網路新聞讀者閱後情感之預測

Sentiment Prediction for Internet News Readers

張昭憲
淡江大學資訊管理系
新北市淡水區英專路151號
jschang@mail.tku.edu.tw

沈育信
淡江大學資訊管理系
新北市淡水區英專路151號
ahsinshen@gmail.com

摘要

隨著社群網路的蓬勃發展，人們已習慣在網路上發表個人看法，留下無數的數位足跡。若能蒐集這些發言，透過系統化的情感分析(sentiment analysis)，便可快速得知群眾的想法或傾向。在吸引群眾參與的網路媒體中，除網路社群與即時通訊平台外，網路新聞亦是非常重要的一環。若能預測大多數讀者閱後之情緒反應，各公私領域(政府、選舉、娛樂、運動等)決策者便可在發布新聞前，事先調整內容，以獲得更多的關注與正面評價。然而，有別於傳統的情感分析，網路新聞因新創詞多、讀者情感變化快速、新聞用語與讀者反應關聯度低等特點，需採用不同的方法來因應。為此，本研究發展了一套有效的讀者情感預測方法，以預測新聞刊登後可能引發之群眾情緒。首先，透過蒐集特定情感分類下的所有新聞，進行N-gram斷詞分析。藉由統計文章中常用字詞排序表，產生各種不同的讀者情感預測模型。為準確預測讀者情感，本研究使用三種相似度計算方式，將待測新聞進行情感分類。為驗證提出方法之有效性，我們蒐集Yahoo奇摩新聞將近一年共193,489筆新聞進行實驗。結果顯示，在相關新聞數量足夠時，本研究提出方法具有良好之預測準確率。其次，當新聞蒐集天數增長時，準確率可獲得明顯提升，但需考量新聞熱度持續時間。此外，當有重大新聞發生時，控制塑模的時間點可獲得更佳的預測結果。上述結果說明本研究發展方法之有效性，若能實際應用於各領域之新聞發布，將可提供有效之決策支援。

關鍵詞: 情感分析、N-gram、資料分類、文字探勘、網路新聞。

Abstract

In the past few years, Internet social community has been growing rapidly. People get used to post their opinions on the Internet and thus leave tremendous amount of digital footprints. To understand what people are thinking, one of the best ways is to perform systematic sentiment analysis on those posts. In addition to social communities and instant messengers, network news sites are also popular for people. If one can predict the sentiment of people after they read news, he can adjust the news content to attract the spot of people and obtain positive feedbacks. However, different from conventional sentiment analysis, predicting the emotion of the news readers is inherently difficult, which is caused by unknown newly-created words, fast sentiment evolvement and the mismatch of used words in news and readers' emotion. Thus, it is required to develop new analysis method to predict the emotion of news readers. To this end, an effective predicting method is proposed in this work to mine the instant public opinions on the Internet. At first, the news under test is segmented by N-gram partition. Then, a frequent word accounting table is constructed to help building the prediction model of readers' emotion. To enhance the prediction accuracy, three difference similarity measures are used to classify the news under test to a proper user emotion category. To verify the proposed method, 193,489 news are gathered from Yahoo!Kimo news site for experiments. The results show that the proposed method can achieve good prediction accuracy when the number of news is large enough for building emotion model. In addition, the accuracy can be promoted by increasing the days of news gathering. Also, when breaking news comes, one can maintain the accuracy by setting shorter data gathering period. Based on the above results, we believe the proposed method can be used as useful assistant tools for news posting.

Keywords: sentiment analysis, N-gram, classification, text mining, Internet news

一、 緒論

隨著社群網路的興盛，人們已習慣在網路上發表個人看法、關注有興趣的議題，甚至因網路意見而改變個人好惡。有鑑於此，各領域的決策者(如公共事務、商業、醫藥、教育等)便可蒐集各網路社群的發言內容，透過系統化分析，獲知群眾的想法或傾向。上述領域之相關研究，經常被稱為意見探勘(Opinion Mining)或情感分析(Sentiment Analysis) [8]。情感分析的應用非常廣泛，在政治上可經由網路留言分析，預測群眾對於政治現況的看法[26]。此外，亦可藉此了解網路訊息對於公共事務的影響，並試圖解答如「哪一個公共領域會成為熱門搜尋項目」或「哪些公共領域發聲者會影響搜尋者的看法」等問題[27]。在新世代的民主選舉中，網路意見更佔有舉足輕重的地位，在最近一次的台北市長選舉中，更實際展現其潛力。因此，選舉人可透過網路留言情感分析，了解群眾的關注議題與投票傾向。例如，Lampos 等人[18]便利用 Twitter 上的留言，分別預測英國與

奧地利民眾對於各主要政黨的投票傾向。在商業領域，He 等人[13]則針對 Pizza 產業，分析 Facebook、Twitter 上之顧客反應，以了解不同 Pizza 廠牌之顧客接受度與競爭關係。在災害應變上，為預測氣象事件(大雪、暴雨等)的發展，Lampos 等人[19]則針對群眾在網路上有關天氣之留言進行分析，以提供更即時、正確的應變訊息。在醫療保健上，Akay 等人[6]則透過網路意見探勘，即時監控民眾對特定藥品之用藥反應，做為改良或開發新藥的依據。由上可知，身處大數據年代，決策者透過網路意見探勘或情感分析，獲得更快速、客觀的決策資訊，將有助於提高決策品質，做出正確決定。

在吸引群眾參與的眾多網路媒體中，除社群網站(如 Facebook, Twitter, BBS 等)與即時通訊軟體平台(如 Line, Messenger 等)之外，網路新聞亦是非常重要的一環。事實上，透過無遠弗屆的網路連結，網路新聞以即時方式報導世界各地發生的事件，早已成為大眾獲得訊息、抒發情緒的重要管道。在瞬息萬變的環境中，正確、快速且多元的訊息(新聞)，能夠幫助管理者做出更正確的決策。以政治為例，網路新聞在民主社會中儼然已成為民情輿論的試煉場，當記者將政府施政措施傳達給民眾時，民眾也能迅速在新聞下方的討論區發表評論、進行投票，讓施政者即時了解民心向背[12]。

以台灣的 Yahoo 奇摩新聞為例(https://tw.news.yahoo.com/)，一篇熱門新聞可在短短幾小時內，便可累積上千篇的回應與投票。相反地，若讀者對某篇新聞不感興趣，可能在刊登的期間，完全沒有任何一篇回應。因此，若決策者同時針對新聞本身與這些回應進行分析統計，便可獲得寶貴的即時民情，進而預測網民對於特定新聞的可能反應。曾有研究對就讀新聞傳播相關科系的三百名香港學生進行調查，發現他們最常使用的新聞媒體前三名依序為:Yahoo 網路新聞(Yahoo!News)佔 34%，第二位為《蘋果日報》佔 24%；第三位為無線電視，僅佔 17%[1]。由此可見，在眾多媒體當中，網路新聞確實已經佔有重要的一席之地，並成為年輕世代獲得資訊的重要管道。若能在發布新聞前，正確預測大多數讀者之情感反應，必能提供各種公私領域重要的決策依據(例如，做為發布有關候選人、政府政策、影視娛樂等相關新聞時之參考)。

情感分析經過近年來的蓬勃發展，在理論與技術上均有明顯進展[9]。從早期單純的字詞正負面意義判斷，發展至推論整個段落、文章之涵義。對於讀者反應的分析，也從原先的總體正負傾向分析(polarity analysis)，延伸至對於特定主題(aspect)的多層次觀點分析(例如，發文對於某廠牌機車「排氣管」設計表達「不信任」)。分析的技術也從單純的關鍵字(keywords)分析，提升為概念(concept)分析。常見的網路意見情感分析經常需將蒐集的文章進行斷詞，此步驟可配合預設字數或字詞庫(語料庫)來分解文章。之後，再根據分析目標，配合迴歸分析、關聯規則探勘、分群、貝式機率模型、SVM(Support Vector Machine)等學習方法，預測網路留言的情感傾向[8]。

然而，在前人的研究中，鮮少對於網路新聞進行情感分析或預測讀者閱後反應。事實上，網路新聞的情感預測具有本質上的困難:

(1) 分析文章時，前人經常利用語料庫來進行斷詞，以利後續分析。此法雖然便利，但對於文章中的新創詞(未知詞)則經常誤判。例如，早期語料庫中並未有「服貿」或「一帶一路」等詞，當新聞第一次出現時，便無法正確斷詞，更遑論提供情感分析支援。

新聞報導中新創詞出現頻率較一般文章更高，且經常是新聞重點。因此，欲透過語料庫進行情感分析，有其本質上之困難。

(2) 其次，網路新聞所造成之讀者反應並非一成不變，隨著新聞發展，讀者情緒也可能發生變化。因此，發展之方法需能動態因應新聞熱度、報導方向的改變，準確預測讀者閱後之情感。

(3) 前人研究經常利用文章中字詞的情緒意涵，進行發文者之情感預測。然而，此種作法卻不適用於新聞讀者閱後情感預測。原因是新聞中描述事件之情緒用詞，可能與讀者閱後情緒完全無關。例如，報導內容為「…怒氣沖沖的駕駛立即下車理論，雙方破口大罵，造成沿路大塞車…」，內容雖含有憤怒情緒之客觀描述，但讀者可能不會感到「憤怒」，而只覺得「無聊」。因此，若以新聞中出現之情緒字詞做為讀者情感分析基礎，易造成預測失準。

(4) 再者，專業新聞記者撰寫報導時，需摒除個人意見，保持客觀中立。然而，縱使面對一篇沒有情緒用語的新聞，讀者閱讀後卻可能產生不同的情緒反應。例如，報導陳述「…德國改口宣布，戰爭結束後，將遣返難民…」，讀者可能產生「快樂」、「憤怒」等迥然不同之反應。

綜合以上討論可知，網路新聞之讀者情感預測具有其本質上複雜性與困難度，需發展有效的方法來因應，才能在大數據年代獲得隱藏在網路中的寶貴智慧。因此，為瞭解網路即時輿情，預測刊登特定新聞後民眾可能的情感反應，本研究提出一套有效的讀者新聞情感預測方法。首先，我們蒐集特定情感分類下的所有新聞，在不考慮字詞情緒意涵前提下，利用改良式 N-gram 進行斷詞分析。之後，透過整理文章常用詞，建立常用字詞出現次數排序表，以產生讀者情感預測模型。為準確預測讀者情感，本研究使用三種相似度計算方式，將待測新聞進行情感分類。為驗證提出方法之有效性，本研究蒐集 Yahoo 奇摩新聞將近一年共 193,489 筆新聞進行實驗。結果顯示，在相關新聞數量足夠時，本研究提出方法具有良好之預測準確率。其次，實驗也發現當新聞蒐集時間增長時，準確率可獲得明顯提升，但需考量新聞熱度持續時間。此外，當有重大新聞發生時，控制塑模時間點可獲得更佳的預測結果。上述結果說明本研究發展方法之有效性，若能實際應用於公私領域之新聞發布，將可提供有效之決策支援。

本論文其餘章節之架構如下: 第二節為文獻探討與相關技術簡介；第三節介紹本研究提出之新聞情感預測方法；第四節為實驗結果，將展示各種狀況下，網路新聞讀者情感之預測結果；第五節為結論與未來工作。

二、 文獻探討與相關技術

本節將介紹與本論文之相關研究，並概述使用之 N-gram、相似度計算等技術，以利後續討論。

1. 文獻探討

如前所述，人們已習慣在網路上蒐集資訊並發表個人意見。根據 Pang 與 Lillian(2008)對 2000 位美國成年人做調查，發現有 81%使用者在購買產品前，會先在網路上瀏覽相關資訊，20%的使用者願意在網路上給予評價。若能對這些發言進行情感分析，將可獲得寶貴的即時輿情，做為各領域改善決策或開發新產品的依據。由於具有高度實用性，吸引學界與業界競相投入，以下將介紹此領域之相關研究。

針對網路上的電影評論，Parkhe 與 Biswas 發展以 aspect 為基礎的情感分析方法，透過與主體相關的單字對評論進行斷詞，再利用不同的貝氏分類器進行分類，最後以投票方式決定評論對於電影的好惡(polarity) [23]。為分析選民投票傾向，Lampos 等人則以特定字詞(terms)對 Twitter 文章進行頻率分析，配合發言者個人基本資料，利用雙線性迴歸(bi-linear regression)進行選民投票傾向預測[18]。Polgreen 等人利用 2004 到 2008 年的 Yahoo 搜尋資料，透過分析類流感關鍵字佔所有搜尋的比率，與 CDC 檢驗類流感陽性反應的比率之相關度，以顯著性差異與決定係數檢定其實驗結果，以建立可用的預測模型[24]。Lampos 等人蒐集取得英國 49 個最熱鬧的城市周圍十公里內的 Twitter 網站 40 週的發文，利用有類流感關鍵字進行頻率統計。最後透過迴歸分析，建立流感就診率的預測模型[19]。Bahrainian 與 Dengel 使用 Twitter 上提供之預設屬性，但加入 Naïve Bays, SVM 等監督式方法找出的 polarity 值做為屬性，再配合詞語的正負面屬性，提出一套正負評論(Polarity Measure)的判別公式[7]。Akay 等人則透過文字探勘常用的 TF-IDF 分析，獲得文章中的重要常用字詞，再根據這些字詞的前後文關係賦予正、負向評價標籤，透過後續的網路分析方法，獲得使用者的意見回饋，以及社群中具有影響性的成員[6]。Weichselbraun 等人則透過 SenticNet 定義的詞語意涵分類(attention, pleasantness, sensitivity, aptitude)進行文章分析，並依照每個分類進行正負意向判別，最後依照詞語出現的情境(context)，產生不同的綜合正負傾向判定[28][29]。

有別於常見的正反意見偵測(polarity detection)，亦有許多學者對網路發言進行概念分析(concept analysis)。Lampos 等人歸納使用者帳號屬性(貼文數量、回文比率、非重複發文等)，使用回歸模型配合學習函數，預測 Twitter 使用者的影響力[17]。Yang 與 Dorbin 利用 Density-based Clustering 方法對網路留言進行分群，以了解網路中的活躍群體，除能發掘群眾興趣外，也可用來辨識罪犯或恐怖分子在網路上的活動[32]。Huang 等人以 ARC-BC 演算法(Association Rule-based Classifier By Categories)對 BBS 的文章進行分類，以了解成員之間的相似度與差異性[14]。Wu 等人則探討社群網路中的意見擴散(opinion diffusion)問題，配合機率公式，了解使用者之間意見相互影響的可能性與程度，並發展視覺化工具以增加分析結果的可讀性[31]。Isidro 等人提出以本體論為基礎，透過語意網路(Semantic Web)分析文章的屬性，以加強輿情探勘的適當性，避免產生無意義的探勘結果[15]。Jiang 等人則運用 MapReduce 架構，透過 PageRank 演算法找出 BBS 群組中的意見領袖[16]。Cao 等人則分析網友發言對於事件與事件間之關聯看法(例如，”沉船根本是超載”，”當初的失言風波根本只是炒作新片的發表”等發言)，並配合機率模型計算其關聯強度(Causal-Strength)[10]。

由上可知，在情感分析相關研究上，學者已提出各種可行的方法，並應用在政治、娛樂醫療等領域。然而，這些研究鮮少對於網路新聞讀者情感進行預測。網路新聞引發之情緒具有多樣性，讀者可能覺得溫暖、生氣、無聊等，而非只有正負評價判定。這些情緒也具有高動態性，讀者可能隨時間改變集體情緒。此外，新聞新創用語多，事先使用辭庫進行分析也容易失準。綜合上述原因，本研究將針對網路新聞之讀者情感，發展有效的預測方法。

2. N-gram 斷詞

為降低分析複雜度，配合語料庫進行斷詞，是文章探勘常見的用法。根據《現代漢語頻率辭典》，中文前 9000 個常用詞數量的統計表明，26.7 %的詞為單漢字，69.8 %的詞為雙漢字，2.7 %的詞為三漢字。然而，根據研究[4]，利用語料庫進行斷詞雖然便利，但對文章中的新創詞或未知詞，無法有效進行分析。在一般文章中，未知詞的比例約 3%-5%。在新聞報導中，常常有特定的名詞或者網路上出現的新詞，因此未知詞出現的比率更高。例如，早期語料庫中並未包含「美牛」一詞，當新聞第一次出現「美牛」時，會將其視為未知詞，無法提供情感分析支援。又如，近年的「服貿」與「安倍三箭」等相關經濟新聞，也均為新創詞，當時並未被包含在語料庫中。其次，傳統做法也常利用中文 stop word(例如，的、是、也、會、一個)來切割中文字詞，然而當新創詞出現時(如”馬習會舉行”)，利用 stop word 斷詞將導致新聞重點字詞被消滅(例如，將”馬習會舉行”切割為{”馬習”、”舉行”})。由上討論可知，使用語料庫斷詞在未知詞的擷取效果不佳，當以網路新聞為分析目標時，便需特別將未知詞納入考慮，以免分析效果大打折扣。為解決上述問題，本研究將採用 N-gram 概念來進行新聞斷詞。

在語言學中，N-gram 代表在文章(或講詞)中連續出現的 n 個項目(item)，由本文的第一個項目開始，到最後一個項目，以 n 個項目長度為限制來斷詞，並逐字向後遞移[11]。若以英文單字(word)為項目，則句子”This is a book.”之 1-gram (或稱為 unigram)序列便是{This, is, a, book}。若本文為”預測新聞情感”，則其 1-gram 序列為{預、測、新、聞、情、感}，其 2-gram 序列為{預測、測新、新聞、聞情、情感}。與西方語系（如英語、法語等拼音文字）相比，漢語最大的特點之一就是字與字、詞與詞之間沒有分隔符號。此外，由於詞的界限不明確，一個詞可能由一個或多個字來組成。使用 N-gram 斷詞的優點為將斷詞、斷句視為一種單純統計的過程，無須在意一個字或多個字是否有意義，亦不需要任何的辭典或語料庫。事實上，將一篇本文直接用 N-gram 分析，透過出現頻率的統計，即可找出本文中所有用過的字詞。因此，N-gram 斷詞對於經常出現未知詞的新聞分析，能提供有效的解決之道。

在實務上，N-gram 斷詞需考量的問題為如何設定 N 的最大值。對此問題，齊夫定律(Zipf's law)可提供部分解答[6]。Zipf's law 是由哈佛大學的語言學家 Zipf 於 1949 年發表的實驗定律，重點為在自然語言的資料庫中，一個單字出現的頻率與它在頻率表(frequency table)中的排名成反比。如此一來，在頻率中排名最高的單字，其出現頻率約為第 2 名單字的 2 倍，第 3 名單字的 3 倍，依此類推。參考圖 1 所示，其線型便是 Zipf's law 特質的展現。在古騰堡計畫中，總共對 421,441 本書(共 160 億個章節，23,923 位作

者，2,880,579,249 個 words)進行數位化及歸檔，並建立了 15 世紀到 19 世紀最常使用的 100,000 個英文字辭典。經過檢測，這 100,000 個字出現頻率及次數確實符合 Zipf's Law。由此可知，當使用 N-gram 對一篇文章進行斷詞時，N 的選擇可詞語出現頻率加以限縮，無需窮舉。

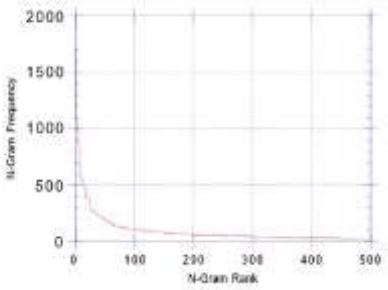


圖 1: 單詞出現的頻率與排名關係 (來源:[11])

3. 相似度計算

常見的文章或郵件分類(classification)問題，經常以監督式學習(supervised learning)方法來進行。一開始，利用既有的歷史資料做為訓練資料(training set)進行分析，找出可用以描述或區分資料類別或概念的模型。根據所獲得的模型，便可以進行未知個體類別的預測。分群(clustering)與分類不同，屬於非監督式的學習法，訓練樣本所屬的群組是未知，必須利用特定點算法來找出相近的群組。為了獲得更佳的預測結果，本研究使用分類與分群演算法中使用之相似度(similarity)計算公式，對新聞內容進行情感預測。使用的方法包含常用之歐氏距離及餘弦相似度（cosine similarity）等方法。為了簡化後續討論，以下將逐一簡介這些方法[31]。

(1) 歐氏距離(Euclidean distance)

歐氏距離廣泛用於各種相似度計算中，例如 k-means 演算法是一種常用的群聚演算法(clustering)，最終可把所有個體歸入事先指定的 k 個群聚中，使群聚內的個體具有較高的相似度。按照設定的距離公式(通常為歐氏距離)，演算法的最佳化目標為使各個群組內部個體的均方誤差總和最小。上述目標可用以下公式(1)來表達，其中 C 為群聚所成集合，$|C|=k$，u_i 為第 i 個群聚 C_i 之群心，$|x-u_i|$則為個體 x 與群心 u_i 之距離，

$$argmin_C \sum_{i=1}^{k} \sum_{x \epsilon Ci} |x - u_i|^2 \qquad (1)$$

(2) 餘弦相似度（Cosine Similarity）

因此在傳統文件分類中，餘弦相似度經常被拿來度量文件間之相似度。此法建立在空間向量的基礎上，透過兩個同維度向量 A 與 B 之間夾角的餘弦值，來判斷兩者的相似程度(參考以下公式(2))。當兩向量角度相差越小代表兩者越相似，餘弦值越接近 1；反之其值則越接近 0，相似度值介於[0,1]。以文件分類而言，假設在 n 維空間中(有 n 個關

鍵詞)，一份文件 A=(A_1,A_2,…,A_n)與文章類別 B=(B_1,B_2,…,B_n)之相似度便可用公式(2)來計算，其值便做為是否將文章 A 歸入類別 B 之依據。

$$\text{similarity(A, B)} = \frac{A \cdot B}{|A| \cdot |B|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (2)$$

餘弦相似度經常應用在文件探勘上，包括廣為人知的 TF-IDF 權重計算，以算出某個詞(word)在文件中的重要性(權重)。搜尋引擎亦經常使用 TF-IDF 做為搜尋結果排序的方法，計算查詢字串與文件之間的相關性。但為顧及效率，計算時可先透過事先建構的反向索引檔，找尋每個查詢詞所在的文件，最後累積與每篇文件的相似度而成。

三、 新聞讀者情感預測方法

本節將介紹本研究之提出網路新聞讀者情感預測方法，內容包含使用改良式 N-gram 對新聞進行斷詞，並以頻詞序表做為塑模工具，最後再運用三種不同相似度比對方法來進行情感預測。

1. 新聞斷詞與字詞縮減

為建立預測模型，我們先使用改良式 N-gram 概念對新聞進行斷詞，以產生字詞頻率排序表(以下簡稱頻詞序表)。傳統 N-gram 利用改變連續單詞(word)長度 n 的方式來斷詞，應用在中文系統可能會出現問題。許多中文名詞長度大於 1，其子字串出現頻率至少會跟該名詞相同，造成字詞數量大增。例如，一篇文章內容為"歐巴馬、歐巴馬、歐巴馬，河馬出版"，<歐巴馬> 出現 3 次，<河馬出版>出現 1 次，則其 N-gram 斷詞結果將如表 1 所示。由表中可發現，既然<歐巴馬>出現三次，其子字串<歐>、<巴>、<馬>、<歐巴>、<巴馬>出現次數必定大於等於 3。這些代表相同意涵(重複)的字詞，將增加分析時的負擔，甚至影響判斷上之準確性。

表 1: 中文文章之 N-gram 分析，括號()中之數字為出現次數

N=1	N=2	N=3	N=4
歐(3), 巴(3), 馬(4), 河(1), 出(1), 版(1)	歐巴(3), 巴馬(3), 河馬(1), 馬出(1), 出版(1)	歐巴馬(3), 河馬出(1), 馬出版(1)	河馬出版(1)

為縮減字詞數量，斷詞後若某單詞的子字串頻率與該單詞相同，可將其子字串由字詞表中移除，以避免產生過多重複意義之單詞。參考表 2 中的範例，表 2 分別列出以 N-gram 與改良式之 N-gram 之統計結果。在數量上，N-gram 斷詞產生 15 個字詞，而改良式 N-gram 僅有 3 個字詞。值得注意的是，出現 4 次的<馬>被保留下來，原因為其同時出現在<歐巴馬>與<河馬出版>中，並非單一意涵字詞。以下使用實際新聞說明斷詞結果，參考表 3(a)中 2014 年初有關台灣油品食安風暴的新聞，透過改良式 N-gram 斷詞後，產生的頻詞序表如表 3(b)所示。由表中可知，讀者關心的「油」、「頂新」、「檢」、「檢方」等語彙，均屬於出現頻率較高的單詞，可能引發讀者特定的情感反應。

表 2: 縮減 N-gram 斷詞結果

N-gram斷詞	縮減後之N-gram斷詞
歐(3), 巴(3), 馬(4), 河(1), 出(1), 版(1), 歐巴(3), 巴馬(3), 河馬(1), 馬出(1), 出版(1), 歐巴馬(3), 河馬出(1),馬出版(1), 河馬出版(1)	馬(4) 歐巴馬(3) 河馬出版(1)

為避免斷詞後單詞數目過多，本研究根據實際搜集的新聞，只保留出現次數前 100 名的字詞。原因為收集新聞(中文)之平均字數為 963 字(每篇)，利用改進式 N-gram 分解後，產生出現頻率大於 2 的斷詞平均有 96 個。因此，文章完成斷詞後，只保留出現次數前 100 名之單詞(長度由 1~4)，將其結果儲存到資料庫。

在計算一篇新聞之頻詞序表時，經常發現介係詞、連接詞及主詞的出現頻率偏高。即使在不同情緒反應的新聞之中，這些常用介係詞、連接詞及主詞的出現頻率可能相去不遠，造成預測失準。為了避免上述狀況，本研究先對不同情緒分類的新聞進行字頻統計，透過中研院資科所發展之 CKIP 系統[5]對於字詞詞性定義，找出最常出現並且重複的介係詞、連接詞或主詞。在建立預測模型時，先剔除這些字詞，以免影響預測準確率。根據本研究蒐集的新聞資料統計，在不同情緒分類下重覆出現的常見單字統計如表 4 所示。以表 3 的範例而言，出現次數 10 次的「的」便會由表中移除，以免影響後續分析的準確性。

表 3: 利用改良式 N-gram 分解新聞產生頻詞序表

(a) 新聞來源

Date	Category	Emotion	Text
2014-01-28 12:00:00	Society	angry	彰化地檢署偵辦大統長基食品公司混油案從查扣的電腦資料庫意外查出頂新製油近七年來串通下游十多家廠商逃漏營業稅...

(b) 左列新聞之頻詞序表

count	text	Rank
15	稅	1
12	油	2
12	檢	3
11	新	4
10	頂新	5
10	的	6
9	查	7
8	案	8
8	頂新製油	9
8	檢方	10
:	:	:

表 4: 不同情緒下最常出現的介係詞、連接詞及主詞統計

心情模型 之頻詞序表	火大	開心	無聊	難過	感人	害怕
1	的	的	的	的	的	的
2	不	在	不	人	人	人
3	是	是	是	在	有	在
4	有	不	人	不	在	不
5	人	人	在	是	不	是
6	在	中	有	有	是	有
7	會	時	台	時	大	年
8	中	有	會	到	中	他
9	國	大	時	中	時	時
10	:	:	:	:	:	:

2. 建立預測模型

接續上一節的做法，本節將介紹本研究如何建立讀者情感預測模型:

(1) 為每篇新聞建立頻詞序表: 利用斷詞產生的字詞表，並依照字詞出現頻率排序(越大的在前)，再依照時間點、相同類別及心情分類分別儲存。參考表 5(a)之範例，在 12/28 下午時段，社會新聞分類中有多篇產生火大(angry)情感的新聞，在此步驟會分別為這些新聞建立頻詞序表(參考表 5(b))。據此，假設一則新聞 *ne* 之頻詞序表中所有字詞所成的集合為 $TERM(ne)=\{t_1,t_2,\ldots,t_n\}$，且字詞 t_i 在新聞 *ne* 中出現的次數為 $f(t_i,ne)=c_i$，則可將新聞 *ne* 之頻詞序表可用以下之序列來表示:

$$FW(ne) = < (t_1, c_1), (t_2, c_2), \ldots, (t_n, c_n) > \quad (3)$$

對序列中任二個字詞，若 $i<j$，則 $c_i \geq c_j$。換言之，$FW(ne)$中的元素，將依照字詞出現的次數，由大至小遞減排序。

(2) 建立預測模型: 此步驟將整合相同心情類別中的各篇新聞的頻詞序表，產生一綜合頻詞序表，做為該心情分類的預測模型。假設某一心情分類 *e* 中共有所有新聞所成的集合為 $Ne=\{ne_1, ne_2, \ldots, ne_m\}$，$ne_i$ 對應的頻詞序表以 $FW(ne_i)$表示，且令任一字詞 *t* 在新聞 *ne* 中出現的次數為 $f(t,ne)$，則心情類別 *e* 之綜合頻詞序表 FW_e(預測模型)中之所有字詞集合 $TERM_e$ 為:

$$TERM_e = \bigcup_{i=1}^{m} TERM(ne_i) \quad (4)$$

且對於 FW_e 中的任一字詞 $t_{e,j}$，其出現次數為

$$f(t_{e,j}, Ne) = \sum_{i=1}^{m} f(t_{e,j}, ne_i) \quad (5)$$

簡言之，綜合頻詞序表 FW_e 包含該心情下所有新聞頻詞序表之字詞，整合時若各表中有相同字詞，則將其出現次數相加。整合後之單詞頻率排序表，即為該時間點、該新聞及心情之預測模型。以表 5 中之範例為例，假設火大(angry)分類只有 2 篇新聞，則根據上述步驟，可建立如表 6 之「社會」新聞之「火大」情感預測模型。

然而，新聞為每天發生的事件，同樣的人物或事件，可能隨著時間發生變化，甚至在其他人、事、物交互作用下，對讀者大眾的心情造成不同的影響。例如，在捷運殺人事件剛發生時，一開始讀者對於新聞內文中出現的"鄭捷"一詞，可能會感到害怕或火大。但隨著時間與其他事件的影響(例如在鄭捷接受司法審判後)，相關新聞出現鄭捷一詞時，讀者心情投票結果可能會產生相當大的變化。換言之，相同的字詞在不同的時間點，可能會引發不同的情緒反應。因此，在建立模型時，必須考慮到時間因素。在本研究中，每個新聞分類及心情分類的模型皆有時間戳記，模型會隨著時間而有所變動。在預測不同時間點之新聞時，使用不同時間的模型，以因應讀者情緒的變化。

表 5: 整合資料庫中的 N-gram 頻率表

(a) 新聞來源資料庫

Date	Category	Emotion	Text	Report No.
2014-01-28 12:00:00	Society	angry	彰化地檢署偵辦大統長基食品公司混油案從查扣的電腦稅...	1
2014-01-28 12:00:00	Society	angry	中國時報鐘武達唐玉麟綜合報導彰化地檢中區...	2
:	:	:	:	3

(b)整合後之頻率表

Count	Text	Rank	Report No.
15	稅	1	1
12	油	2	1
12	檢	3	1
11	新	4	1
10	頂新	5	1
9	查	7	1
8	案	8	1
8	頂新製油	9	1
8	檢方	10	1
:	:	:	1
6	檢	1	2
6	稅	2	2
5	頂新	3	2
5	關	4	2
:	:	:	2

表 6: 利用各新聞之頻詞序表建立讀者對於新聞情感之預測模型

Date	Category	Emotion	count	text	Rank
2014/1/28	society	angry	21	稅	1
2014/1/28	society	angry	18	檢	2
2014/1/28	society	angry	15	頂新	3
2014/1/28	society	angry	:	:	:

3. 讀者情感預測

在本研究中，讀者情感預測指的是預測讀者對於一篇新發布新聞之情感反應(投票結果)。為此，對於一篇待測新聞(剛發佈的新聞)，我們先將此新聞以 N-gram 斷詞，以產生頻詞序表。之後，再與同時段內中相同分類中之所有情感分類模型進行比對，以預測讀者閱讀後可能的情感反應。根據上述說明，針對一則待測新聞 *news*，若已知之情感類型集合 $E=\{e_1,e_2,\ldots,e_l\}$，任一情感 e 對應之預測模型(綜合頻詞序表)為 M_e，則將 *news* 之頻詞序表(M_{news})與所有情感模型 M_e 進行比對，找出相似度最高之情感，即為預測之讀者閱後情感 *emo*(*news*)，換言之:

$$emo(news)=argmax_{e\in E}Sim(M_{news},M_e) \quad (6)$$

比對時，本研究使用頻詞序表交集大小、歐氏距離以及第二節提到之餘弦相似度等三種相似度計算方法，以預測文章之可能讀者情感反應。以下將逐一說明各種相似度計算在本研究之應用方式，並舉例說明之。

(1) 使用頻詞序表交集大小做為相似度($Sim_{intersect}$)

$Sim_{intersect}$ 以待測新聞與情感模型之頻詞序表交集大小為相似度判斷依據，交集越大表示相似度越高。已知待測新聞 *news* 與情感模型 e 之頻詞序表分別以 *FW(news)* 與 FW_e 來表示(此處 $M_{news}\equiv FW(news)$, $M_e\equiv FW_e$)，則

$$Sim_{intersect}(M_{news},M_e)=|TERM(news)\cap TERM_e| \quad (7)$$

其中，|S|運算表示集合 S 之元素個數，若公式(7)計算結果的值越大(交集元素越多)則代表 *news* 與情緒 e 之相似度越高。參考表 7(a) 之範例文章，若欲對該文章進行讀者的情感預測，則可先將該文章進行 N-gram 斷詞，產生如表 7(b)第一欄之頻詞序表。而後，根據之前建立的情感預測模型(表 7(b)其他欄位)，進行文章單詞匹配，最後產生如表 7(c)之字詞交集個數統計，並根據該表預測此文章發布後之讀者反應可能為「火大」。

表 7: 以 $Sim_{intersect}$ 做為新聞情感預測之相似度

(a) 待預測之新聞(Date: 2014-01-28 12:00:00 新聞分類:社會)

Date	Category	Emotion	Text
2014-01-28 12:00:00	Society	Angry	彰化地檢署偵辦大統長基食品公司混油案從查扣的電腦資料庫意外查出頂新製油近七年來串通下游十多家廠商逃漏營業稅...

(b) 待測新聞之頻詞序表(第一欄)，以及先前建立的情感預測模型(頻詞序表)

待測新聞	**火大模型**	**難過模型**	**開心模型**	**感人模型**	**害怕模型**
稅	油	女童	劉政	阿嬤	女兒
油	案	警方	查	性	次
檢	女童	姓	頂新	少女	鄒雅婷
新	警方	案	稅	所	遭
頂新	檢	死	案	無	綑
查	姓	發現	業	親	打
案	劉政	遭	劉政池	嚴男	性
檢查	公司	妻	辦	男子	庭
頂新製油	嫌	屍	速	法院	檢
檢方	七	路	檢方	大甲	鹽
:	:	:	:	:	:

(c) 對(a)中之文章進行字詞交集個數統計

	火大模型	難過模型	開心模型	感人模型	害怕模型
交集大小	18	11	7	16	10

(2) 利用歐氏距離做為相似度(Sim_{edist})

本方法採用常用之歐氏距離公式，計算待測新聞與情感模型之相似度。已知待測新聞 *news* 與情感模型 *e* 之頻詞序表分別為 $M_{news}=FW(news)$與 $FW_e(=M_e)$，則利用歐氏距離計算 *news* 與 M_e 之不相似度方式如下:

$$Sim_{edist}(M_{news}, M_e)=-\sqrt{\sum_{ti\in p}(r(ti, FW_e)-r(ti, TERM(news))^2+|TERM_e-p|\cdot L} \quad (8)$$

其中 $p=TERM_e\cap TERM(news)$，$r(t_i,FW_e)$表示字詞 t_i 在頻詞序表中之排名，L 為一大整數。參考表 8 之範例，$TERM_{angry}$={油,案,女童,警方,檢,姓,檢方,公司,嫌,七}，$TERM(news)$={稅,油,檢,新,頂新,查,案,檢查,頂新製油,檢}，因此 p={油, 案, 檢}。若使用歐氏距離來計算待測新聞與火大模型(angry)之相似度，並假設 L 為 9999，則

$$Sim_{edist}(M_{news}, M_{angry})=-\sqrt{(1-2)^2+(2-7)^2+9999+9999+(5-3)^2+\cdots}\ 。$$

表 8: 利用歐式距離計算待測新聞與情感模型之相似度

待測新聞頻詞序表 FW_{news}	稅	油	檢	新	頂新	查	案	檢查	頂新製油	檢方
	1	2	3	4	5	6	7	8	9	10
火大模型頻詞序表 FW_{angry}	油	案	女童	警方	檢	姓	檢方	公司	嫌	七
	1	2	3	4	5	6	7	8	9	10

(3) 使用餘弦相似度(Sim_{cos})

本法先將頻詞序表中的單詞視為屬性(attribute)，並將單詞出現次數視為屬性值，便可將其表示為一向量。假設情緒 e 之頻詞序表中各字詞($TERM_e$)出現的次數可用一向量 B=(b_1,b_2,…,b_n)來表示，再查詢 $TERM_e$ 中每個字詞 t_i 在待測新聞 $news$ 之頻詞序表 $FW(news)$ 中出現的次數，產生另一向量 A=(a_1,a_2,…,a_n)，若 t_i 並未出現在 $FW(news)$ 中，則令 a_i=0。根據以上說明，則 $news$ 引發之讀者情緒與 e 之餘弦相似度為

$$Sim_{cos}(M_{news}, M_e) = \text{cosine similarity}(A, B) = \frac{A \cdot B}{|A| \cdot |B|} = \frac{\sum_{i=1}^{n} a_i \times b_i}{\sqrt{\sum_{i=1}^{n}(a_i)^2} \times \sqrt{\sum_{i=1}^{n}(b_i)^2}} \quad (9)$$

上式之計算結果將介於[0,1]之間，越接近 1 則表示二者相似度越高。例如，若 A=(1,2,3), B=(2,4,6)，則 cosine similarity(A,B)=1，亦即此二向量方向相同；若 A=(1,0,3), B=(0,2,0)，則相似度為 0，表示此二向量正交。以表 9 之簡化後之頻詞序表為例，火大模型之 $TERM_e$={油，案，女童，檢，姓，稅}，則上式中的 A=(12,9,0,12,0,15)，B=(52,48,46,41,39,30)，此時新聞之情緒與火大(angry)之餘弦相似度為

$$\frac{(12*52)+(9*48)+(0*46)+(12*41)+(0*39)+(15*30)}{\sqrt{12^2+9^2+0+12^2+0^2+15^2}\cdot\sqrt{52^2+48^2+46^2+41^2+39^2+30^2}} = 0.7737$$

計算結果顯示此新聞之讀者情緒與火大模型之相似度頗高。

表 9: 以餘弦相似度做為新聞情感預測之相似度

待測新聞之頻詞序表: $FW(news)$		火大模型(angry): FW_{angry}	
字詞	出現次數	字詞	出現次數
稅	15	**油**	52
油	12	**案**	48
檢	12	女童	46
新	11	**檢**	41
頂新	10	姓	39
案	9	**稅**	30

4. 情感預測運作流程

綜合上述所述，本研究提出之分析預測流程分為兩大部分(參考圖 2):

(1) 建立預測模型: 利用爬蟲程式(web crawler)收集新聞內文當為資料，將資料進行 N-gram 斷詞，依照讀者投票之心情分類建立情感分類模型。

(2) 情感預測: 先對待測新聞進行斷詞，再與所屬新聞分類的所有情感模型(步驟(1)建立)進行比對，以預測該新聞之讀者可能情緒。然後再與實際的新聞投票結果做比較，測得預測準確率。

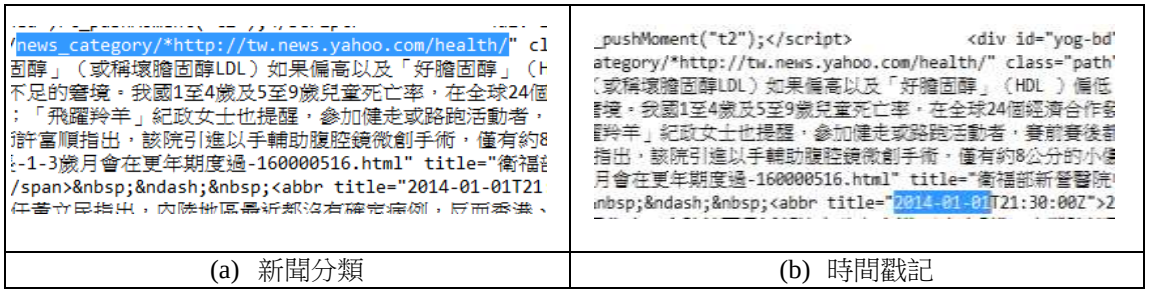
(a) 建立預測模型　　(b) 情感預測

圖 2: 本研究提出之網路新聞讀者情緒分析與預測流程

本研究資料蒐集來源為 Yahoo 奇摩新聞，此網路平台包含了中央社、中時電子報、自由時報、聯合新聞、TVBS 新聞、NowNews 與中廣等不同媒體新聞。Yahoo 新聞平台提供讀者閱讀後進行心情投票，抒發己見。為進行讀者情感預測，本研究收集每日心情新聞分類前一百名新聞，並根據感人(warm)、開心(happy)、無聊(boring)、害怕(worried)、難過(depressing)及火大(angry)等六類心情進行分析。蒐集資料時，本論文利用網路爬蟲程式進入雅虎奇摩心情分類排行榜新聞網頁(https://tw.news.yahoo.com/sentiment/)，並執行以下步驟:

(1) 資料蒐集: 每日早上及晚上六點各一次，利用爬蟲程式至奇摩熱門心情新聞，擷取前一百篇新聞內文，並將新聞內文依照心情分類，放置在不同資料夾。

(2) 分析新聞內文: 擷取每篇新聞時間標記及新聞分類標記(政治、財經、影劇、運動等)，並把所有標點符號及影音串流連結資料刪除(參考圖 3)。同一篇新聞在網路上可能隨著時間而有不同的讀者情緒變化，若擷取之新聞與之前重複時，則以較晚的時間點之新聞為準。



(a) 新聞分類　　(b) 時間戳記

圖 3: 利用本文將新聞分類並擷取時間戳記

(3) 建立資料庫: 將新聞內文儲存到資料庫，同時儲存該內文的投票心情結果、新聞分類及時間點(資料表格式如表 10 所示)。每筆新聞均有其所在時段，我們將每日分為上午與下午時段，各 12 小時。所謂的「同時段」新聞，指的是同一個上午或下午之所有新聞。

表 10: 將新聞內文儲存至資料庫之格式

Date	Category	Emotion	Text
2014-01-01 12:00:00	health	worried	美國研究指出體內壞膽固醇偏高以及好膽固醇低會造成腦部形成斑塊提高...

四、 實驗與評估

為驗證本研究提出方法之有效性，本節將使用 Yahoo 奇摩新聞做為資料來源進行實驗。配合各種實驗設定，比較不同相似度計算方式對情感預測準確度之影響。此外，我們也討論各種準確率影響因素，期望能進一步提升總體預測效能。

1. 實驗設計

本研究所使用之 Yahoo 新聞資料，蒐集期間自 2013 年 12 月 8 日起，至 2014 年 11 月 12 日止，共有 193,489 則新聞。實驗時，使用本研究提出之情感模型，配合三種相似度計算方法，對於情感或新聞分類進行準確率比較。為更清楚呈現實驗結果，除呈顯預測準確率外，亦分別對各種影響準確率因素設計實驗，重點如下:

(1) 塑模時使用之歷史新聞數量多寡，對於預測準確率之影響。
(2) 塑模時使用之歷史新聞擷取時間長短，對於預測準確率之影響。
(3) 如何準確預測重大新聞發生時讀者情感? 重大新聞具有特殊性，與同時段或前幾日之其他新聞具有明顯歧異度，容易造成預測失準。
(4) 某類型新聞在連續時間區間可能產生相同情感模型，若二個模型之字頻表重複字詞出現的數量越多，是否能提升預測準確率?

有關預測準確率的評估方式，乃將本論文之分析結果與 Yahoo 新聞讀者心情投票結果進行比對，若結果一致則視為準確。實務上，如何評估情感分析結果好壞並不容易，同樣一篇文章給多位讀者閱讀，可能產生完全不同的情感反應(例如，有人認為是正面表述，有人認為是反諷文)。因此，所謂「預測結果正確」，只能根據讀者投票來決定。例如，10 位讀者閱讀同一篇 iPhone 6s 之使用者心得，有 7 位認為發文者對此產品為正面評價，3 位認為是負面或中立，則此發文之預測正確答案便是正面(positive)。若自動情感分析結果為「正面」，則被歸為準確。但對其餘 3 位讀者而言，這樣的結果是錯誤的。因此，所謂 100%正確之情感分析結果，可能只有 70%的人同意。事實上，根據 Ogneva(2010)分析 Amazon's Mechanical Turk 中的內容，讀者同閱讀一篇文章之情感一致性僅有 79%。換言之，若自動情感分析準確性能超過 70%，便幾乎和人類讀者之情感辨識力一樣好[25]。

2. 結果與評估

根據上述設定，以下各節將呈現並討論各項實驗結果。此外，亦將探討情感分析在網路新聞之應用下，產生的特殊問題與解決方式。

(一) 預測準確率

本節驗證本研究提出方法之有效性，並比較各種相似度計算方式($Sim_{intersect}$, Sim_{edist}, Sim_{cos})，在新聞讀者情感上之預測準確率。參考表 11 之實驗結果，其中 T_{a-b} 表示與預測新聞同時段(每日分為上午、下午兩個時段)之新聞數量介於[a,b]篇，若同時段沒有任何相同心情新聞，則不對此待測新聞進行預測。由表中數據可知，在各種狀況下 $Sim_{intersect}$ 之準確率最高，顯示以頻詞序表之交集來預測新聞情感最為準確。此外，當與待測同時段之新聞篇數愈多時，預測準確率越高。當篇數由 1-5(T_{1-5})增加至 5-10(T_{5-10})時，三種方法之準確率均有大幅提升，分別為 56.5%(+17.1%), 50.9%(+15.5%)與 43.2%(+11.9%)，其中仍以 $Sim_{intersect}$ 表現最佳。當篇數提升至 25 篇新聞以上時($T_{>25}$)，$Sim_{intersect}$ 之情感預測準確率已高達 73.3%，超過一般情感分析準確度可接受之 70%門檻[25]。上述結果顯示本研究提出以頻詞序表做為情感預測工具之有效性，當配合 $Sim_{intersect}$ 相似度計算方法時，在變化快速的網路新聞應用中，能獲得良好的預測準確性。

表 11: 網路新聞讀者情感預測準確率統計

訓練集大小 / 預測方法	塑模時之歷史新聞篇數					
	T_{1-5}*	T_{5-10}	T_{11-15}	T_{16-20}	T_{21-25}	$T_{>25}$
$Sim_{intersect}$	39.4%	56.5%	65.3%	65.8%	69.1%	**73.3%**
Sim_{edist}	35.4%	50.9%	59.3%	60.3%	62.4%	67.8%
Sim_{cos}	31.3%	43.2%	51.6%	52.3%	53.1%	64.2%

* T_{a-b} 表示與預測新聞同時段之其他新聞數量介於[a,b]篇之間

有關三種相似度計算方法之準確度差異討論如下: $Sim_{intersect}$ 的表現最佳，可能原因為新聞中特定字詞的出現即可主導讀者情緒。例如，在某一時期，當「頂新」出現在一則新聞中時，就算次數很少，便可能性造成讀者之「火大」情緒。因此，可推估網路新聞中字詞出現的頻率非預測之最重要因素，而是字詞引發讀者情緒之強弱程度[1]。據此，便可推論其餘二種方法因將字詞出現次數做為相似度計算主要依據，造成預測效果不如 $Sim_{intersect}$。

(二) 新聞蒐集區間對於預測準確率之影響

由上述結果可知，塑模之新聞數量明顯的影響預測準確率。然而，有時在特定時段內，相同新聞類型與心情分類的新聞確實太少，便會導致其預測準確率無法提升。此時，可考慮延長塑模所需之新聞蒐集區間，例如延長至待測新聞發布前 5 天、前 10 天，甚至前 15 天內，以獲得更多的訓練資料(新聞)。

[1] 有關字詞引發讀者情緒之強弱程度，非本論文之探討範疇，需進行更深入之研究。

為了解新聞蒐集區間對於準確率之影響，我們對不同時間區間進行實驗(只使用 $Sim_{intersect}$ 相似度)，結果如表 12 所示。從表 12(a)可知，當模型建立所使用新聞資料由與待測新聞同一時段，延長至前 5 天的時候，平均預測準確率提升了 9.4%以上(由 49%提升至 58.4%)，延長至前 10 天，準確率只再提升 1.4%。此結果顯示，延長蒐集時間確實可提升預測準確率。但由於新聞有其討論熱度，因此蒐集區間由 5 日延長 10 日，準確率僅有少量提升。當延長到 15 天時，雖然其使用的新聞數量增加，預測準確率不升反降。可能的原因為:雖然建立模型所使用的時間區段延長，蒐集的新聞數量增加，但是相同的新聞字詞，在不同時間可能導致有不同的情緒反應，因此預測準確率反而下降。若再將表 12(a)根據情感分類統計，可獲得如表 12(b)之結果。由表中可知，以前 10 天為新聞蒐集區間仍可獲得最高之預測準確率。「無聊」分類之所以可獲得如此高的準確率，推測原因可能是記者用詞越來越辛辣，改變讀者閱讀習慣，導致許多新聞無法引起讀者興趣，而給予「無聊」評價。此外，由於新聞具有延續性，令民眾「火大」的新聞可能延燒多日，因此亦可獲得相對較高預測準確率。

表 12: 使用不同新聞資料蒐集區間產生情感分類模型所獲之預測準確率統計

(a) 總體預測準確率結果

$Sim_{inersect}$	塑模取用之歷史新聞擷取區間(以待測新聞發布時間為準)			
	同時段	前5天	前10天	前15天
準確率	49.0%	58.4%	59.8%	50.4%

(b) 依照情感分類預測準確率結果

塑模取用之新聞集合 (以待測新聞為基準點)	各種新聞讀者情緒分類之預測準確率(使用$Sim_{intersect}$)					
	火大	無聊	開心	害怕	難過	感人
同時段	58.0%	55.0%	48.7%	45.1%	43.0%	41.0%
前5天	68.9%	74.4%	57.0%	50.0%	41.3%	53.2%
前10天	69.5%	74.3%	59.6%	51.2%	42.5%	55.8%
前15天	58.3%	67.0%	50.00%	42.8%	38.9%	40.0%

(三) 突發之重大新聞對於預測準確率之影響

接下來，將探討重大新聞發生時，對於預測準確率之影響。在正常狀況下，相同新聞分類下的某個心情分類之新聞篇數呈現穩定數量。以圖4為例，在2014/8/9~2014/8/11早上為止，娛樂新聞中少有讓讀者覺得沮喪的新聞(每半天低於3篇)。但在8/12日早上，突然有58篇娛樂新聞讓讀者出現沮喪反應，代表該時間點前可能發生重大新聞。經過查證，原來美國知名喜劇演員羅賓.威廉斯在8/11日自殺身亡，引發新聞界大量報導。若利用8/9-8/11的資料來塑模，將羅賓威廉斯過世的新聞進行情感分類，則無論使用那一種預測方法(參考表13第二列)，只會將該新聞分類為「快樂」或「無聊」，準確率為0%。原因很簡單，因為在8/10日前，並無羅賓威廉斯的新聞，且娛樂類新聞大都引發「快樂」或「無聊」情緒。但若在隔天(8/12)再行塑模，則預測準確率便可大幅提升(參考表13的第三列)。

由上可知，在重大新聞發生當下，若以之前同類型、相同情緒的新聞來塑模，將無法提供可信的預測結果。此時，若使用過長的回溯期間(如表12中之5天或10天)，將使準確率無法迅速提升。因此，面對重大新聞之情感預測，應考慮使用短期或同一發布時間區間內的新聞來塑模。

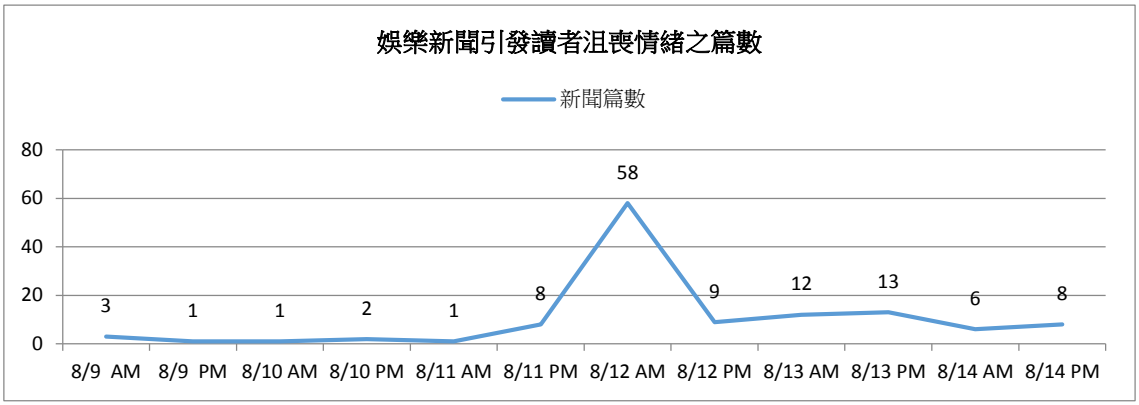


圖 4: 2014/8/9~2014/8/14 娛樂新聞引發讀者沮喪情緒之新聞篇數，8/11 日發生美國知名喜劇演員羅賓威廉斯不幸過世之新聞事件。

表 13: 重大新聞(2014/8/11)發生當時之預測準確率

日期/預測方法	$Sim_{intersect}$	Sim_{edist}	Sim_{cos}
2014/8/11	0%	0%	0%
2014/8/12	70%	68%	50%

(四) 連續模型中重覆字詞對於準確率之影響

每個新聞及心情分類在不同時間點都會產生不同的模型，而在前一個時間點與該時間點所產生的模型，兩個模型中重複的字詞越多，表示兩個模型的相似度越高。這也表示同樣的字詞在前一個時間點與該時間點，都能讓民眾產生類似的情緒反應。因此，可推論當出現的重複字詞越多，預測準確率可能便會提高。參考表14中之範例，顯示2014/3/28與2014/03/29連續二天之「政治」類別之「火大」情感模型，計有97個重複字詞。當考慮各種不同情感模型，統計前一個時間點與該時間點之重複字詞出現的次數，便可獲得如表15之數據。對照表12(b)之「同時段」列，可看出當連續時段之預測模型內的重複字詞數量越多，其預測準確率可能會相對提高。

表 14: 2014 年 3/28~3/29 二天之「政治」類別之「火大」情感模型

日期	政治類火大情緒模型(字頻序表)													
2014/3/29	總統	議	們	學生	台灣	服貿	立	院	政	協	協議	意	主	:
2014/3/28	服貿	議	政	院	黨	立	協	進	江宜樺	協議	學生	表示	意	:

表 15: 重複字詞出現平均數量

火大	無聊	開心	害怕	感人	難過
41.8744	37.13618	32.59152	33.93809	32.42597	31.66951

五、結論與未來工作

Web 2.0的影響方興未艾，各類型網站再也不以仰賴經營者提供內容為主，參與群眾留下的網路足跡才是網站重要的資產。事實上，隨著社群網路的興起，群眾開始習慣在網路上發表意見，並進行評論，因此留下了大量的公開資料。面對這些「大數據」，若能仔細加以分析，便可獲得寶貴的訊息，了解民眾的喜好與需求。本研究以網路新聞讀者情感預測為目標，了解讀者對於刊登新聞之可能反應，做為當局發布新聞、制定決策時之重要參考。本研究長時間大量蒐集網路新聞，使用N-gram技術對於網路新聞進行斷詞，對於常用字詞進行次數統計，並配合讀者的情緒投票，產生新聞與讀者情感之預測模型。對待測新聞進行預測時，本研究亦嘗試各種不同的相似度計算方法，以提升準確率。本研究蒐集Yahoo奇摩新聞網站2013年12月8日至2014年11月12日止共193,489筆新聞進行實驗，結果顯示本研究提出之方法具有良好準確率。其次，透過結果可知，每一個模型中所用到文章數量影響了預測準確率。此外，我們也發現新聞蒐集時間增長時，預測準確率更可獲得明顯提升。但若有重大新聞發生時，延後塑模的時間點可獲得更佳的預測結果。

由於新聞數量影響預測準確率，未來可考慮只針對某些新聞分類及情感分類來進行分析，除能提高文章數量外，亦有助於提升分析結果之針對性。其次，雅虎奇摩新聞上的心情新聞為讀者投票結果，但並不代表大多數讀者閱讀後均產生類似情緒。因此，未來亦可將所有情緒的投票票數納入分析，以獲得讀者的綜合情感，而非單一情感。此外，隨著新聞熱度的消失及新事件的發生，人們對於一個新聞字句的情緒可能會有所轉變，因此情感模型必須要能隨著時間而有所變動。最後，若能將資料蒐集的來源擴展至國內各大入口網站，將可進一步改善預測結果。

參考文獻

[1] 馬偉傑，新聞媒體及網絡行為習慣調查，傳媒透視 4 月號，2012, pp. 12-14。
[2] 《現代漢語頻率辭典》，北京語言文化大學出版社出版，1986 年。
[3] 馬偉雲，未知詞擷取作法， http://ckipsvr.iis.sinica.edu.tw/，最後擷取日期 2016/02/01。
[4] 馬偉雲、謝佑明、楊昌樺、陳克健，中文語料庫構建及管理系統設計，第十四屆計算語言學研討會論文集，2001, pp. 175-191，台南成功大學。
[5] 中研院資訊科學研究所，中文斷詞系統，http://ckipsvr.iis.sinica.edu.tw/，最後存取日期 2015/5/23。
[6] Akay, A., et al., A Novel Data-Mining Approach Leveraging Social Media to Monitor Consumer Opinion of Sitagliptin, IEEE Journal of Biomedical and Health Informatics, Vol. 19, No. 1, Jan. 2015, pp. 389-396.
[7] Bahrainian, S.-A., Dengel A., Sentiment Analysis and Summarization of Twitter Data, IEEE 16th International Conference on Computational Science and Engineering, 2016, pp. 227-234.
[8] Balazs, Jorge A., Juan D. Velsquez , Opinion Mining and Information Fusion: A survey, Information Fusion, Vol 27, Jan. 2016, pp. 95-110.
[9] Cambria, E., et al., New Avenues in Opinion Mining and Sentiment Analysis, IEEE Intelligent Systems, March/April 2013. pp. 15-21.
[10] Cao, Y.-N., et al., Mining Large-scale Event Knowledge from Web Text, Procedia - Computer

Science, Vol. 29, 2014, pp. 478-487.
[11] Cavnar , W. B. and Trenkle , J. M., N-gram-Based Text Categorization, In Proceedings of 3rd Annual Symposium on Document Analysis and Information Retrieval, 1994.
[12] Chen, H., and Zimbra, D., AI and Opinion Mining, IEEE Intelligent Systems, May/June 2010, pp. 74-76.
[13] He, W., Zha S., and Li, L., Social media competitive analysis and text mining: A case study in the pizza industry, International Journal of Information Management, vol. 33, 2013, pp. 464-472.
[14] Huang, W., Zhao, Y., Yang S., Lu, Y., Analysis of the user behavior and opinion classification based on the BBS, Applied Mathematics and Computation, vol. 205, 2008, pp. 668-676.
[15] Isidro, P.-M., et al., "Feature-based Opinion Mining through Ontologies," Expert Systems with Applications, vol. 41, 2014, pp. 5996-6008.
[16] Jiang, L., et al., "BBS Opinion leader Mining based on an improved PageRank Algorithm using MapReduce," 2013 Chinese Automation Congress, pp. 392-396.
[17] Lampos, V., et al., "Predicting the Characterizing User Impact on Twitter," The 14th conference of the European Chapter of the Association for Computational Linguistics, 2014, pp. 405-413.
[18] Lampos, V., Preotiuc-Pietro, D., and Cohn, T., A user-centric model of voting intention from social media, The 51st annual meeting of the association for computational linguistics, 2013, pp. 993-1003.
[19] Lampos, V., and Cristianini, N., Nowcasting Event from the Social Web with Statistical Learning, ACM Trans. On Intelligent Systems and Technology, Vol. 3, No. 4, Sep. 2012, pp. 72:1~72:22.
[20] Lampos, V., Tijl De Bie, Nello Cristianini, Flu Detector - Tracking Epidemics on Twitter, Lecture Notes in Computer Science Volume 6323, 2010, pp. 599-602.
[21] Niesler, T. R. and P. C. Woodland, A variable-length category-based N-gram language model, Computer Speech & Language, Vol. 13 (1), January 1999, 99–124.
[22] Pang, Bo and Lillian Lee, Opinion Mining and Sentiment Analysis, Foundations and Trends in Information Retrieval Vol. 2, No 1-2, 2008 , pp. 1–135.
[23] Parkhe, V., Biswas, B. Aspect based sentiment analysis of Movie Reviews - finding the polarity directing aspects, 2014 International Conference on Soft Computing & Machine Intelligence, pp. 28-32.
[24] Polgreen, P. M., Chen, Y., Pennock, D. M. & Forrest, N. D., Using internet searches for influenza surveillance, Clinical Infectious Diseases 47, 2008, pp. 1443-1448.
[25] Roebuck, K., Sentiment Analysis: High-impact Strategies - What You Need to Know: Definitions, Adoptions, Impact, Benefits, Maturity, Vendors, Emereo Publishing, 2012。
[26] Sobkowicz, P., Kaschesky, Bouchard, G., Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web, Government Information Quarterly, vol. 29, 2012, pp. 470-479.
[27] Urban, J., and Bulkow, K., Tracing public opinion online - An example of use for social network analysis in communication research, Procedia - Social and Behavioral Sciences, vol. 100, 2013, pp. 108-126
[28] Weichselbraun, A., Gindl, S., and Scharl, A., Enriching semantic knowledge based for opinion mining in big data applications, Knowledge-Based Systems, vol. 69, 2014, pp. 78-85.
[29] Weichselbraun, A., Gindl, S., and Scharl, A., Extracting and Grounding Contextualized Sentiment Lexicons, IEEE Intelligent System, March/April, 2013, pp. 39-46.
[30] Witten, I. H., Frank, E., Hall M. A., Data Mining, Practical Machine Learning Tools and Techniques (3/e), Morgan Kaufmann series in data management systems, 2011.
[31] Wu Y., et al., OpinionFlow: Visual Analysis of Opinion Diffusion on Social Media, IEEE trans.

on Visualization and Computer Graphics, Vol. 20, No 12, Dec. 2014, pp. 1763-1772.

[32] Yang, C. C. and Dorbin, T., Analyzing and Visualizing Web Opinion Development and Social Interactions with Density-based Clustering, IEEE trans. on Systems, Man, and Cybernetics-Part A, Vol. 41, No. 6, Nov. 2011, pp. 1144-1155.