛評網(ipeen)提供每一發布者帳號一組實用率計算，透過單一帳號的發文總數與發文被系統判定為實用文所占之比例，計算出發布者帳號實用率以裁定此帳戶之可信任度，實用率越高者表示其可信任度越高，反之則越低，經由上述之過濾過程，達到避免文本之使用文字剛好與實用文之重疊率太高的問題，為了提高分類準確性，我們認為這是必要的過濾步驟。

為了驗證使用者信用判斷的用戶級別檢測，本研究將以兩組資料集進行實驗對照，第一組為原始資料集除去發布者帳號之實用率小於 95%之評論，第二組則為原始資料集除去發布者帳號之實用率小於 99%之評論，如表 2、表 3 所示。

表 2、發文章發布者實用率 95%以上之有效文章數量整理表

		美食評論文章數量整理			
年份		2008-2012	2013-2015	2016-2018	總計
總文章數量	實用文	470	420	329	1219
	非實用文	142	147	189	478
95%有效文章數	實用文	407	371	264	1042
	非實用文	142	147	189	478

資料來源：本研究整理

表 3、發文章發布者實用率 99%以上之有效文章數量整理表

		美食評論文章數量整理			
年份		2008-2012	2013-2015	2016-2018	總計
總文章數量	實用文	470	420	329	1219
	非實用文	142	147	189	478
99%有效文章數	實用文	319	291	214	824
	非實用文	142	147	189	478

資料來源：本研究整理

若經本研究實驗結果得正，可推得發布者帳號之實用率與發布者之文章為實用與否之正向關係，本研究將事先將資料集內發布者實用率小於 100%之實用文予以去除，以此提高並收斂實用文之構句特徵。

(四)、 特徵選取及關鍵字詞庫建立

此章節是提取原始文本中有意義的單詞作為每篇文章特徵的主要過程，主要分為二大步驟，一為主題模型分析，二則為字詞頻率計算與特徵選取。

1. 主題模型(Topic Model)分析

(1) 主題模型定義

主題模型(Topic Model)指以統計的方式，計算文檔內所包含詞語出現之頻率，從大規模文件中發現隱藏的主題結構與抽象主題的統計模型，主要應用於機器學習和自然語言處理的領域中，在主題模型分析之研究中，主要以隱含狄利克雷分布(Latent Dirichlet Allocation，LDA)最為普及。以主題性的分群專注於較小且相似之領域，而後進行關鍵字詞的提取以獲取較大之分類準確性，進而更精準分析消費者行為、文章構句特徵與消費者對文章閱讀的喜好，以利支援企業產品與服務產銷決策制定與未來趨勢分析。

隱含狄利克雷分布(Latent Dirichlet Allocation，LDA)為 Blei et al.於 2003 提出之一種非監督式學習的聚類演算法，在模型訓練時僅需指定主題的數量參數 k 即可，每

篇文章通常都包含數個主題(Topic)，且每個主題(Topic)所占的比例各不相同，每個主題(Topic)都可以用數個關鍵字詞來描述，且相同的字詞可同時出現在不同的主題之間；以圖 3 之範例所示，其表文章內含 Arts、Budgets、Children 和 Education 四個主題，設定 n_top_words 取得各主題(Topic)的前 n 個關鍵字詞以描述並推得主題名稱與特性，並於文章內文中以不同的顏色標示各個主題之佔比。

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

圖 3、單篇文章之 LDA 主題及包含的字彙示意圖。

資料來源：Latent dirichlet allocation. (Blei, Ng, & Jordan, 2003)

(2) 隱含狄利克雷分布(LDA)運作原理

隱含狄利克雷分布(Latent Dirichlet Allocation，LDA)是一種典型的詞袋模型(bag of words)，其將整個文本當作一組詞的構成的集合，將所有字詞當作獨立的個體，忽略各個字詞的先後順序關係，利用詞頻-逆向文件頻率 (Term Frequency-Inverse Document Frequency，TF-IDF)計算字詞出現的頻率以制定主題特徵，是一種無監督式的學習演算法。

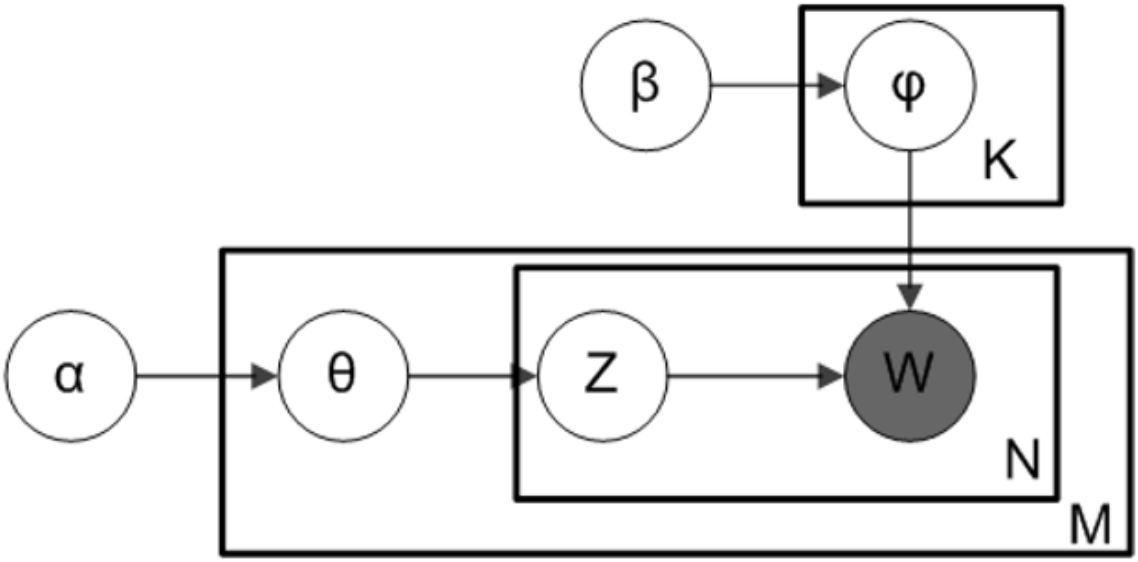

圖 4、LDA 貝葉斯網絡結構。

資料來源：Wikipedia - Latent Dirichlet Allocation

以圖 4 進行 LDA 模型參數說明，K 為主題的總數量，N 為文檔中的總字詞數量，M 為資料集內所含之文本數量，α 為主題在文檔中分佈的 Dirichlet 先驗參數，β 為字詞在主題中分布的 Dirichlet 先驗參數，θi為第 i 篇文件的主題分佈，φk為字詞在主題 k 的分佈，Zij 為文件 i 中第 j 個單字的主題，W 則為最終產生的字詞。

LDA 主題模型的建立流程可分為四個步驟，首先從主題在文檔中分佈的 Dirichlet 先驗參數 α 中取樣生成文檔 i 的主題分布θi，第二從主題的多項式分布θi中取樣生成文檔 i 第 j 個詞的主題 Zij，第三從字詞在主題中分布的 Dirichlet 先驗參數 β 中取樣生成主題 Zij 的詞語分布φk，最後，從詞語的多項式分布φk中採樣最終生成詞語 W，重複上述過程 N 次，就產生主題 K 及主題 K 所包含之關鍵字詞 (Wikipedia:隱含狄利克雷分布, 2017)。

LDA 模型在估計模型內參數的方法有很多種，最初提出的時候主要使用 Expectation-maximization algorithm(Bilmes, J. A. , 1998) 演算法求解，近期則普遍採用 Gibbs sampling(Griffiths, T. L.& Steyvers , 2004) 進行參數估計，最終可得主題－詞語矩陣 φ 與文檔－主題矩陣 θ 。

(3) 本研究之 LDA 演算法

LDA 主題模型扮演在大量文本中發掘潛在主題的工具，在上述的定義之中有提到 LDA 主題模型為一種非監督式學習的分群演算法，僅需在訓練前設定主題數量的

參數即可，因此需進行 LDA 最適合之分群數量分析，即為主題數量，而後再進行文檔－主題之機率矩陣θ與主題－字詞之機率矩陣φ計算，前者可得該篇文本之各主題佔比以推得該篇文章所屬主題，後者則可印出代表各主題之重要關鍵字以推得主題名稱。

而如何評估主題模型建構的好壞一直以來都是個值得關注的議題，蕭昱維(2014)所提出之研究表示可透過制定適當的主題(Topic)個數並計算這些主題之間的組內相似度來評估主題建模的好壞。從歷史文獻中可得知一般設定主題(Topic)個數之方式為依照經驗法則居多，其根據數據集大小來設置主題(Topic)個數，數據集越大則設定較大主題(Topic)數量，反之則較小，再經由反覆地過程調整主題(Topic)個數來觀察分群結果的好壞，而 Blei et al.(2003)提出主題(Topic)個數的設置可以透過計算 Perplexity-Topic Number 值來設定，Perplexity 的值表示主題模型對於單一文檔屬於哪個主題(Topic)的不確定因素有多大，也就是說主題模型的 Perplexity 值越小其模型的分群效能越高。

基於 Blei et al.(2003)發表之研究中提到，基於類別複雜度的 Perplexity 算法公式如公式 1 所示：

$$\text{perplexity(D)} = \exp\left\{\frac{-\sum d\log(p(w_d))}{\sum dN_d}\right\}$$

(公式 1)

以公式 1 進行 Perplexity 參數說明，D 代表文本資料集，Wd代表第 d 篇文本所含之字詞個數，N 則表示文本資料集內含總字詞個數，而 $P(w) = \Sigma z p(z|w) * p(w|z)$，代表對詞袋模型(Bag of Words)裡的任意字詞 W 來說，該字詞在主題上的分佈機率和該字詞所在文檔的主題分佈機率乘積。

由上述敘述可以得知，當主題(Topic)個數設置越多，則 Perplexity-Topic Number 值將會越小，但可能會因使用過多參數的關係出現 overfitting 的問題，且將使研究者不容易理解各主題所代表的含意；因此同時考量主題(Topic)個數與 Perplexity-Topic Number 值以達到綜合平衡，才能建構較佳之主題模型。

綜上所述，本研究欲探討愛評網(ipeen)之美食類文章之實用與非實用文辨別，由於美食類文章又可隨著餐廳類別、發布時間或餐廳位置等細節向下分成不同主題，其所含之關鍵字詞也會有所不同，若以原始文本直接加以分類分辨可能會因雜訊太多，

導致特徵關鍵字的擷取不夠精準，而若採用人工方式從大量文本中以單一類別進行分類，例如餐廳種類，可能會因只考慮單一變數而造成分類不夠全方位，最終也將影響關鍵字詞庫的提取。

因此，本研究以 LDA 主題模型分析進行主題分群，採用先分群後分類之模式，將文章中的字詞進行聚類，根據文本相似性自動化整理、歸納出數個族群，利用上述 perplexity-topic number 值的計算方式，得到並掌握資料集內被評論發布者們討論的主題數量並依照群內關鍵字推得群集特性與名稱，再依各主題以本研究之 TF-IDF 演算法快速建立各群之特徵關鍵字詞庫，以進行下一步的分類與預測作業，藉此規模化的整理以提高分類準確度。

2. 詞頻-逆向文件頻率(Term Frequency-Inverse Document Frequency，TF-IDF)

(1) 詞頻-逆向文件頻率定義

Salton and Buckley et al.於 1988 提出的「詞頻與逆向文件頻率(Term Frequency /Inverse Document Frequency，TF-IDF) 」是一種常用於資訊檢索與文字探勘的常用統計加權技術，用來評估一個「字詞」對於一個「文件」或「資料集」的重要程度。利用 TF-IDF 法，計算詞彙的權重，萃取出權重較高之關鍵詞，以降低特徵向量的維度。

字詞重要性隨著其出現在文件的次數成正比增加，但同時會隨著其在資料集中出現的頻率成反比下降，以 TF(Term Frequency)－詞頻來說，詞頻計算的是該「字詞」出現在該文件中的次數，但因每篇文章內容長短將影響詞頻計算，所以須將詞頻進行標準化，以比例來代表重要性。公式如下：

$$tf_{i,j} = \frac{n_{i,j}}{\sum k n_{k,j}}$$

(公式 2)

其中， ni,j 指的是該「字詞 i」在該文件 dj 中的出現次數，而分母則是在文件中所有字詞出現次數之和。也就是說 tfi,j 表示字詞 i 在文件 j 中出現的頻率，表示出現在某一特定文件中的詞 ti 的重要性。

矩陣的行代表文本文章的數量，矩陣的列代表初步詞庫內所有的字詞，也就是上述自動分詞後各文件之相異字詞的聯集，矩陣之值則為特定字詞在特定文件中之詞頻。

$$\begin{matrix} word_1 \\ \vdots \\ word_t \end{matrix}\begin{bmatrix} n_{1,1} & \cdots & n_{1,d} \\ \vdots & \ddots & \vdots \\ n_{t,1} & \cdots & n_{t,d} \end{bmatrix} \rightarrow \begin{matrix} word_1 \\ \vdots \\ word_t \end{matrix}\begin{bmatrix} tf_{1,1} & \cdots & tf_{1,d} \\ \vdots & \ddots & \vdots \\ tf_{t,1} & \cdots & tf_{t,d} \end{bmatrix}$$

圖 5、tfi,j 之計算示意圖

資料來源：本研究整理

而 IDF(Inverse Document Frequency)－逆向檔案頻率則計算字詞出現在所有文件之頻率次數，用來判別該字詞在整體資料集中的重要程度。公式如下：

$$idf_i = \log(\frac{|D|}{|\{d : t_i \in d\}|})$$

(公式 3)

其中，|D|指的是資料集中的所有文件的總數，分母為包含字詞 ti 的文件數目，由公式可推得該字詞在越多文件中出現之idf將趨近於 0，表示其重要性較低，可能為「我、然後、而且」這一類型的詞，而idf很大時，表示該字詞只在少數文件中出現，鑑別度與重要性則較大。

$$\begin{matrix} word_1 \\ \vdots \\ word_t \end{matrix}\begin{bmatrix} n_{1,1} & \cdots & n_{1,d} \\ \vdots & \ddots & \vdots \\ n_{t,1} & \cdots & n_{t,d} \end{bmatrix} \rightarrow \begin{bmatrix} idf_1 \\ \vdots \\ idf_t \end{bmatrix}$$

圖 6、idf之計算示意圖

資料來源：本研究整理

綜上所述，TF-IDF 便是透過 TF 和 IDF 相乘做為字詞的權重值，特定文件內的高字詞頻率，以及該字詞在整個資料集中的低文件頻率，可以產生出較高權重的 TF-IDF。公式如下：

$$tfidf_{i,j} = tf_{i,j} \times idf_i$$

(公式 4)

以上可得出每一個字詞 i 對每一篇文件 j 的權重值，因此，可藉由 TF-IDF 來計算特定關鍵字於資料集內的權重值，並設定權重門檻以利關鍵字詞篩選，過濾掉常見的字詞，保留重要的關鍵字詞。

(2) 本研究之 TF-IDF 演算法

自動分詞的分詞成效將影響模型之分類準確度，在文本分析中注重的是關鍵字的

品質而非數量，過多的字詞將導致效率降低以及雜訊過多的現象。

透過 TF-IDF 的計算可得知在此資料集內所有曾出現過的字詞的權重值，有鑑於某些字詞可能僅出現過一次，或是只出現在單一文件內，而 TF-IDF 的做法是須將文本內容結構化，將文章內文轉換成特徵向量，而特徵向量的維度數便是關鍵字詞的數量，因此如何提取重要性高的關鍵字以及提取多少數量，以適度降低特徵向量的維度為本研究所主要探討的議題。

本研究分別計算實用文集與非實用文集之所有字詞頻率($tfidf\ \ i,j$)，並列出 TF-IDF 矩陣，相較於非實用文集而出現在實用文集中頻率較高的字詞，我們視為高重要性之關鍵字詞。以下圖示之：

$$\begin{matrix} word_1 \\ \vdots \\ word_t \end{matrix}\begin{bmatrix} Useful\_tfidf_{1,1} & Useless\_tfidf_{1,2} \\ \vdots & \vdots \\ Useful\_tfidf_{t,1} & Useless\_tfidf_{t,2} \end{bmatrix}$$

圖 7、實用與非實用文集之$tfidf\ i,j$計算示意圖

資料來源：本研究整理

將包含在實用文集與非實用文集中所出現的所有字詞列出，若只出現於單一文集，則以零表示。將實用文集之各字詞的字詞頻率($Useful\_tfidf\ i,j$)減去對應的非實用文集的字詞頻率($Useless\_tfi\quad f\ i,j$)，則可得此字詞對於判別評論是否為實用文的重要性，相減之權重值越大則代表字詞重要性越高，反之則越低，而後設定權重值門檻以篩選關鍵字詞數量，即為特徵向量之維度數。

3. 關鍵字詞閥值設定與特徵選取

選取之特徵過量將導致維度向量太高，雜訊過多將增加文章分類的難度，而維度過低又將造成重要詞彙遭到刪除的問題，透過設定適當的閥值，以找出特徵集中信息量最大的特徵量。

本研究使用實用文集之字詞頻率 Useful_tfidfi,j 與非實用文集之字詞頻率 Useless_tfidfi,j 相減後之 Diff_tfidf 為衡量值，其值涵義等同於相較非實用文集而出現在實用文集中頻率較高的字詞頻率，值越高則代表重要性越高。

$$\begin{matrix} word_1 \\ \vdots \\ word_t \end{matrix}\begin{bmatrix} Useful\_tfidf_{1,1} - Useless\_tfidf_{1,2} \\ \vdots \\ Useful\_tfidf_{t,1} - Useless\_tfidf_{t,2} \end{bmatrix} \rightarrow \begin{bmatrix} Diff\_tfidf_1 \\ \vdots \\ Diff\_tfidf_t \end{bmatrix}$$

圖 8、實用文關鍵字詞之 tfidfi,j 計算示意圖。

資料來源：本研究整理。

本研究以設定字詞權重門檻作為閥值，以決定詞庫之關鍵字數量及維度向量，根據本實驗結果，如果閾值太低，會導致關鍵字詞彙不足以判別分類，如果閾值太高，則會導致實用詞彙和非實用詞彙之間的重複性過高。

(五)、 批次時間更新模型

上述關鍵字特徵選取之步驟為一個定期更新的過程，使用小量已標記數據集對分類器進行訓練後，再對測試數據集進行實用性與非實用性兩項分類，評論文章以每三年批次增加，實用文之關鍵字詞庫也會依次更新，更新後之已分類數據則視為下一次迭代的實例，重複以上過程直到分類完成，觀測擴增資料集後之分類準確度變化。

表 4 、本研究數據集描述。

	數據集描述
原始數據集	ipeen 愛評網的美食頻道排名前15間推薦餐廳2008-2012年之評論文章
1st迭代更新	ipeen 愛評網的美食頻道排名前15間推薦餐廳2008-2015年之評論文章
2nd迭代更新	ipeen 愛評網的美食頻道排名前15間推薦餐廳2008-2018年之評論文章

資料來源：本研究整理。

(六)、 本研究分類器與評估方法

1. 分類器介紹

在網路評論類型的分類過程中，本研究根據 Sebastiani 於 2002 提出之研究整理，以常見之資料探勘分類技術：支持向量機(Support Vector Machine，SVM)，單純貝氏分類器(Naive Bayes classifier，NB)，和隨機森林樹(Random Forests，RF)等三項演算法執行模型訓練與評論分類。

以上三種分類演算法皆屬於監督式學習方法，以下將依序介紹：

(1) 支持向量機(Support Vector Machine，SVM)

SVM 是一種知名的二元分類器(Binary classifier)，運作核心概念為在空間中建構一個超平面(Hyperplane)做為資料的分類基準，讓資料能夠被區分成兩個不同的集合，資料分隔在二維空間中指的是一條線，在三維空間中指的則是一個平面，而在更高維度的空間便使用「超平面(Hyperplane)」來概括，如圖 3-12 所示，可得出 H3 為最

好的分類間隔線。

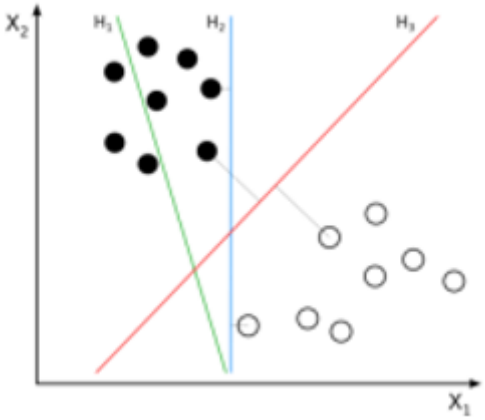

圖 9、二維空間中的 SVM 超平面示意圖

資料來源：Wikipedia - Support Vector Machine

此技術被廣泛使用於分類(Classification)和回歸分析(Regression)之統計理論，由統計學家 Vapnik 等人於 1995 年所提出，為一種監督式學習的演算法，主要使用於高維或無限維空間中，常用於文本圖像分類、手寫字型辨識和人臉識別等。 分類評估的原理為認定最靠近超平面(Hyperplane)的訓練資料點即為該資料類別的邊界，稱之為 Margin，而 Margin 值越大代表此模型能更更明確的分辨未知的資料點是屬於哪個集合，以圖 10 所示。

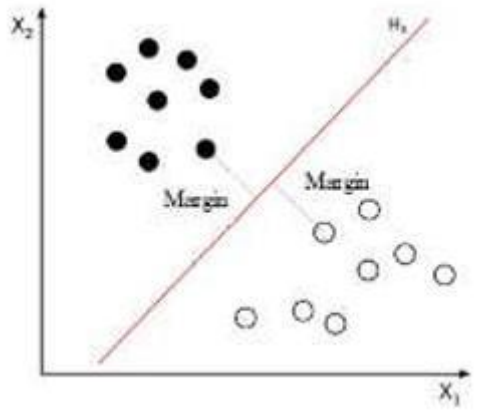

圖 10、SVM 之 Margin 邊界

資料來源：Wikipedia - Support Vector Machine

(2) 單純貝氏分類器(Naive Bayes classifier，NB)

NB 是一種基於條件機率的分類方法，最常使用於文本分類，為一種監督式學習的演算法。 以條件機率解釋，我們假設 A、B 為樣本空間中的兩獨立事件，在 B 情況下發生 A，與在 A 情況下發生 B 的機率是不一樣的。則在給定 B 事件

發生之下，A 事件的條件機率以 P(A|B)表示，反之則以 P(B|A)：

$$P(A|B) = P(A^B)/P(B)$$

(公式 5)

$$P(B|A) = P(A^B)/P(A)$$

(公式 6)

從公式 5 與公式 6 可推導出貝氏定理 P(A|B) = P(B|A) * P(A)/P(B)，透過機率的計算，結合條件獨立假設，假設樣本中每個特徵字詞皆互相獨立且同等重要，去計算詞出現在文本集 A 與文本集 B 的關聯，以計算未標記資料屬於何類別之機率。

(3) 隨機森林樹(Random Forests，RF)

RF 是一種組合學習演算法，包含多個決策樹的分類器，利用 Ensemble Method（集成方法）將多個分類器組合起來，以圖 3-14 為例，並必須滿足兩項限制，一為各個分類器之間須具有差異性，二則為每個分類器的準確度必須大於 0.5，而每一棵決策樹之間都是互相獨立的，輸出類別由各別樹所輸出的類別之眾數而定，此技術也被廣泛使用於分類(Classification)和回歸分析(Regression)。

使用 Bagging 採樣方式將訓練樣本進行分化，產生具有差異性的決策樹，做法為從訓練資料集中取出 M 個樣本並訓練出 M 個分類器，樣本特徵採用後會放回，因此這 M 個樣本之間會有資料重複的部分，而每個分類器的樣本不同，輸入的樣本也並非全部的樣本，使其相對不容易出現 over-fitting，訓練出的分類器也會具有差異性。

綜上所述，因為各個分類器皆為從 M 個樣本特徵中選擇 m 個樣本進行學習，因此我們可將各個分類器視為精通某一領域的專家，利用不同的決度去分類分析新的未知數據，最終再由投票得出結果，其結果將比以單一分類器之分類效能來的佳，並且能夠處理具有高維特徵的輸入樣本且不需要進行降維。

2. 評估方法

最後，以混淆矩陣(Confusion Matrix)如表 3 所示，定義文件分類的準確率 accuracy、精確率 precision、召回率 recall 和 F-score。

表 5 、混淆矩陣（Confusion Matrix）

真實/預測	condition positive	condition negative
Predicted condition positive	True positive(TP)	False positive(FP)
Predicted condition negative	False negative(FN)	True negative(TN)

資料來源：整理自 Provost et al. (1998)。

在混淆矩陣(Confusion Matrix)中，四個象限代表四種分類結果，以本研究之分類來說明：

- True positive 表示樣本評論與模型預測結果皆為實用文的個數。
- False positive 則表示預測為實用文，實際為非實用文的個數。
- True negative 為樣本評論與模型預測結果皆為非實用文的個數。
- False negative 則為預測為非實用文，實際為實用文的個數。

檢測分類是否正確最直觀的評估方法為準確率(Accuracy)，將正確分類之 TP 和 TN 總個數除以所有驗證資料的總個數，但此評估方法可能受到不平衡資料之影響造成分類判斷誤差。

為了解決以上問題而衍生之精確率(precision)、召回率(recall)計算則受到廣大使用，前者為以預測結果為基準，實際預測正確的精準度為多少，後者為以實際情形為正向的情形下，能預測多少正確的正向答案，因此本研究預計採用同時考量其二者之 F-score 進行分類效能評估。

四. 研究成果

在本章節中，將依據研究方法之章節架構說明本研究之實驗過程與結果，以驗證本研究改良之 TF-IDF 演算法所提取之關鍵字詞庫之於實用文辨別之成效。4.1 將分別說明資料收集與預處理結果與用戶級別檢測過濾與閥值設定結果；4.2 則進行文本分類的實驗並產出結果，進而探討研究目的。

表 7 、用戶級別檢測過濾之 F1-score 結果。

		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
99%	SVM	0.6163	0.6183	0.5978	0.6085	0.5856	0.5726	0.6047	0.5813	0.6
	NB	0.6692	0.6704	0.7041	0.6299	0.6054	0.6102	0.6144	0.5949	0.5725
	RF	0.568	0.5809	0.5747	0.5343	0.495	0.568	0.5422	0.5454	0.5754
95%	SVM	0.5294	0.5501	0.5536	0.5307	0.5399	0.5294	0.5342	0.5339	0.5277
	NB	0.5966	0.5973	0.5921	0.5821	0.5737	0.5966	0.5596	0.5404	0.5396

	RF	0.4519	0.5	0.4881	0.5189	0.4623	0.4607	0.473	0.4986	0.4607

資料來源：本研究整理。

(一)、 資料預處理

1. 資料收集結果

本研究採用 ipeen 愛評網之美食頻道評論作為實用文與非實用文檢測之實驗資料，共計採用 ipeen 愛評網推薦之前 15 間餐廳，包含美好年代、隨意鳥地方與螺絲瑪莉等，且發表時間為 2008 年 1 月至 2018 年 12 月內，共 1219 篇實用文評論與 478 篇非實用文評論作為資料來源。

表 6、原始資料集之實用與非實用文比例。

	實用文篇數	非實用文篇數	總計
數量	1219	478	1697
比例	72%	28%	100%

資料來源：本研究整理。

2. 用戶級別檢測過濾與閥值設定

在本節實驗中，為了驗證除了評論階級檢測，使用者階層之信用判斷是否也為分類評估標準重要的一環，本節以兩組資料集進行實驗對照，以下為資料集介紹：

- 原始資料集除去發布者帳號之實用率小於 99%之評論。
- 原始資料集除去發布者帳號之實用率小於 95%之評論。

發布者帳號之實用率定義為以下，愛評網(ipeen)透過單一帳號的發文總數與發文被系統判定為實用文所占之比例，計算出發布者帳號實用率以裁定此帳戶之可信任度，實用率越高者表示其可信任度越高，反之則越低。

本節實驗依照第三章研究方法之研究設計流程，將原始資料集經過正則表達

式去除英文字母與標點符號等字元，再依精確模式去除停用字詞後進行文本分詞步驟，最後以本研究設計之 TF-IDF 演算法，分別計算出實用文集中之所有字詞頻率($Useful\_tfidf\ i,j$)與非實用文集中之所有字詞頻率($Useless\_tfidf\ i,j$)，並使相對應之字詞相減，得到各個字詞的 Diff_tfidf 作為關鍵字詞之權重衡量值。

依照上述步驟計算出 Diff_tfidf，其含義為相較於非實用文集而出現在實用文集中頻率較高的字詞，我們視為高重要性之關鍵字詞。在此我們將權重閥值設定於 0.1-0.9 區間進行實驗觀測，使用時間區間設定為 2008-2018 之文本資料集，觀察支持向量機

(Support Vector Machine，SVM)與單純貝氏分類器(Naive Bayes classifier，NB)與隨機森林樹(Random Forests，RF)三種分類器之 F-score 在上述兩組實驗中的分類效能變化結果。

由表 7 可得知用戶級別檢測過濾之 F1-score 結果，原始資料集經過用戶級別檢測，過濾發布者帳號之實用率較低之評論後，三種分類器在 0.1-0.9 實驗區間設定中皆能得到較佳的分類效果，確認使用者帳戶之發文實用率將會影響實用文的分類判定。

因此，本研究將各篇評論之發布者帳號之實用率小於 100%之評論皆於以過濾，藉此降低訓練資料集中的雜訊影響力，經由此步驟，原始文本資料集由 1,219 篇實用文評論與 478 篇非實用文刪減至 549 篇實用文評論與 478 篇非實用文評論所組成，共過濾掉 670 篇發布者帳號之實用率小於 100%之實用文評論，並依照評論發布日期整理成表，如表 8 所示。

表 8、發文章發布者實用率 100%以上之有效文章數量整理表

		美食評論文章數量整理			
年份		2008-2012	2013-2015	2016-2018	總計
總文章數量	實用文	470	420	329	1219
	非實用文	142	147	189	478
100%有效文章數	實用文	225	165	159	549
	非實用文	142	147	189	478

資料來源：本研究整理

刪除後得之資料集再進行訓練集與測試集區分，區分方式以隨機給予每篇評論文章一個介於 0-1 的數值的方式，以 80%：20%的比例原則將其區分成訓練集與測試集，再依照評論發布日期的時間區間整理成表，可分為原始數據集、1st 迭代更新與 2nd 迭代更新，時間間隔為三年，將於下一章節進行擴增資料集後之分類準確度變化驗證，如表 9 所示。

表 9、用戶級別檢測過濾後之資料集介紹。

	原始數據集	1st迭代更新	2nd迭代更新	測試集
評論篇數	290	534	815	211
實用文數	184	311	442	107
非實用文數	106	223	373	104
時間區間	2008-2012	2008-2015	2008-2018	2008-2018

資料來源：本研究整理。

(二)、 研究結果

本章節依用戶級別檢測過濾結果，將各篇評論之發布者實用率小於 100%之評論皆於以過濾得之資料集作為訓練與測試集，如表 4-10 用戶級別檢測過濾後之資料集介紹所示，在關鍵字詞庫提取後，實用與非實用文集之 tfidfi,j 計算後得之關鍵字詞庫包含 33,334 個相異的字詞聯集。

以時間區間設定為 2008-2018 之 2nd 迭代更新數據集進行支持向量機(Support Vector Machine，SVM)，單純貝氏分類器(Naive Bayes classifier，NB)，和隨機森林樹(Random Forests，RF)三種分類器的最佳關鍵字詞之閥值找尋，再依照各分類器之最佳閥值進行批次時間更新模型之準確度觀測，根據關鍵字閥值設定結果可得三種分類器於閥值設定為 0.3 時分類效果最佳，因此以$Diff_tfidf$為 0.3 作為主要文本分類實驗之設定。

表 10、本實驗$Diff_tfidf$為 0.3 之分類結果。

	分類器	accuracy(%)	precision	recall	F-score
2008-2018	SVM	70.05	0.783	0.613	0.687
	NB	74.62	0.759	0.773	0.754
	RF	76.63	0.777	0.608	0.682
2008-2015	SVM	62.25	0.58	0.648	0.613
	NB	73.1	0.757	0.735	0.746
	RF	61.76	0.586	0.574	0.58
2008-2012	SVM	61.14	0.619	0.548	0.547
	NB	65.4	0.663	0.6	0.633
	RF	59.24	0.68	0.326	0.441

資料來源：本研究整理。

由分類分析結果可得知，當使用來自 2nd 迭代更新數據集(2008-2018)作為 NB 分類器的訓練集時，我們可實現的最佳模型之 F-score 為 0.754，準確度為 74.62%，精確度為 0.759，召回率為 0.773。

伍、結論與未來發展方向

本研究中透過提出一監督式學習的迭代模型框架，並以三種不同之分類器(Support Vector Machine、Naive Bayes classifier、Random Forests)來探討並辨別數據集中的實用文評論，預期可以有效減少消費者在蒐集與整理文章資料上所花費的效力，使消費者迅速得到對自身有用之資訊，證實網路評論文章之影響力，找出實用文與非實用文之間的文章構句特徵差異，綜合整體資訊以分析相關評論。

而本研究實驗結果證明定期擴增數據集以及使用者層級過濾在檢測實用文中的有效性，參照網路評論發布者之來源可信度，提高實用文辨別門檻，並以批次更新關鍵字詞庫內容的方式，可得評論隨著時間變化所指涉的關鍵字變化，進而提升實用文與非實用文之辨別分類的準確性。

未來希望可以針對跨領域之評論文章進行辨別分類，不須局限於美食推薦文當中，使得消費者能夠透過此研究優化整體搜尋有效資訊的流程，在實務上提供真正符合消費者需求的服務。

陸、參考資料

[1] 王力弘. 社群媒體新詞偵測系統 以 PTT 八卦版為例. Diss. 王力弘, 2015.

[2] 江義平、溫演福、廖奕翔、陳靖翔、陳佳駿，2012，網路文字探勘技術運用於智慧型手機口碑之分析研究，國立台北大學資訊管理研究所。

[3] 任柏衛（2015）。基於文章分析的美食推薦系統。國立清華大學通訊工程研究所碩士論文，新竹市。 取自 https://hdl.handle.net/11296/vj93b7

[4] 李啟誠, & 李羽喬. (2010). 網路口碑對消費者購買決策之影響—— 以產品涉入及品牌形象為干擾變項. 中華管理評論學報, 第十三卷一期, 1-23.

[5] 吳珮菁（2012）。意見探勘分析顧客行為之研究。國立成功大學資訊管理研究所碩士論文，台南市。 取自 https://hdl.handle.net/11296/8h9d86

[6] 林名彥（2015）。應用文字探勘技術於客訴資料之研究-以台大 PPT 論壇為例。龍華科技大學資訊管理系碩士班碩士論文，桃園縣。 取自 https://hdl.handle.net/11296/8u7ft9

[7] 林國仲（2017）。運用情緒分析結合產品多面向自動分類於消費者評價之研

究。國立臺南大學數位學習科技學系數位學習科技碩士在職專班碩士論文，台南市。取自 https://hdl.handle.net/11296/r4fdnz

[8] 陳世榮. "社會科學研究中的文字探勘應用: 以文意為基礎的文件分類及其問題." 人文及社會科學集刊 27.4 (2015): 683-718.

[9] 劉力華. "應用資料探勘於手機評論文章分類之研究." 電子化企業經營管理理論暨實務研討會 (2010): 294-303.

[10] Bickart, Barbara, and Robert M. Schindler. "Internet forums as influential sources of consumer information." Journal of

2019 資訊管理暨電子商務經營管理研討會論文集

18

interactive marketing 15.3 (2001): 31-40.

[11] Chen, Zhouhan, et al. "Hunting Malicious Bots on Twitter: An Unsupervised Approach." International Conference on Social Informatics. Springer, Cham, 2017.

[12] Chavoshi, Nikan, Hossein Hamooni, and Abdullah Mueen. "DeBot: Twitter Bot Detection via Warped Correlation." ICDM. 2016.

[13] Davis, Clayton Allen, et al. "Botornot: A system to evaluate social bots." Proceedings of the 25th International Conference Companion on World Wide Web. International World Wide Web Conferences Steering Committee, 2016.

[14] Eagly, Alice H., Wendy Wood, and Shelly Chaiken. "Causal inferences about communicators and their effect on opinion change." Journal of Personality and social Psychology 36.4 (1978): 424.

[15] Garg, Ranjna. "Study of text based mining." Proceedings of the International Conference on Advances in Computing and Artificial Intelligence. ACM, 2011.

[16] Narayan, Rohit, Jitendra Kumar Rout, and Sanjay Kumar Jena. "Review Spam Detection Using Semi-supervised Technique." Progress in Intelligent Computing Techniques: Theory, Practice, and Applications. Springer, Singapore, 2018. 281-286.

[17] Ott, Myle, et al. "Finding deceptive opinion spam by any stretch of the imagination." Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011.

[18] Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis." Foundations and Trends® in Information Retrieval2.1–2 (2008): 1-135.

[19] Provost, Foster, and R. Kohavi. "Glossary of terms." Journal of Machine Learning 30.2-3 (1998): 271-274.

[20] Salton, Gerard, and Christopher Buckley. "Term-weighting approaches in automatic text retrieval." Information processing & management 24.5 (1988): 513-523.

[21] Sebastiani, F. (2002). Machine learning in automated text categorization. ACM computing surveys (CSUR), 34(1), 1-47.

[22] Sedhai, Surendra, and Aixin Sun. "Semi-supervised spam detection in Twitter stream." IEEE Transactions on Computational Social Systems 5.1 (2018): 169-175