

應用 LDA 及 BPN 在長篇電影評論分析

賀信華 沈培輝

國立中興大學科技管理研究所

pfsun@nchu.edu.tw

摘要

本研究針對電影影評進行本文分析，以原本就具有意見分數的電影評論本文，先以 LDA 模型找出劇情內容、演員演技及畫面音效有關的主題字詞，再找出修飾主題字詞的意見字詞，以這些意見字詞針對不同的電影影評計算出影評的主題意見特徵值，再以 BPN 進行演算，得到最後的影評意見特徵值，這個意見特徵值便為該電影影評的意見表象。

關鍵詞：意見字詞、主題字詞、BPN、LDA、電影影評。

1. 緒論

線上口碑(online word of mouth)被認為是對於消費者而言，最具影響力的自然及非正式的交流之一，被定義為消費者對產品的任何陳述，並於網路上向其他消費者提供 [1]，研究指出線上口碑的重要性，特別是消費者產生的產品評論，產品評論對於消費者的購買決策和產品的銷售有顯著影響 [2]，也有研究指出線上口碑的總量以及效量是兩個重要的指標，這可以影響評論的有用性，總量為線上口碑的總體數量，效量為線上口碑的意見表象，不論是正面還是負面的 [3]。

然而線上口碑也已經在電影業中流行起來了，消費者通常會參考過去消費者的意見再進行消費。過去的研究已經研究過線上口碑與電影銷售之間的關係，研究也顯示線上口碑是電影消費者的一個強大的影響因素 [4][5]。但研究也顯示單純只使用線上口碑的總量以及效量來分析對產品銷量的影響，並不能有準確的結果，主要是因為這些聚合的口碑訊息，並無法為消費者足夠的信息來進行產品購買選擇，簡單來說單純只看消費者給予的評價分數，並無法有足夠的資訊讓消費者決定是否購買，他們可能想要審查評論文本 [6][7][8]。然而要是消費者遇到太多積極的評論時，他們可能會懷疑用戶評價的可信度，甚至會覺得可能是公司刻意刪除負面評論或發布正面評論。

關於消費者決策的研究已經確定，消費者認為線上評論比傳統信息來源更具可信度 [9][10]，消費者在面對評論的效量時，儘管用戶評級無法將各種類型的情緒反應為匯總的信息，但當消費者遇到信息分布較為平均的評論本文時，他們會對此用戶評級有較好的可信度。由於用戶對同一部電影具有不同的偏好，例如劇情面向、演技面向，因此均勻分布的評論文本會較具有意義，所以當線上評論在正面、中立以及負面的類型呈現較均勻的分布時，混合的評論文本可以為線上口

碑變量提供更高的可信度。

情緒分析(Sentimental Analysis)早已研究多年，在定義中又可分為廣義及狹義的定義，廣義的情緒分析可稱為意見分析(Opinion Analysis)，而狹義的情緒分析又跟意見分析並不完全相同，情緒分析與意見分析有某些部分相互重疊，但最大的不同是，意見分析多了一樣意見區塊的判斷 [11]。

本研究已經了解到電影評論對電影產業的重要性，其中不同層面的情緒分析能更準確地了解用戶評級，因此本研究希望建立一個模型系統幫助尚未擁有用戶評級的評論，給予一個系統評級分數。

2. 文獻探討

2.1 情緒分析

情緒分析主要有兩個方法用來做情緒辨識，字典法(Lexicon-Based Method)及機器學習法(Machine Learning Approaches)。字典法是利用事先已經具有標示情感訊息的情緒辭典，對未知情緒的字詞以此辭典進行字詞比對，來決定未知情緒字詞的情緒特徵，甚至推斷出句子、段落或文本的情緒特徵 [12]。機器學習法是利用以分類的語料進行訓練，歸納出語料的特徵，再依據特徵來判定學習字詞、句子、段落或文本的情緒特徵。而在這個研究領域上，常見的情緒分析機器學習法大多使用支援向量機(Support Vector Machine，簡稱 SVM)[13]。

2.2 LDA 主題模型

LDA 主題模型是由 Blei、Andrew 及 Jordan 提出 [14]，最初提出利用 EM 演算法設計原始的 LDA 主題模型，之後 Griffiths 及 Steyvers 利用 Gibbs 抽樣來預設 LDA 主題模型中的參數 [15]。

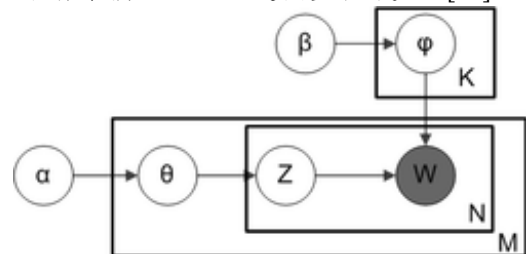


圖 1 LDA 貝氏網絡結構

假設有 M 個文檔，當中出現 N 種不同的字詞，其中可分成 K 個不同的主題，第 i 個文檔第 j 種的字詞詞類及主題類別以 $W_{i,j}$ 、 $Z_{i,j}$ 表示，由於這兩個皆為一個機率形式，因此可以計算出文檔 i 對應主題的機率分布 θ_i 以及主題 $Z_{i,j}$ 對應字詞的機率分布 $\phi_{z_{i,j}}$ ，其中可以用 Dirichlet 計算各別的參數

值 α 、 β ，然而整體的聯合分布為(1)式

$$p(w_i, z_i, \theta_i, \Phi | \alpha, \beta) = \prod_{j=1}^N p(\theta_i | \alpha) p(z_{i,j} | \theta_i) p(\Phi | \beta) p(w_{i,j} | \phi_{z_{i,j}}) \quad (1)$$

其中字詞分布的最大似然估計(Maximum Likelihood Estimation)可以通過將(1)式的 θ_i 及 Φ 積分和對 Z_i 進行求和得到(2)式

$$p(w_i | \alpha, \beta) = \int_{\theta_i} \int_{\Phi} \sum_{z_i} p(w_i, z_i, \theta_i, \Phi | \alpha, \beta) d\Phi d\theta \quad (2)$$

(Blei et al, 2003)[14]。

LDA 貝氏網絡結構中，LDA 將模型參數氏作隨機變量，將 Dirichlet 分布的共軛先驗作為參數的先驗分布，並使用 Gibbs sampling 方法對主題進行採樣。

2.3 BPN

類神經科學的研究顯示，深層的特徵可以從簡單的概念到更高級的複雜模式進行分層次學習[16]。這也推動了深度學習(deep learning)方法的突破，也運用到各個領域，然而倒傳遞類神經網路(Back Propagation Neural Network，簡稱 BPN)是一種結合多層感知機(Multilayer Perceptron，簡稱 MLP)及誤差倒傳遞演算法(Error Back Propagation，簡稱 BP 演算法)的監督式學習類神經網路模型，而訓練類神經網路模型大致分兩個階段，訓練階段跟測試階段，先將一些已知結果的資料當作學習資料來訓練模型，訓練結束後，再以測試資料來測試此模型的準確度。

在 BPN 網絡架構中會存在輸入層、輸出層以及隱藏層，隱藏層可以依學習資料的類型來調整層數，從隱藏層到輸出層皆是由多個神經單元組成，第 n 層第 j 個神經單元會接收 $n-1$ 層的數據資料、路徑權重參數 W 以及該神經元的殘差 θ 來進行分析，然後將結果 A_j^n 傳輸到下一層，如同(3)(4)式

$$net_j^n = \sum_i W_{ij} \cdot A_i^{n-1} - \theta_j \quad (3)$$

$$A_j^n = f(net_j^n) \quad (4)$$

(4)式中，本研究所使用的函數為 H 型函數(Hyperbolic Tangent Function)，如式(5)

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (5)$$

在學習階段，整個架構計算到輸出層時，會將輸出層的結果比對正確的結果，並藉由誤差函數計算其誤差 δ ，再藉由這個誤差倒回進行 W 及 θ 的修正，而修正的公式如式(6)(7)

$$W_{ij} = W_{ij} - \eta \cdot \delta_j \cdot A_i^{n-1} \quad (6)$$

$$\theta_j = \theta_j - \eta \cdot \delta_j \quad (7)$$

其中 η 為設定的學習速度係數，而 δ 會隨著所在的層類型不同而有所不同，若是計算輸出層的 δ 會運用輸出值 Y_j 與正確值 T_j 經由式(8)得到，而計算隱藏層的 δ 則以式(9)

$$\delta_j = (T_j - Y_j)(1 + Y_j)(1 - Y_j) \quad (8)$$

$$\delta_j^n = \left(\sum_k \delta_k^{n+1} \cdot W_{jk} \right) (1 + A_j^n) (1 - A_j^n) \quad (9)$$

每學習一次便會修正一次，值到達收斂條件便停止，再以測試資料進行測試。

3. 資料辭典建立

本研究架構分為兩大部分，資料辭典的建立以及意見分析演算模型，在此以兩個章節分別詳細說明。

詞庫建立的過程中，主題詞以劇情、演技及畫面三個詞為主題種子詞，類聚三個類別的主題詞庫，情感詞主要依據影評的原始評分找出正向情感詞、負向情感詞及中立情感詞。

由於本研究採取的是監督式學習類神經網路模型，資料來源需要已有意見分數的影評文本，因此使用已有影評撰寫者自己評分制度的評論網站(簡稱網站 A)上的影評，透過爬蟲程式抓取影評文本以及意見原始分數，在此一共蒐集了16,271篇影評，其中涵蓋了824部不同的電影。

3.1 影評原始分數分析

網站 A 中一共有劇情面、演技面、視覺面以及總體評分四個分數，三個項目的評分模式為1到5分，以每1分為一個級距，總共5個分數，而總體評分的評分模式為1到5分，以每0.5分為一個級距，總共9個分數。計算四個種類個別的平均分數及標準差，並計算其次數性統計分析。

四個類別的分析結果分別是，總體評分的平均分數為4.277、標準差為0.5658，劇情分的平均分數為4.07、標準差為0.923，演技分的平均分數為3.98、標準差為0.965，視覺分的平均分數為3.98、標準差為0.915，由此可以看出四個類別的平均分數皆位在4分左右，也可推估4分可能為中立意見分數。次數性統計中，有一個值得討論的現象，整體分數分布是極為不平均的，4到5分佔了全部的75%。

3.2 影評文本前處理-jiebaR 的應用

在開始進行分析前須要將這些文本進行前處理，這包含了中文斷詞以及文本斷句。本研究在 R 語言環境下使用 jiebaR 套件進行中文斷詞，並將每個字詞標記上詞性，由於此套件可以標記標點符號以及換行符號，因此再藉由標點符號及換行符號將整篇文章斷成一句一句。

3.3 主題詞資料庫建立-LDA 的應用

由於影評原始分數是三個類別評分外加一個總體評分，因此在後續的意見分析中，也希望以這三個類別為分析基礎，最後計算出一個最終的意見分數，在此之前需要先將這三個類別的主題詞資料庫建立起來。

我們會先將3.2節處理過的所有字詞中挑選名詞出來，接著設定三個類別的種子字詞(劇情、演技、畫面)，再經過多次操作 LDA 來挑選保留的字詞，直到能將挑選的字詞能完整的分到3個組別或者無法再刪除字詞。

每次的 LDA 過程，會先進行 LDA 最適分群數

分析，透過 R 語言的 ldatunung 套件分別計算四個數值，以找出最適分群數的範圍，然後我們選用最最小值為分群數量進行 LDA 模型的建立，模型中會得到一個主題對應字詞的關係矩陣，再透過計算單一字詞對每個主題機率的平均值，再以這個平均值為標準，檢查每個主題對該字詞的數值是否高於此平均值，若是則此字詞便可歸納在此主題中，以此方法將每個字詞做主題歸納，歸納完後，以三個種子字詞為基準，找出包含種子字詞的主題，將這些主題內的字詞保留，其餘的刪除掉。重覆操作直到分到3個組別或者無法再刪除字詞停止，也就完成主題詞資料庫的建立。

於此我們一共操作了10次才完成主題詞資料庫的建立，將字詞由原本的31,498個字詞刪除到3,348個字詞，其中劇情類別的字詞共有1,055個、演技類別有1,385個、畫面類別有1,220個，其中也有部分重疊，但重疊的部分只占全部的9%。

3.4 意見詞資料庫建立

我們希望透過影評的原始分數進行意見詞資料庫建立的依據，透過界定正向、中立及負向的原始分數範圍，搭配主題詞資料庫的輔助，來挑選意見字詞。

以3.2節分詞及斷句後的資料型態，搭配主題詞資料庫找尋每句中是否出現主題字詞，若有出現主題字詞再找尋同句中是否有出現形容詞，以此方法將所有文檔中每一句進行處理，挑選出所有與主題字詞有關聯的形容詞。

由於3.1節中，分析出各項原始分數的平均值大致位在4分左右，因此我們假定三個模式，依序為下：1到3.5分為負面意見分數，4到5分為正面意見分數，沒有中立意見分數；1到3.5分為負面意見分數，4到4.5分為中立意見分數，5分為正面一見分數；1到2.5分為負面意見分數，3到4.5分為中立意見分數，5分為正面意見分數，以此三組分別探討形容詞的分布情形。

在挑選意見字詞中，我們找尋正面意見分數以及負面意見分數內的文檔，將這些文檔內出現與主題詞有關聯的形容詞挑選出來，並歸納至正面及負面，接著觀察兩組是否有重複字詞，將這些重複字詞挑出當作中立意見詞，剩下的個別為正面意見詞及負面意見詞。

三組結果分別為：負面680個，正面5,825個，中立5,766個；負面2,115個，正面2,279個，中立4,331個；負面311個，正面1,574個，中立5,036個。由於第一組及第三組的正負意見詞數量差距過大，因此以第二組為我們的意見詞資料庫，而設定1分到3.5分為負面意見分數，4分到4.5分為中立意見分數，5分為正面意見分數。

4. 意見分析演算模型-BPN 的應用

經過第3章建立完所需的資料辭典，接著需要透過這些辭典進行影評文本的分析，以得到最終的意見特徵值，其中會透過 BPN 的網路模型，經過訓練及測試以達最終的模型。

在此先定義意見分數於模型內的數值，將負面意見分數定義為-1，中立意見分數定義為0，正面意見分數定義為1，也因此採用 H 型函數作為神

經元的運算函數。

4.1 影評主題特徵值

在此先將每篇文檔依照三個主題個別建立兩個項目，分別是正面意見詞數量(number of positive, 簡稱 NP)及負面意見詞數量(number of negative, 簡稱 NN)，所以一共有六個項目。接著將文檔依3.2節斷詞、斷句後，比對每句中出現哪個主題的主題詞，在比對是否出現意見詞，並計算整篇文本再每個項目出現的字詞數，舉例來說，若是該句出現劇情類別的主題字詞，也出現正面意見詞，便在劇情-正面該項目加一，但如出現中立意見詞，便同時在正面及負面的項目加一，以此類推計算整篇最終六個項目的分數。

前一部分得到六個項目的項目分數，但由於下一節中的 BPN 演算模型，其神經單元的函數我們是使用 H 型函數，而這個函數的輸出值為1到-1之間，因此將每個類別的 NP 及 NN 經由式(10)計算成1到-1之間的主題意見特徵值(Opinion Value of Topic, 簡稱 OVT)。

$$OVT_k = \frac{NP_k - NN_k}{NP_k + NN_k} \quad (10)$$

計算出三個 OVT 後，會發現有些 OVT 為零，表示該文本中無法判斷該主體的意見特徵，如果該文本的三個 OVT 皆為零的話，便表示無法擷取該評論的意見特徵，因此會將該文本刪除，而我們一共刪除了8339篇，保留了7932篇。

4.2 模型資料挑選

類神經模型中原則上須要經過訓練階段以及測試階段，因此需要挑選訓練階段所需的資料樣本，然而在3.1節中以分析出整體原始分數的資料樣本分布極度不均勻，4到5分佔了全部的75%，為了不因為這樣的現象影響模型的訓練，所以我們在挑選訓練資料時，以每個中挑選等數量的樣本數來當作訓練資料，其餘的當作測試資料，而我們在每個級距下挑選狀況如下：1分到1.5分從4.1節保留的35筆中挑20筆，2分到2.5分從4.1節保留的154筆中挑80筆，3分到3.5分從4.1節保留的1,375筆中挑100筆，4分到4.5分、5分別從4.1節保留的文本中各挑200筆，這樣形成正面、負面及意見分數各有200筆，以上600筆為訓練資料，其餘的7,332筆為測試資料。

4.3 意見分析模型的建立-BPN 演算法

最終的意見特徵值(Opinion Value, 簡稱 OV)我們藉由 BPN 網路來計算，以每篇的三個 OVT 為輸入值，中間隱藏層設定兩層，經由測試以第一層隱藏層5個神經單元、第二層隱藏層6個神經單元最為適合，最後輸出層只有一個神經單元，而每個神經單元內的運算函數為 H 型函數。

經由 BPN 網路模型計算出來的輸出值，由於神經元的運算函數是使用 H 型函數，所以輸出值會界於1到-1之間，而要怎麼界定哪些範圍是正面意見分數、負面意見分數以及中立意見分數的範圍，在此經由調整範圍，以訓練資料的正確率最高為選取範圍。

5. BPN 分析結果

在此分別討論兩個實驗模型的結果，實驗一：三個意見級別分數共600筆的訓練資料導入訓練階段，再以7,332筆的測試資料進行測試；模型二：去除中立意見分數的資料，只導入正面及負面意見級別分數共400筆的訓練資料到訓練階段，再以其餘的7,532筆測試資料進行測試。

5.1 實驗一

訓練的資料筆數共有600筆，測試的資料筆數為7,332筆。BPN 訓練階段一共疊代了8,330次，誤差值收斂到0.049，而經由測試以數值範圍-1到-0.575為負面意見、-0.575到0.575為中立意見、0.575到1為正面意見，會有最高的正確率72.17%，若是只單看正面、負面意見的資料，以這個範圍的正確率為85%。BPN 測試階段，其正確率雖然只有49.33%，但若是只看正面、負面意見的資料，正確率可以到達78.03%。

表 1 實驗一結果

(筆) 原始 OV	訓練資料			測試資料		
	-1	0	1	-1	0	1
-1	172	54	0	1065	1551	0
0	28	93	32	264	1646	291
1	0	53	168	0	1609	906

5.2 實驗二

訓練的資料筆數共有400筆，測試的資料筆數為7,532筆。BPN 訓練階段一共疊代了304次，誤差值收斂到0.045，由於只用正面意見及負面意見的資料來訓練，因此訓練資料輸出的OV皆為1或-1，在此無法界定三個面向意見的OV範圍，因此採用實驗一的範圍值，其中訓練資料的正確率為100%。BPN 測試階段，包含中立意見資料的結果正確率只有33.72%，但不包含中立意見資料的正確率為100%。

表 2 實驗二結果

(筆) 原始 OV	訓練資料			測試資料		
	-1	0	1	-1	0	1
-1	200		0	1329	2426	0
0				0	14	0
1	0		200	0	2546	1197

由實驗一及實驗二可以看出正面意見及負面意見的影評，分析結果的效果相較於中立意見的影評來的有效，觀察中立意見的分析結果發現，有些中立意見資料的OV值會很極端的逼近1或-1，這可以顯示其實很多影評撰寫者在給予意見分數是過於保守的。

6. 結論及貢獻

本研究除了提出意見分析用於分析長篇文章的系統架構，也驗證期的可行性。主要的貢獻有：(1)LDA 模型可以用於尋找特定字詞的相關字詞。(2)現今通用的情緒辭典或意見辭典並非適用於各種領域，因此建立適用於電影影評的意見辭典。(3)提出不同領域意見辭典建立的系統模型。(4)提出藉由意見字詞進行長篇文章的意見分析模型。(5)此分析模型不單能將長篇文章的意見表象量化，亦可針對即使人為判斷不出來的意見表象(意見表象不明顯)進行意見表象量化，進而降低觀看影評的人不知撰寫的意見推薦程度。

7. 參考文獻

- [1] T. Hennig-Thurau, K.P. Gwinner, G. Walsh, D.D. Gremler, "Electronic word-of-mouth via consumer-opinion platforms: what motivates consumers to articulate themselves on the internet?" *Journal of Interactive Marketing*, 2004, pp. 38-52
- [2] Chen, P.Y., Dhanasobhon, S., & Smith, M.D., "All reviews are not equal: The disaggregate impact of reviews and reviewers at Amazon.com.", 2008, Available at <http://ssrn.com/abstract=978083>.
- [3] S.M. Mudambi, D. Schuff, "What makes a helpful review?. A study of customer reviews on Amazon.com", *MIS Quarterly*, 2010, pp. 185-200
- [4] S. Basuroy, S. Chatterjee, S.A. Ravid, "How critical are critical reviews? The box office effects of film critics, star power, and budgets.", *J. Marketing*, 2003, pp. 103-117
- [5] W. Duan, B. Gu, A.B. Whinston, "Do online reviews matter? An empirical investigation of panel data", *Decis. Support Syst.*, 2008, pp. 1007-1016
- [6] M. Li, L. Huang, C.H. Tan, K. Wei, "Helpfulness of online product reviews as seen by consumers: source and content features.", *Int. J. Electron. Commerce*, 2013, pp. 101-136
- [7] S.G. Moore, "Attitude predictability and helpfulness in online reviews: the role of explained actions and reactions." *J. Consum. Res.*, 2015, pp. 30-44
- [8] Yin, D., Bond, D.S., Zhang, H., "Keep your cool or let it out: nonlinear effects of expressed arousal on perceptions of consumer reviews, Forthcoming at *Journal of Marketing Research*.", 2017
- [9] B. Bickart, R.M. Schindler, "Internet forums as influential sources of consumer information.", *Journal of Marketing*, 2001, pp. 31-40
- [10] C. Li, J. Bernoff, "Groundswell: Winning in a World Transformed by Social Technologies." Harvard Business Press, Boston, MA, 2008
- [11] Kim, S. M. & Hovy, E. "Determining the Sentiment of Opinions.", In *Proceedings of 20th international conference on Computational Linguistics*. Retrieved October 14, 2011
- [12] Murthy Ganapathibhotla & Bing Liu, "Mining Opinions in Comparative Sentences.", *International Conference on Computational Linguistics*, 2008, pp.241-248
- [13] Corinna Cortes, Vladimir Vapnik, "Support-Vector Networks", *Machine Learning*, 1995, pp.273-297
- [14] David M.Blei, Andrew Y. Ng and Michael I. Jordan. "Latent Dirichlet Allocation", *Journal of Machine Learning Research* 3, 2003, pp.993-1022
- [15] T.L. Griffiths, M. Steyvers, "Finding scientific topics", *Proceedings of the national academy of sciences*, 101 (2004), pp. 5228-5235
- [16] I. Arel, D.C. Rose, T.P. Karnowski, "Deep machine learning – a new frontier in artificial intelligence research.", *IEEE Comput. Intell. Mag.*, 5 (2010), pp. 13-18