

## 領域相關詞彙極性分析及文件情緒分類之研究

# Domain Dependent Word Polarity Analysis for Sentiment Classification

游和正\*、黃挺豪\*、陳信希\*

Ho-Cheng Yu, Ting-Hao (Kenneth) Huang, and Hsin-Hsi Chen

### 摘要

情緒分析乃近年來發展迅速之一熱門研究領域，旨在透過文本分析技術探討作者之意見傾向與情緒狀態。其中，以情緒詞與情緒詞典為基礎之各種方法尤為知名。然而，情緒詞之情感傾向及其行為於不同領域文本中之行為並不盡然相同。本研究聚焦於情緒詞彙於不同領域文本中之行為，對房地產、旅館、和餐廳等三種不同領域之文本進行分析，並發現部分情緒詞彙於不同領域文本中的情緒傾向非但有差異，甚至彼此衝突。此外，部分未收錄於情緒詞典中之「非情緒詞」，在特定領域中亦可能成為「領域相依」之詞彙，影響情緒分類。本研究進而提出不同詞彙權重計算方式，將此資訊加入舊有情緒分類系統中。在使用 LIBSVM 的線性核函數方式，對房地產、旅館、和餐廳等三種語料使用 5 次交叉驗證方式進行分類。實驗結果顯示所提出之 TF-S-S-IDF 分類方法，結合 TF-IDF、臺灣大學情感詞典，及計算語料之領域極性情感傾向程度(SO)，強化領域相關及領域不相關之情緒詞之權重，通過 t 檢定有效提升各領域中文文件分類之效能。

**關鍵詞：**文件情緒分類、詞彙極性分析、機器學習

### Abstract

The researches of sentiment analysis aim at exploring the emotional state of writers. The analysis highly depends on the application domains. Analyzing sentiments of

---

\* 國立臺灣大學資訊工程學系 Department of Computer Science and Information Engineering,  
National Taiwan University  
E-mail: {p98922004, r96944003, hhchen}@ntu.edu.tw

the articles in different domains may have different results. In this study, we focus on corpora from three different domains in Traditional and Simplified Chinese including real estate, hotel and restaurant, then examine the polarity degrees of vocabularies in these three domains, and propose methods to capture sentiment differences. Finally, we apply the results to sentiment classification with LIBSVM (linear kernel). The experiments show that the proposed method TF-S-S-IDF which integrates TF-IDF, NTU Sentiment Dictionary, and word sentiment orientation degree in each specific domain can effectively improve the sentiment classification performance.

**Keywords:** Document Sentiment Classification, Word Polarity Analysis, Machine Learning.

## 1. 緒論

隨著資訊科技發展，如今任何人皆可在任何時間、任何地點，透過智慧型手機或電腦輕易連上網路。而網路使用者除由網上獲得所需資訊外，亦可進一步發表意見、評論議題，甚或公開抒發內心的情緒。隨著社群網站興起，此類攜帶主觀意見或情緒之文本資料量大幅提升，情緒分析技術也日益受到重視，並發展出許多重要之應用，諸如產品趨勢分析、關連性商品挖掘、推薦系統、議題偵測等等。

昔日的網站內容提供者僅具有單方向的資訊傳播行為，但隨著 Web 2.0 概念興起，如今已漸由單向轉為雙向的互動傳播，使用者間的相互影響因而益加顯著（高，2012; Tang & Chen, 2011, 2012）。（Pang & Lee, 2008）整理近年之研究而歸納得出下列幾項重要現象：

1. 81%的網路使用者會對有興趣的產品進行線上研究。
2. 20%的人每天都會查看有興趣產品的評論（Review）。
3. 網路上對餐廳、旅館、或其他服務（醫生、旅行業者）的使用者評價，有 73% 到 87%的程度影響使用者是否購買的慾望。
4. 相較於評價為四顆星的商品，消費者更願意多支付 20% - 99%去購買評價為五顆星的商品。

這些現象也說明產品的意見評論對消費者的購買意願有著重大的影響力，讓企業可以發現其商品與服務的優劣，進行改善與強化。

由於情緒分析的結果會隨著不同領域或主題而有不同，在過去相關文獻中，中文方面的多領域之情緒分析相關研究並不多見，大都是對主題式相關之單一領域文章進行研究。本研究嘗試以不同領域之語料進行分析，探討文章領域、詞彙極性、與情緒分類三者之間之關聯，繼而達到下列兩項研究目的：（一）分析不同領域、不同語言文章中詞彙極性之程度，以了解在不同的語料中，詞彙極性變化的情形。（二）研究如何將詞彙極性之變化應用於 SVM 情緒分類器，以提高文件情緒分類之準確度。

## 2. 相關研究

情緒分析由文章、句子、詞彙等三個不同層面切入 (Ku & Chen, 2007; Ku, Huang & Chen, 2009, 2010)，由單語到多語 (Seki *et al.*, 2007)。情緒分類常將輸入文本分為「正面」與「負面」兩個極性 (Turney & Littman, 2003; Turney, 2002; Chaovalit & Zhou, 2005)，或分為特定情緒，如：快樂、難過、溫馨、有趣、驚訝等類別 (Sautera, Eisner, Ekman, & Scott, 2010)。情緒分類也因實際應用上之需求，亦常伴隨情緒強度標註 (Lu, 2010)。諸如一極佳之商品與一尚可之商品，雖皆歸為正面評價之範疇，但強度卻不盡相同。

過去常用之情緒分類方法，主要可以分成非監督式學習法 (Unsupervised Learning) 與監督式學習法 (Supervised Learning) 兩類。Chaovalit and Zhou (2005) 比較這兩類方法，發現監督式學習法具有較高的準確率，但需花費大量時間對標記完成之語料進行訓練；而非監督式學習法的效能則仰賴其所參照的詞性標記程式 (POS tagger)。該實驗結果顯示後者準確率不及監督式學習法，卻具有即時性 (Real-Time) 的優勢。

Tan and Zhang (2008) 則對各種監督式學習法進行比較，發現支援向量機 (SVM) 分類效果較佳。Lan *et al.*, (2005) 提出了 SVM 的 TF-RF 權重計算方式並與其他方法比較，發現該方法具有最佳的準確率。Martineau and Finin (2009) 也提出了 SVM 的 DELTA TF-IDF 權重計算方式，分類效能勝過了 SVM 的 TF-IDF 權重計算方法。

## 3. 語料

### 3.1 資料蒐集及標記

本研究採用三種不同領域的語料，分別為房地產新聞摘要、大陸旅館評論、以及臺灣餐廳評論，這些語料分屬正體與簡體中文，除編碼不同外，亦可探討語言使用上的差異。

#### 3.1.1 房地產新聞摘要

房地產新聞摘要為內政部建築研究所與政治大學臺灣房地產研究中心每季公布之房地產景氣動向季報內新聞摘要 (Chin, 2010)，並由該房地產相關研究人員對新聞摘要進行標記，該語料由 2001 年至 2010 年共 2,389 篇，與其他語料相比，篇幅較少，表 1 列出正面和負面的範例。

表 1. 房地產新聞摘要範例

正面	加速活絡市場，不動產邁向證券化。為加速活絡不動產市場，行政院院會訂今日通過「不動產證券化條例」草案，受益證券買賣均可免徵證交稅；信託財產為土地者，地價稅按基本稅率千分之十計徵；不動產投資信託基金或不動產資產信託持有之土地或建築物，得依地方政府規定減免地價稅及房屋稅等。
負面	再創新高，全台貧富差距飆至 66 倍。過去 11 年來，貧富差距一路狂飆，毫無減緩的趨勢。據財稅資料中心統計，1998 年，最富有的 5% 與最窮的 5% 的平均所得相差 32 倍，11 年後，這項所得差距已擴大至 66 倍，非常驚人。在金融風暴的衝擊下，國內失業率大幅上揚，預估這項所得差距仍會進一步竄升。

### 3.1.2 旅館評論

旅館住宿心得評論為譚松波先生搜集之中文情感挖掘語料之一<sup>1</sup>，以進行學術研究使用，其內容為一般民眾住宿旅館後之心得評論，再由第三方標記者標記其情緒極性。其範例如表 2。

**表 2. 旅館評論範例**

正面	酒店很干净，服务员会推荐我到女士无烟层，设施也比较好，餐厅的小点心味道也可以。吸取了之前美酒店不好停车的经验，在这的三天还是我比较满意的。
負面	这个酒店，隔音太差，都是用木板隔的。有纯净饮用水口，但是坏了,也不配给免费矿泉水。房间设置也是不好的。我的同事房间淋浴口竟然不出水。奉劝大家不要去住。

### 3.1.3 餐廳評論

餐廳評論由愛評網<sup>2</sup>撈取 2006 年至 2009 年之餐廳使用者評論，從其中抽取正負面各 2,000 篇評論進行實驗，其語料文章篇幅較長，表 3 是正面及負面評論之範例。

**表 3. 餐廳評論範例**

正面	無意間看到牛肉麵節的評審，便迫不及待的跑來 白鐵製的煮麵台，感覺就是將一般麵店放在外面的煮麵台，搬進店裡來煮，很有意思 價目表，點麵時就要先決定辣度，那的牛肉麵是用紅燒的 大碗牛肉麵，中辣，135，牛肉給的很多，而且就如王瑞瑤小姐所說：『採用不限部位的本地牛肉，所以吃起來有硬 有軟，富嚼勁又甘甜，』有肉，有助條，也有筋 空心菜有點太爛了 麵條是一般的麵條，但是煮的非常好，咬起來有一股韌性，非常好 接下來喝一口湯，哇，好濃啊，吃過這麼多牛肉麵，這是我喝過牛肉味最濃的湯了，...
負面	台--市----路的--真的很濫，之前吃就有點感覺不好，最近聚餐，朋友剛好又找這家，感覺真的差啦， 8 人吃，點 10 人份，結果感覺端來只有 2-3 人份;再加點 10 人份，結果只來 5 片肉，服務生說吃不夠再點，這擺明有點故意啦.因為已跟不同服務生都反應，我們桌上都空啦;烤肉烤到最後在吃火鍋，因為沒東西烤.我覺得既然開吃到飽，至少不要怕人家點吧，不然就開單點的方式就好啦.花錢是高興的，結果吃的超不高興的.真的差勁到極點.各位網友，台--店都倒光，也不要再去這吃...

<sup>1</sup> [http://www.searchforum.org.cn/tansongbo/senti\\_corpus.jsp](http://www.searchforum.org.cn/tansongbo/senti_corpus.jsp)

<sup>2</sup> <http://www.ipeen.com.tw/>

表 4 整理這三份語料之特性，由不同屬性比較他們之相同和差異處。

表 4. 旅館評論範例

語料	房地產新聞摘要	旅館評論	餐廳評論
文章類型	新聞摘要	個人評論	個人評論
作者背景	記者	一般民眾	一般民眾
標記方式	由房地產專家標記	由語料製作者標記	評論者自行標記
文章內容	房地產	旅館	餐廳
平均文章長度	約 130 字	約 70 字	約 300 字
文字	正體中文	簡體中文	正體中文
文章數	正面：1,418 篇 負面：971 篇	正負各 2,000 篇 共 4000 篇	正負各 2,000 篇 共 4000 篇
期間	2001~2010	未提供	2006~2009

### 3.2 斷詞方法

本研究採用mmseg4j<sup>3</sup>進行斷詞，mmseg4j乃為一基於詞典之斷詞方式。mmseg (Tsai, 1996) 演算法主要區分為簡單 (Simple) 與複雜 (Complex) 兩種方式進行解析，此兩種方式都是使用最大匹配演算法 (maximum matching algorithm) (Chen & Liu, 1992) 進行處理，其簡單的方式準確率達 95%，而複雜方式準確率達 98%。由於mmseg4j之詞典可自行擴充，故本研究除了將臺灣大學情感詞典 (NTUSD) (Ku & Chen, 2007) 導入外，也將語料透過中研院CKIP<sup>4</sup>斷詞系統斷開之詞彙導入，如圖 1。

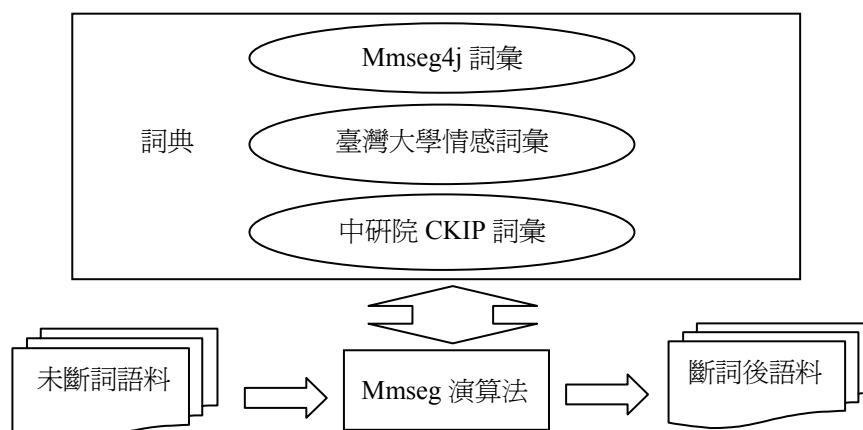


圖 1. 斷詞方法

<sup>3</sup> <http://code.google.com/p/mmseg4j/>

<sup>4</sup> <http://ckipsvr.iis.sinica.edu.tw/>

表 5 是對照擴充前後斷詞結果的比較，擴充後有效地將領域特殊詞彙斷出，例如：地價稅、房屋稅等專有名詞，故比未擴充前詞庫之斷詞效果更佳。

**表 5. 詞庫擴充前後斷詞比較 (差異處以粗體表示)**

原文	為加速活絡不動產市場，行政院院會訂今日通過「不動產證券化條例」草案，受益證券買賣均可免徵證交稅；信託財產為土地者，地價稅按基本稅率千分之十計徵；不動產投資信託基金或不動產資產信託持有之土地或建築物，得依地方政府規定減免地價稅及房屋稅等。
擴充前斷詞結果	[為, 加速, 活絡, 不動產, 市場, 行政院, 院會, 訂, 今日, 通過, 不動產, <b>證券, 化</b> , 條例, 草案, 受益, 證券, 買賣, 均可, 免, 徵, <b>證, 交稅</b> , 信託, 財產, 為, 土地, 者, <b>地價, 稅</b> , 按, 基本, 稅率, 千分之, 十, 計, 徵, 不動產, 投資, 信託, 基金, <b>或不, 動產</b> , 資產, 信託, 持有, 之, 土地, 或, 建築物, <b>得, 依</b> , 地方政府, 規定, 減免, <b>地價, 稅</b> , 及, <b>房屋, 稅</b> , 等]
擴充後斷詞結果	[為, 加速, 活絡, 不動產, 市場, 行政院, 院會, 訂, 今日, 通過, 不動產, <b>證券化</b> , 條例, 草案, 受益, 證券, 買賣, 均可, 免, 徵, <b>證交稅</b> , 信託, 財產, 為, 土地, 者, <b>地價稅</b> , 按, 基本, 稅率, 千分之十, 計徵, 不動產, 投資, 信託, 基金, <b>或不動產</b> , 資產, 信託, 持有, 之, 土地, 或, 建築物, <b>得依</b> , 地方政府, 規定, 減免, <b>地價稅</b> , 及, <b>房屋稅</b> , 等]

## 4. 分析

### 4.1 情緒度程度計算方式

本研究以情緒詞彙於情緒語料庫中之正負面文章頻率 (Document Frequency) 來分辨該詞彙之極性程度，亦即當一詞彙在較多正面文章中出現、在較少負面文章中出現時，此一詞彙便較偏向正面；反之則越偏向反面。利用此特性，吾人可進而計算文章中詞彙之情感傾向程度 (Sentiment Orientation)，並區別文章的正負面極性。其計算方式如公式 (1)、(2)、及(3)所示。

$$PSO = \frac{|D_{w \text{ in } P}| + 1}{|D_P|} \quad (1)$$

$$NSO = \frac{|D_{w \text{ in } N}| + 1}{|D_N|} \quad (2)$$

$$SO = \log_e \left( \frac{PSO}{NSO} \right) \quad (3)$$

PSO 為正面情感傾向程度 (Positive Sentiment Orientation)，是由詞彙所在的正面文章數量  $|D_{w \text{ in } P}|$  除以所有正面文章數量  $|D_P|$ 。為了避免該詞彙在某類文章中未出現，我們將  $|D_{w \text{ in } P}|$  加 1，以避免 PSO 為 0。同理，NSO 為負面情感傾向程度 (Negative Sentiment Orientation)，是由詞彙所在的負面文章數量  $|D_{w \text{ in } N}| + 1$  除以所有負面文章數量  $|D_N|$ 。

得到 PSO 與 NSO 後進行比較，如 PSO 大於 NSO 則情感傾向為正面，反之情感傾向為負面。當  $\frac{PSO}{NSO}$  等於 1 時，代表 PSO = NSO，應為無情緒程度，故  $\log_e(\frac{PSO}{NSO})$  即為 0。本研究將  $\log_e(\frac{PSO}{NSO})$  定義為情感傾向程度 (Sentiment Orientation, SO)，其絕對值代表情感強烈程度，而正負號代表正負向。

## 4.2 詞彙分析

在表 6、表 7、和表 8 中，我們分別由房地產、旅館、和餐廳語料中，各抽取出 SO 最正面及最負面之五個詞彙，進行分析和討論。詞彙後方的「+」符號，代表情感詞典裡之正面詞彙，「-」符號則為情感詞典裡之負面詞彙。

除特定情感詞彙外，還有許多領域相關的特定詞彙，例如：「搶標」、「證券化」、「重劃區」等皆在其中。旅館評論之正負面詞彙較多為情緒詞典內的詞彙，吾人可注意到詞彙「美中不足」似乎是完美、但仍有缺陷的意思，以句子而言，可能有負面的傾向，但若探討整篇文章的情緒傾向，考慮語境時則明顯為一個正面的詞彙。另外，可觀察到餐廳評論之情緒傾向正面之詞彙大都為餐廳名稱或是食品名稱，而情緒辭彙較少，推測應為當一間餐廳或是一項食品，當較常出現於正面文章內，其名稱就會使得文章偏向正面，例如，他做的蛋糕和阿默一樣，而阿默經常出現在正面文章，使得阿默這個辭彙就有讚揚的正面情緒。類似這類型的詞彙是依照語料的特性所賦與詞彙的極性，可以稱作為領域相關的情緒詞彙，依照不同的語料可能會有不同的極性，反觀傾向負面之詞彙則較多情緒詞典詞彙。

**表 6. 房地產正負極性詞彙**

SO 最高之 5 個詞彙				SO 最低之 5 個詞彙			
詞彙	SO	$ D_w \text{ in } P $	$ D_w \text{ in } N $	詞彙	SO	$ D_w \text{ in } P $	$ D_w \text{ in } N $
搶標	2.99	28	0	警訊-	-3.64	0	25
發行+	2.82	25	0	下修	-3.39	2	60
證券化	2.64	43	1	跌幅	-3.37	1	39
新高價	2.54	19	0	不吃不喝	-3.21	0	16
商機+	2.49	18	0	慘	-3.09	0	14

表 7. 旅館評論正負極性詞彙

SO 最高之 5 個詞彙				SO 最低之 5 個詞彙			
詞彙	SO	$ D_{w \text{ in } P} $	$ D_{w \text{ in } N} $	詞彙	SO	$ D_{w \text{ in } P} $	$ D_{w \text{ in } N} $
一流+	3.47	31	0	爛-	-3.99	0	53
美中不足	3.00	19	0	最差-	-3.95	1	103
略顯	3.00	19	0	要我	-3.33	0	27
相當不錯+	2.94	18	0	惡劣-	-3.26	1	51
首選+	2.94	18	0	折騰-	-3.22	0	24

表 8. 餐廳評論正負極性詞彙

SO 最高之 5 個詞彙				SO 最低之 5 個詞彙			
詞彙	SO	$ D_{w \text{ in } P} $	$ D_{w \text{ in } N} $	詞彙	SO	$ D_{w \text{ in } P} $	$ D_{w \text{ in } N} $
饗食	3.22	24	0	不耐煩-	-3.14	0	22
竹筴	3.09	21	0	有待加強-	-3.04	0	20
阿默	3.04	20	0	無言	-2.97	1	38
荷蘭	3.00	19	0	臉色	-2.94	0	18
京兆尹	3.00	19	0	老街	-2.83	0	16

表 9. 極性衝突的詞彙

詞彙	房地產	旅館	餐廳	詞彙	房地產	旅館	餐廳
才能+	-2.32	-1.45	0.13	很快+	-1.48	2.04	0.31
不夠好-	0.00	1.10	-2.40	按照+	0.00	-1.53	2.08
不會有	0.00	-2.20	0.71	恐怕-	-2.58	-0.92	0.16
及時+	-1.48	1.12	1.61	酒鬼-	0.00	-1.10	2.08
少數-	1.76	1.10	1.01	高檔	-1.29	1.10	1.53

進一步對這三種語料之詞彙極性交叉比對後發現，某些詞彙在某一語料傾向正面，但在其他語料卻傾向負面，或是在某語料傾向中性，但在其他語料卻有正負傾向。部分具有衝突的詞彙，如表 9 所示。以表 9 之「恐怕」一詞為例，於房地產新聞及旅館評論中，該詞多為負面意思；然而於餐廳評論中，卻多用於「恐怕得排隊」之意，故多出現於正面文章。由此可知，部分詞彙儘管並無強烈「領域相關性 (domain relatedness)」——如「恐怕」一詞與房地產、旅館或餐廳皆無明顯關聯——卻仍會因領域的不同而產



生不同的正負情緒傾向，即具有「領域相依性 (domain dependency)」。

以下表10 和表 11 以「不夠好」與「遙遙領先」為例，進一步說明此現象。「不夠好」雖為負面的情緒詞彙，但在旅館評論中表現出「盼望未來能夠更好」的正面情緒，故文章多傾向正面；但在餐廳評論中則多有責備之意，故情緒傾向負面。而「遙遙領先」雖為正面詞彙，在餐廳評論裡亦多用於形容環境與美味；但在房地產新聞中，該詞多用來形容房價、油價及其他財經指標高漲，因此多用於負面文章。

表 10. 衝突詞彙範例「不夠好」(以粗體標示)

旅館正面評論	房間很清潔，洗手間裝修好。早餐非常丰盛，豆腐腦很好吃！交通很方便。隔音 <b>不够好</b> ，我们的邻居夜里 1 点多电视声音开的很大，投诉后酒店服务人员去与隔壁邻居协商解决了。
	房間还算干净整洁,服务也可以,以这个价格来说,不错了,建议大家要定有窗户的房間,但是餐廳 <b>不够好</b> ,有一次吃的蘑菇是酸的,附近没有超市,不太方便,买水果都要走很远,
餐廳負面評論	丫勒勒~可能是我好料吃多了! 所以這家就相形失色了吧! 或著是這家分店不比老店好吃! 品管 <b>不够好</b> 嚕~ @0@ 套餐所附的小菜! 我也來介紹一下吧~ 都是一碟有兩種小菜! ...可是是這家分店完全沒有提供! 或是告知是否有這些好康... 真的是弱掉了~~ ...
	用餐時間從 3 點到 5 點 送完餐之後沒有再來加過一次水 也沒有任何人來收吃完後的盤子 價格不美麗，東西不好吃，服務又 <b>不够好</b> 這樣的餐廳，如何讓人再想踏進一步.....

表 11. 衝突詞彙範例「遙遙領先」(以粗體標示)

餐廳正面評論	這雪片冰果真名不虛傳的啊！好吃的程度真的是 <b>遙遙領先</b> ，尤其是咖啡口味的呢！巧克力口味有點讓我心痛就是了，下再來吃要點抹茶紅豆啊！炎炎夏日吃冰果真很透心涼耶 咖啡口味：★★★★★ 巧克力口味：★★★ 炎炎夏日一定要好好享受吃冰的喜悅阿！
	嚴格說起來 到今天造訪野之前 在臺灣還沒吃過令我驚豔的螢烏賊 猶如新鮮荔枝般的滑脆口感 加上肚子裡的爆漿海味 備材選料的水平果然 <b>遙遙領先</b> 能用螢烏賊捏製握壽司 可見深厚功底 玉子 口感豐厚扎實 混入滿滿磨碎白蝦 明顯的海洋鮮香 不輸給築地的專業玉子舖
房地產負面新聞	高學歷失業惡化，創 23 年新高。國內勞動市場出現警訊！今年八月失業率為四·一四%，創下三年來新高；高學歷、高失業率問題更加速惡化，八月大學以上失業率五·二六%，不僅 <b>遙遙領先</b> 其他教育程度，更創下二十三年來新高
	2010 年第 1 季全台新屋成交價突破歷史新高，北市以每坪 56.6 萬元 <b>遙遙領先</b> 其他城市，但銷售率小幅萎縮，議價空間拉大.17.16%。學者擔憂，房價反彈無基本面支撐，且市場觀望氣氛逐漸形成，V 型反彈極可能會轉成 M 型修正，最快今年下半年到明年就會看到。

## 5. 研究方法

### 5.1 實驗設定

透過前述的研究，本研究提出以下三種不同類型之的特徵值進行分類：一，使用單一詞彙（Unigram）。二，使用臺灣大學情感辭典（NTUSD）之詞彙資訊。三，使用二元詞彙（Bigram）。

在以下實驗中，我們將比較這三種方法的優劣。而特徵值則兼採詞頻（Term Frequency）與情感傾向程度（SO），詳細計算方式如表12。表12 第一直行為方法代碼，第二行是計算方式，第三行是公式中所使用的符號說明。

表12. 權重計算方式

方法	計算方式	說明
TF-SO	$TF_{SO} = TF \times ( SO  + 1)$	TF：詞彙在該文章出現的數量除以該文章的詞彙總數； SO ：情感強烈程度 <sup>5</sup> ，數值越大，代表情感程度越強烈時，也越為重要。在以SVM分類時，當 SO 為0，則維持TF之權重，因此將 SO +1。
TF-SO-IDF	$TF_{SO-IDF} = TF_{IDF} \times ( SO  + 1)$	這是一種結合 SO 與 TF-IDF 的方法，使得 IDF 與 SO 相互影響。若該詞彙情緒程度越強，則 TF-IDF 越重要，實驗中並與領域無關之情感詞典比較。
TF-SD-IDF	$TF_{SD-IDF} = \begin{cases} TF_{IDF} \times k & \text{if } w \text{ in NTUSD} \\ TF_{IDF} & \text{otherwise} \end{cases}$	結合 TF-IDF 和情感辭典，若該詞彙在情感詞典裡出現，代表越重要，則進行加權(乘上 k)。由於搭配情感辭典，因此在 Bigram 分類並不使用。
TF-S-S-IDF	$TF_{S-S-IDF} = \begin{cases} TF_{SO-IDF} \times k & \text{if } w \text{ in NTUSD} \\ TF_{SO-IDF} & \text{otherwise} \end{cases}$	結合 TF-SO-IDF 和情感辭典，若該詞彙在情感詞典裡出現，代表越重要，則進行加權(乘上 k)。由於搭配情感辭典，因此在 Bigram 分類並不使用。

TF-SD-IDF及TF-SS-IDF皆須搭配加權之k值，而k值之決定，本研究以房地產評論，採用不同的k值來進行分類，並觀察最後的準確率以決定k值，分類的結果如表13。

表13. 房地產語料不同k值分類結果

k	1.0	1.5	2.0	3.0	5.0	10.0
Accuracy	0.848	0.851	<b>0.852</b>	0.840	0.816	0.7972

<sup>5</sup> 線性 SVM 分類方式在尋找參數過程會自動將正負號進行調整，故取絕對值差異影響不大，這裡是以情緒強烈程度來對應 TF-IDF，故取絕對值較容易理解。

由上表可以發現，k 值在 2.0 時所得到的準確率最佳，而在後續的 3.0 與 5.0 及 10.0 皆不斷遞減，故本研究後續之 k 值皆以 k = 2.0 來進行分類。

實驗結果並與 TF-IDF、TF-RF (Lan, Sung, Low, & Tan, 2005)、及 DELTA TF-IDF (Martineau & Finin, 2009) 等三種方式進行比較，這三種方法的說明如表 14。

表 14. 基準方法權重計算方式

基準方法	計算方式	說明
TF-IDF	$TFIDF = TF \times \log_e\left(\frac{ D }{DF}\right)$	TF：為詞彙在該文章出現的數量除以該文章的詞彙總數； D ：文章的總數；DF：含有該詞彙的文章數量。
TF-RF	$TFRF = C_{t,d} \times \log_e\left(\frac{ D_{w \text{ in } P} }{ D_{w \text{ in } N} } + e\right)$	$C_{t,d}$ 為詞彙在該文章出現的數量， $ D_{w \text{ in } P} $ 為詞彙所在的正面文章數量， $ D_{w \text{ in } N} $ 為詞彙所在的負面文章數量，兩者相除後，加上 e。原文以 2 為底，故加 2，但本研究之 log 皆以 e 為底，故修改成加 e。
Delta TF-IDF	$Delta\ TFIDF = C_{t,d} * \log_e\left(\frac{ D_P  \times  D_{w \text{ in } N} }{ D_{w \text{ in } P}  \times  D_N }\right)$	$C_{t,d}$ 為詞彙在該文章出現的數量， $ D_{w \text{ in } P} $ 為詞彙所在的正面文章數量， $ D_{w \text{ in } N} $ 為詞彙所在的負面文章數量， $ D_P $ 為正面文章數量， $ D_N $ 為負面文章數量。

透過上述方法，計算出權重後，再使用 LIBSVM<sup>6</sup> 的線性核函數 (Linear Kernel Function) 方式進行分類。由於線性方式僅需單一參數 (Cost)，不但速度優於其他核函數，文件分類效能也不會輸於其他核函數 (Lan, Sung, Low, & Tan, 2005)，故選擇此一方式。本研究將依序嘗試不同的參數 (Cost)，以最佳的結果為主。

實驗中將三種語料使用 5 次交叉驗證 (5-Fold Cross Validation) 方式進行，即計算權重時(包括上述的所有權重值)僅使用訓練資料集來計算，以避免內部測試 (Inside Test) 的問題。同時，本研究使用準確度 (Accuracy) 以評估效能，並 t 檢定進行有效性檢定。

## 6. 實驗結果

### 6.1 單一詞彙 (Unigram) 分類結果

表 15 是採用單一詞彙 T 檢定分類結果，評估的標準是準確率，對 TF-S-S-IDF 與其他方式進行比較，若顯著程度達到 95% 的信心水準則標記為 \*，若達到 99% 則標記為 \*\*，而達到 99.5% 則標記為 \*\*\*。吾人發現僅使用單一之 TF-SO 或 TF-IDF 的效果很接近，但是將 IDF

<sup>6</sup> <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>

與SO相乘，也就是TF-SO-IDF，其效果更好，TF-SO-IDF優於其他兩種。若結合情感辭典，可將分類效果更進一步提昇。表中呈現TF-S-S-IDF優於TF-SO-IDF，TF-SD-IDF優於TF-IDF。總結，Unigram的結果以TF-S-S-IDF為最佳，TF-SO-IDF與TF-SD-IDF次之，接著是TF-IDF，與其他方法。這種情況以餐廳的語料最為顯著。

**表 15. Unigram 分類 T 檢定綜合比較表(Accuracy)**

語料	TF-IDF	TF-RF	Delta	TF-SO	<b>TF-SO-IDF</b>	TF-SD-IDF	<b>TF-S-S-IDF</b>
房地產	0.848*	0.849	0.853	0.847*	0.854	0.852	<b>0.863</b>
旅館	0.916	0.906***	0.914*	0.915	<b>0.924</b>	0.918	0.923
餐廳	0.861**	0.839***	0.849***	0.854***	0.871	0.869	<b>0.875</b>

## 6.2 僅用情緒詞典詞彙分類結果

表16 顯示僅使用情緒詞彙來進行分類，同樣的，也對TF-SO-IDF與其他方法進行T檢定，標記方式同表15。與前一方式分類結果比較，各方法之準確度皆顯著下降。這樣的結果顯示，不在情緒辭典內的領域相關情緒辭彙也在分類器中扮演著重要的角色，因此造成此分類結果效能顯著下降。另外，可以觀察到僅用情緒詞的結果並不一致，在房地產分類中，以Delta最佳，在旅館分類中則是以TF-SO-IDF最佳，而在餐廳的評論中則是以TF-IDF最佳。

**表 16. Unigram 僅用情緒詞分類綜合比較表(Accuracy)**

方法	TF-IDF	TF-RF	Delta	TF-SO	<b>TF-SO-IDF</b>
房地產	0.784	0.786	<b>0.796</b>	0.784	0.787
旅館	0.883	0.877*	0.878	0.880	<b>0.887</b>
餐廳	<b>0.806</b>	0.795	0.792	0.783	0.804

## 6.3 二元詞彙 (Bigram) 分類結果

表17 是使用二元詞彙 (Bigram) 的分類結果，同樣的，也對TF-SO-IDF與其他方法進行T檢定，標記方式如前所述，與僅用情緒詞彙分類相比，除餐廳語料外，其他語料的分類效能又下降些許。而在房地產是以TF-IDF最佳，其他為TF-SO-IDF。

**表 17. Bigram 分類綜合比較表(Accuracy)**

語料	TF-IDF	TF-RF	Delta	TF-SO	<b>TF-SO-IDF</b>
房地產	<b>0.777</b>	0.769	0.756	0.767	0.767
旅館	0.879	0.858***	0.879	0.883	<b>0.886</b>
餐廳	0.805	0.778***	0.766***	0.793	<b>0.811</b>

## 7. 結論

本研究以文章頻率來計算詞彙極性程度，探討了各領域詞彙的差異，應用詞彙極性程度結合情緒辭典進行情緒分類，得到更佳的分類效果。從實驗過程的分析與討論，可以歸納以下幾點結論：

1. 情緒詞彙可以歸類為兩類，即「領域不相關」之詞彙與「領域相關」之詞彙。「領域不相關」之詞彙主要為語言學上帶有情感之詞彙，大都為情緒詞典內之詞彙。而「領域相關」之詞彙則是透過語料訓練的方式取得，依照各領域不同而有不同。例如，京兆伊、鼎泰豐、捷運、重劃區等詞彙，原本並未帶有情緒程度，但從訓練的語料得出為正面之情緒詞彙，即屬於此類。這類「領域相關」之情緒詞彙，若用於其他領域則可能不成立，即「鼎泰豐」若用於房地產及旅館則非此類情緒詞彙。  
「領域相關」之情緒詞彙在具有對立關係之事物尤為顯著，舉凡在宗教上或在政治上甚至是社會上，都可見到，舉例而言：某政治人物之姓名，若出現在其政治立場相左之報紙社論之語料內，則該姓名雖非為情緒詞，但是極有可能成為本研究之「領域相關情緒詞彙」。這類由原本的「非情緒詞彙」，轉變成「領域相關情緒詞彙」，在未來甚至也有可能成為各領域共通之情緒詞，這類的轉換，本研究稱之為「詞彙情感注入」現象，例如：三國時「諸葛亮」與「阿斗」原本也僅僅是人物名稱，但透過三國歷史的語料不斷習得，皆使得「諸葛亮」與「阿斗」成為情緒詞彙。
2. 不同領域間詞彙情緒程度可能會有衝突，可歸因不同語料所用之匹配詞彙不同，造成差異。例如，「遙遙領先」在房地產語料內與失業率及高房價搭配，造成情緒極性相反，故此類詞彙雖在情緒詞典，但其仍可歸類為領域相關之情緒詞彙。
3. 若比較「Unigram 所有詞彙」與「Unigram 僅用情緒詞彙」之結果，發現語料內若領域相關情緒詞彙較多時（房地產與餐廳語料），僅用情緒詞典之分類效果較差。反之，若語料內之領域相關情緒詞彙較少時（旅館語料），僅用情緒詞典之分類效果較好。但橫向比較「Unigram 所有詞彙」之 TF-IDF 與 TF-SD-IDF（TF-IDF 情緒辭典加權）以及 TF-SO-IDF 與 TF-S-S-IDF（TF-SO-IDF 情緒辭典加權）卻可以發現，情緒辭典作用在房地產與餐廳語料的效果更佳。
4. 透過詞彙 SO 的計算，找出領域相關之情緒詞彙之極性程度，加入 TF-IDF，再搭配 SVM 分類方法，與僅用 IDF 或僅用 SO 相比之下，兩者相互的結合，更可以提昇分類效能。
5. 本研究提出之 TF-S-S-IDF 分類方法，結合 TF-IDF、臺灣大學情感辭典，及計算語料之領域極性 SO，強化領域相關及領域不相關之情緒詞之權重，得出更佳的分類效能。
6. 中文在預處理上的斷詞結果不同，也將會影響詞彙所內含的意義，能斷出越長的詞彙，並不代表斷詞效果越佳。因為較長的詞彙出現的頻率越少，造成資訊的不全。例如，「房間的設施」與「房間設施」，若皆斷成一個詞彙則會造成不同。但是相對的，若是所斷出的詞彙越短（單一字），則會造成斷出的詞彙僅具字面意義而造成誤判。

## 致謝

Research of this paper was partially supported by National Science Council (Taiwan) under the contract NSC 98-2221-E-002-175-MY3 and 2012 Google Research Award.

## 參考文獻

- Chaovalit, P. & Zhou, L. (2005). Movie review mining: a comparison between supervised and unsupervised. In *Proceedings of the 38th Hawaii International Conference on System Sciences*.
- Chin, Y.-L. (2010) A review and discussion of real estate cycle indicators analysis and publication method. Research Project Report. Architecture and Building Research Institute, Ministry of the Interior, Taiwan. <http://www.abri.gov.tw/>
- Ku, L.-W. & Chen, H.-H. (2007). Mining opinions from the web: beyond relevance retrieval. *Journal of American Society for Information Science and Technology*, 58(12), 1838-850.
- Ku, L.-W., Huang, T.-H., & Chen, H.-H. (2009). Using Morphological and Syntactic Structures for Chinese Opinion Analysis. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 1260-269.
- Ku, L.-W., Huang, T.-H., & Chen, H.-H. (2010). Construction of Chinese Opinion Treebank. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, 1315-319.
- Lan, M., Sung, S.-Y., Low, H.-B., & Tan, C.-L. (2005). A comparative study on term weighting schemes for text categorization. In *Proceedings of International Joint Conference on Neural Networks*, 1032-033.
- Lu, Y., Kong, X., Quan, X., Liu, W., & Xu, Y. (2010). Exploring the sentiment strength of user reviews. In *Proceedings of 11th International Conference on Web-Age Information Management*, 471-482.
- Martineau, J. & Finin, T. (2009) Delta TFIDF: An improved feature space for sentiment analysis. In *Proceedings of the Third AAAI International Conference on Weblogs and Social Media*, 258-261.
- Pang, Bo & Lee, Lillian (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2) 1-135
- Sautera, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. In *Proceedings of the National Academy of Sciences*, 107(6), 2010.
- Tan, S. & Zhang, J. (2008). An empirical study of sentiment analysis for Chinese documents. *Expert System with Applications*, 34(4), 2622-2629.
- Tang, Y.-J. & Chen, H.-H. (2011). Emotion modeling from writer/reader perspectives using a Microblog dataset. In *Proceedings of IJCNLP Workshop on Sentiment Analysis where AI Meets Psychology*, 11-19.

- Tang, Y.-J. & Chen, H.-H. (2012). Mining sentiment words from microblogs for predicting writer-reader emotion transition. In *Proceedings of 8th International Conference on Language Resources and Evaluation*, 1226-1229.
- Tsai, C.-H. (1996). MMSEG: A word identification system for Mandarin Chinese text based on two variants of the maximum matching algorithm. Available at <http://technology.chtsai.org/mmseg/>.
- Turney, P. D. (2002). Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 417-424.
- Turney, P. D. & Littman, M. L. (2003). Measuring praise and criticism: inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21, 315-346.
- Seki, Y., Evans, D.K., Ku, L.-W., Chen, H.-H., Kando, N., & Lin, C.-Y. (2007). Overview of Opinion Analysis Pilot Task at NTCIR-6. In *Proceedings of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies*, 265-278.
- 高宏宇 (2012)。文字探勘於社群網路研究之發揮。NCP Newsletter, 38, 11-16.

