Plurk 的情緒分析

施怡如
Department of
Information management
National Taiwan University
R99725003
R99725003@ntu.edu.tw

陳瀅如
Department of
Information management
National Taiwan University
R99725015
R99725015@ntu.edu.tw

郭岱茵
Department of
Information management
National Taiwan University
R99725043
R99725043@ntu.edu.tw

ABSTRACT
在本研究中，我們實作了一個結合分類與情緒分析的 IR 系統，並且評估這個系統對於情緒分析的精確度。

General Terms
Algorithm, Information retrieval.

Keywords
Plurk, Classification, Opinion mining.

1. INTRODUCTION
一句話、一百四十個字的力量有多大？答案是，它讓環法自由車手阿姆斯壯(Lance Armstrong)在三天半內找回價值新台幣三十三萬元的單車；讓星巴克在一百天內多了十四萬跟隨者；讓美國總統歐巴馬(Barack Obama)，在競選期間就透過它吸引了三十萬粉絲，在網友們的支持下，他也比對手多了一億兩千兩百萬美元的小額政治獻金。這是你不能忽略的新科技；微網誌。

微網誌從三年前於美國誕生的 Twitter 開始盛行，有別於一般部落格長篇圖文型式，它強調快速即時、有字數限定，短短的一句話也能引發網友互動討論，讓使用者可以拓展交友圈、快速交流資訊與分享心情，國內外最紅的微網誌還包括 Facebook、Plurk 等。

在台灣，一般使用者在尋找與電影、美食等相關的訊息時通常會由 PTT 上其他人的反應來評估一部電影或一家餐廳的好壞，可能會有資訊過量或可信度不高的問題。然而，若是透過 Plurk 或 Facebook 取得相關資訊，由於大部分的好友名單不是自己的親人就是認識的朋友，對該部電影或該美食餐廳的評價可信度就提高了。

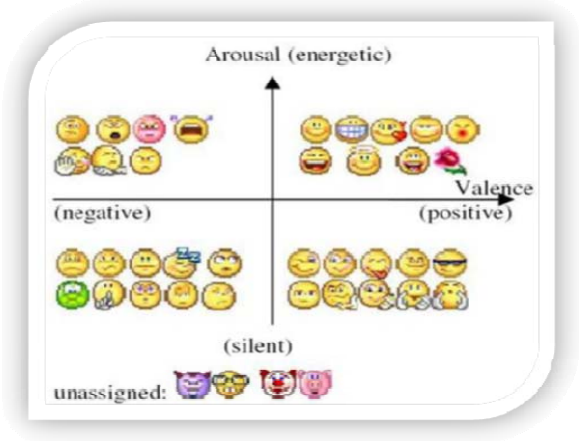
因此，此次研究針對 Plurk 上與電影及美食有關的訊息，做 opinion mining，找出正面及負面的的評價。本系統透過資訊檢索的技術，將這些訊息分類，並標記正面與負面，呈現給使用者。

我們的系統使用 Plurk 當成我們的資料集，我們蒐集 Plurk 上出現的中文訊息，先將這些中文訊息經過斷詞之後，利用關鍵字將這些訊息分成三個類別，美食、電影及心情，而 Plurk 使用者可點選這些類別，看到過往朋友談論過的內容，譬如當使用者想看一部好看的電影卻不知該看哪片時，他就可以進入“電影”此類別裡，看別人過去曾發表過的相關意見，來做決定；除此之外，我們還會進一步對這些訊息做 opinion mining，利用 SVM，我們計算這些訊息的分數，依據這些分數將訊息區分成正面評論與負面評論，讓使用者挑出他想看的訊息內容。

2. Related work
楊鼎等人於 2010 年[4]提到，中文不像英文一樣每個字之間都有空格分開，需要進一步的斷詞才能做資料檢索的應用。利用斷詞字典，將中文以詞語的方式分開，再擷取出具有意義的字，比如：喜歡、天氣、電腦等等，利用擷取出的字進行分類或是情緒分析。

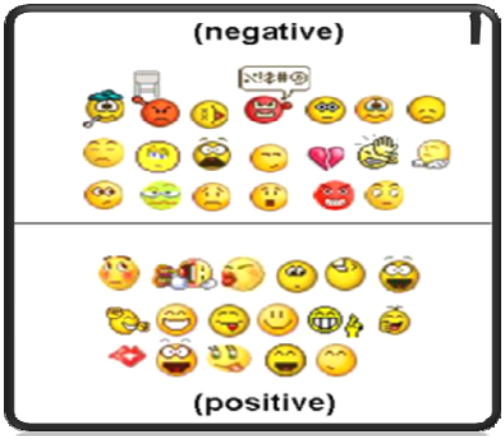
2.1 表情符號
在進行情緒分類或辨識之前，首要任務就是要確定甚麼是情緒以及情緒的類別要如何界定(周嵩能,2009)[2]。Thayer 在 1989 年[8]提出一個基本的二維空間情緒模型，它由壓力(stress)以及能量(energy)兩個維度來組成，並把情緒分成四類，壓力是指情緒的正向(positive)/負向(negative)的程度，能量則是指充滿活力(energetic)/寂靜(silent)的程度。而 Yang 等人 在 2007 年[9]基於 Thayer 所提出的心理學模型，將表情符號映射至正面(positive)、負面(negative)、激動(energetic)和安靜(silent)四個座標軸所組成的四個象限，部分的表情符號與

四個象限沒有任何關聯性分類為 unassigned，如圖一所示。



圖一 Yang 所提出的表情符號分類

而在情感分析(sentiment analysis)研究中的 polarity 分類器則是只將案例(instance)分成正面類別和負面類別兩種類別。所以，本研究不考慮 energetic-silent 軸，只將表情符號簡化為正面(positive)和負面(negative)兩大情緒類別，如圖二所示。



圖二 正負向的表情符號分類

2.2 分類方法

近年的研究，針對情緒分類或辨識的問題， 一些 Blog 的資料提供了 training data 和 testing data(楊昌樺,2009)[3]，Mishne(2005)[7] 首先利用了 Blog 資料，利用 Machine Learning 的方式來預測發表這些 Blog 的作者的心情是哪一種。另外，Bing Liu 等人(2005)[5]提到，supervised pattern discovery 讓較短的訊息比較長的訊息能夠更準確地被分類。而 Yang 與 Lin[10]、Joachims[6]特別以統計考驗的方法，去比較五種分類方法：LLSF(Linear Least Square Fit)、K-Nearest Neighbors(KNN) 、 Support Vector Machines(SVM) 、 Naïve Bayes(NB)、Neural Network(NNet)。實驗結果發現：{SVM，KNN} >> LLSF > NNet >> NB，可看出在這五種方法中，SVM 和 KNN 幾乎差不多，都証明了 SVM 及 KNN 都比其它的方法來得好，於是本研究採用 SVM 這個分類方法來做情緒分析。

3. Methodology

我們系統的架構，分成資料前處理、建立情緒辭典以及機器學習情緒辨識方法三個部分，分成三節討論。另外，在資料前處理的部分之前，我們會介紹語料的來源，也就是 Plurk 的內容。

3.1 Plurk Content

Plurk 是一個免費的社交網站，它允許每一個人擁有類似自己部落格的文字輸入介面，可以在任何時間點留下自己的心情狀態和當下的想法。

Plurk 限制每則文字的訊息長度在 140 個字內，有別於一般部落格的長篇文章。使用者所發的訊息也以短句或甚至幾個簡單的字詞所組成。另外，Plurk 的使用者編輯介面有一個特別的設計，當使用者要發出訊息時，有一組表情符號提供大家選擇，表情符號以圖片的形式呈現，不同的表情符號具有不同的情緒和狀態意涵。使用者加註表情符號除了讓訊息更加生動活潑外，也意味著使用者對於自己所發表的文字透過表情符號標記使用者當下的情緒狀態，或者單純透過表情符號顯示目前狀態。

我們利用 Plurk API 來抓取 Plurk 上面的訊息，總共蒐集了 1000 則的 Plurk 訊息，而其中 200 筆是 training data(已分好情緒的正負向)。

3.2 Pre-Processing

搜集出 Plurk 上所出現的中文訊息，利用關鍵字將在 Plurk 上搜集到的中文訊息篩選出電影、美食及心情三個類別，並進行斷詞。

我們先利用中央研究院(CKIP)小組所開發的中文斷詞系統進行斷詞，中研院的斷詞系統是利用詞典中收錄的詞和我們所輸入的中文訊息做比對，找出可能包含的詞。斷詞系統會將我們所輸入的中文訊息，利用他們所定義的辭典並將斷詞結果顯示，除此之外，並過濾掉不必要的 stop words 如「的」「了」、「吼」等等。

在表情符號的部分，我們利用 Plurk 所提供的 77 種表情符號(如圖三所示)，除此之外，我們還整理了一些 Plurk 未提供，但是卻是我們常用的表情符號，並將其區分為正負向(如圖四所示)。

ID	Emotion	Code	ID	Emotion	Code	ID	Emotion	Code
1		:-))	27		(tongue)	53		(evil_grin)
2		:-)	28		(unsure)	54		(headspin)
3		:-D	29		(highfive)	55		(heart_beat)
4		(LOL)	30		(dance)	56		(ninja)
5		:-P	31		(blush)	57		(haha)
6		(woot)	32		(doh)	58		(evilsmirk)
7		;-)	33		(brokenheart)	59		(eyeroll)
8		:-o	34		(drinking)	60		(muhaha)
9		X-(	35		(girlkiss)	61		(taser)
10		:-(	36		(rofl)	62		(banana_ninja)
11		:'-(	37		(money)	63		(beer)
12		:-&	38		(rock)	64		(coffee)
13		(K)	39		(nottalking)	65		(fish_hit)
14		(angry)	40		(party)	66		(muscle)
15		(annoyed)	41		(sleeping)	67		(smileydance)
16		(wave)	42		(thinking)	68		(bigeyes)
17		B-)	43		(bringit)	69		(funkydance)
18		(cozy)	44		(worship)	70		(idiot)
19		(sick)	45		(applause)	71		(lonely)
20		(:	46		8-)	72		(scenic)
21		(goodluck)	47		(gym)	73		(hassle)
22		(grittongue)	48		(heart)	74		(panic)
23		(mmm)	49		(devil)	75		(okok)
24		(hungry)	50		(lmao)	76		(yahoo)
25		(music)	51		(banana_cool)	77		(cry)
26		(tears)	52		(banana_rock)			

圖三 Plurk 官方提供的 77 種表情符號

正向表情符號	負向表情符號
:)	==+
:-))	:-o
:-)	X-(
:-D	:-(
:-P	:'-(
;-)	:-&
XD	(:
B-)	==
>W<	==|||
:D	>M<
>//<	=□=
>///<	orz
>////<	囧
>/////<	冏
	窘
	>_<
	QQ

圖四 額外整理的正、負向表情符號

3.3 Construct Emotion Dictionary

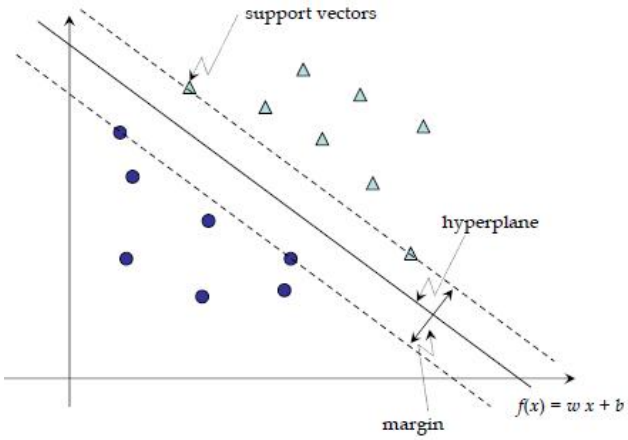
在表情符號的部分，我們利用官方所提供的 77 種表情符號以及我們建立的一些表情符號，並將其分成正負向，當作我們表情符號的情緒辭典。

在 Plurk 文章的部分，我們在網路上收集了一些情緒的詞彙，並將其區分為正負向，我們將這些詞彙視為較重要且有影響力的詞彙；而 training data 中出現的詞彙，則視為中度重要的詞彙，利用這些詞彙做為我們文字部分的情緒辭典。

3.4 Machine Learning in Emotion Classification-SVM

Support Vector Machines是以統計學習理論所延伸發展出來Machine Learning的演算法，是一種監督式學習(supervised learning)，能夠透過訓練(training)，找出效能更好的模型(Model)，提升之後測試(testing)的成效，對於樣本小的資料可預測出較準確且快速的結果，並且擁有風險最小化之特性(林卓彥，2006)[1]。

SVM 往往是拿來做資料分類(Data Classification)，通常會輸入一些訓練資料，根據訓練資料的特性來將資料分為兩類，因此當預測資料輸入時，會對資料做預測分類。



圖表 五 SVM

圖五為 SVM 一個最簡單線性情況，SVM 的最好分類面將 training dataset 分隔開來，其中兩類的文件皆與分類面的距離為最大。而那些距離與分類面最近的元素，我們稱為 support vectors(有圈圈的)。

SVM 在許多應用上幾乎都有很好的表現，它一個明顯的缺點是訓練時間過長。另外，因為 SVM 只能輸出兩種結果(Yes or NO)，是一個二分法的分類器，所以針對多個類別的分類問題時，我們可以採用一個簡單的方法---對每個類別都建立一個分類器，將所有測試文件輸入到每個分類器中，來判別是否屬於該類別。而此研究是利用台灣大學資訊工程系林智仁教授等人開發的 Libsvm 來實做 SVM 的部分。

利用 Libsvm 對 Plurk 文章做情緒分析，有兩個步驟，首先，要先建立 Training Data，我們針對斷詞後的 Training Data，只擷取副詞、動詞及未知詞來判斷情緒，在網路上找到的正負向情緒詞彙，我們給予較高的權重，而若是在 Training Data 中，所擁有的詞，我們就給予較低的權重，會這樣做的原因是因為怕在 Training Data 中，會存在其他的 Stop words(例如，CKIP 所提供的 Stop words list 中沒有

的。)，或是一些比較中立、沒有明確情緒的詞彙，而這些詞彙可能會影響 Training 的結果。在 Libsvm 中，Training Data 必須要弄成指定的格式，如圖六所示。

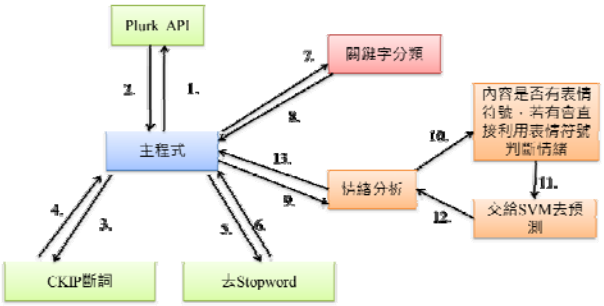
[label] [index1]:[value1] [index2]:[value2] ...
[label] [index1]:[value1] [index2]:[value2] ...

圖六 Libsvm 指定 Training Data 的格式

其中 label 為分類的編號，例如，+1 或-1。而 index 為 term 的 index，value 則是那個 term 的權重。

接著，我們就可以直接利用 Libsvm 將 Training Data 訓練出一個 Model，並利用這個 Model 去對其他的資料做預測。

4. System Overview



圖七 System Overview

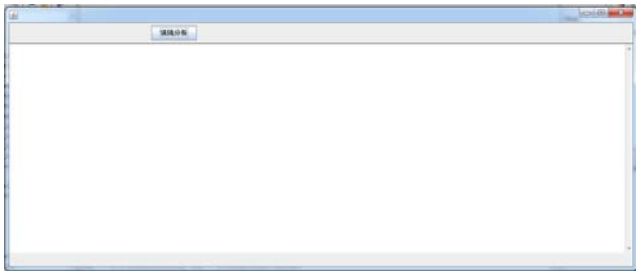
4.1 系統步驟：

我們會先從 Plurk 上面抓取我們的 Data，共 1000 筆，其中 200 筆為 Training data，800 筆為 Testing data。先將抓好的 Data 分成美食、電影及心情三個類別，將分好類別的 Training data 利用 SVM train 出 model，最後再將 Testing data 做情緒分類。

1. 利用 Plurk API 到 Plurk 上抓某個帳號與其好友的 Plurk。
2. Plurk API 每次會回傳 20 則 Plurk 給主程式。
3. 將抓到的 Plurk 利用 CKIP 斷詞。
4. CKIP 傳回斷詞的結果。
5. 將 CKIP 斷詞的結果，去掉 Stopwords。
6. 傳回去掉 Stopwords 結果。
7. 將抓到的 Plurk DATA 利用關鍵字分類。
8. 傳回分類結果(美食、電影、心情)。
9. 將分類好的 Plurk 利用 SVM 做情緒分析。
10. 先判斷 Plurk 內容是否有表情符號，若有會直接利用表情符號判斷情緒。
11. 如果沒有表情符號，會交給 SVM 去預測
12. 回傳情緒分析結果。
13. 在 GUI 上顯示分類結果。

4.2 System Result

我們將所收集到的 Data 透過系統做情緒分類(如圖八)。系統中依照不同類別(電影、美食、心情)顯示每則 Plurk 訊息，而在訊息後面的(+)/(-)/(中立)則是代表這則 Plurk 訊息所代表的情緒。(如圖九)



圖八 系統介面(情緒分析)

圖九 美食情緒分類結果

4.3 System Precision

在系統準確度的部分，我們先將我們所有的 Plurk Data 用人工的方式標記成正向評論、負向評論以及中立評論。當系統將 Testing data 分出結果之後，我們把這些結果與我們標記的結果比對。將所有的結果經過計算之後，我們的系統準確度為 70%。

$$\text{Precision} = \frac{\text{系統結果(正/負/中立)}}{\text{人工標記資料結果(正/負/中立)}}$$

由於是用人工方式標記 Plurk 正向、負向以及中立評論時，我們有時較無法判斷這則 Plurk 的使用者真正想表達的情緒。因此，我們認為準確度為 70%是在可接受的範圍內。

5. Conclusions

5.1 Conclusion

本研究希望透過搜集到的 Plurk 語料，將 Plurk 訊息分成電影、美食與心情三大類，進一步探究其文句的情緒正負向，讓使用者透過此系統得到在電影、美食上相關的正評與

負評，解決一般評鑑網站資訊過量或可信度不高的問題，幫助使用者在想看電影或想吃美食時有一個方向可供參考。

過往的研究在情緒分析上大多集中在文字部分，然而，隨著社群網站的蓬勃發展，文字已不足以表達使用者的情緒，更多時候使用者會加入表情符號來表達其情緒；根據一些文獻探討，通常表情符號更能表達出該句的情緒是正向抑或是負向，因此在此研究中，我們整合 Plurk 官方提供的 77 種表情符號與自行蒐集的常見表情符號，將表情符號加入考量，進而提高了情緒分析準確率。

除此之外，不僅是文句情緒的正負兩種面向，此研究亦考慮情緒中立的面向，此考量是基於有時候評斷者對於某些電影或美食的看法並非正面卻也非負面、可能沒有特別建議但也不至於批評，這種情況，我們視之為一情緒中立的狀況，因此也希望能讓使用者將此情況列為其參考依據之一；過去所探討的文獻中，似乎並未做出中立面向，希望此研究多一面向之探討可提供未來相關研究考量。

以上是屬於資料前處理的部分，而在情緒分類的機器學習方法上，此研究做了一部分文獻探討，最後選定 SVM 做為所使用的分類器，我們將文句以向量形式表示，並且利用 LIBSVM 此工具來進行實作 SVM，最後透過實驗結果，平均準確率達到 70%。

5.2 Future Work

本研究的系統實驗仍存有許多限制的部分，可就情緒詞的權重、情緒辭典的資料來源、詞彙演化、搜尋功能等來做說明。首先，在情緒詞的權重部份，此研究將原先所建立情緒辭典內的詞認定為較重要的詞，因此會賦予之較高的權重；另一方面，透過 Training data 亦可得到一些情緒詞，由於不在原先所建之辭典內，我們認定其為一般情緒詞，故針對 Training data 得到的詞設定較低的權重。此作法屬二分法，考量尚不夠周全，未來，若可將搜集而來的詞畫分為「重要」、「一般」與「不重要」三個權重去考慮，在準確率上應該會提高不少。

此外，在中文情緒辭典資料來源的部分，所參考之文獻其情緒辭典大多來自台灣大學自然語言處理實驗室所建立的語意辭典 NTUSD，然而，此研究礙於時間的考量，僅以網路上搜集到的情緒辭典當做我們正負向情緒詞的資料集，所建立出的情緒辭典並不夠龐大，某些重要的情緒詞可能未被考量進去。倘若使用 NTUSD 的情緒辭典，利用更完善的自然語言知識庫，相信會在文句情緒分類上有更好的表現，進而提升系統的準確率。

此研究已加入表情符號的考量，然而中文詞彙和網路用語的演化與發展迅速，使得許多新詞彙如雨後春筍般出現，這也是對情緒分類準確率的一大挑戰，因此未來，如何自動化維持情緒辭典也是應納入考量的重要議題。

最後，此研究本希望提供 Plurk 使用者一些關於電影與美食的訊息，分析出正負評價，進而輔助使用者在電影與美食方面做決定；然而，這樣的作法卻也同時侷限了使用者所能獲得的資訊，也許該使用者也會想得到如音樂、書籍等的相關正負評價；因此，未來若增加搜尋的功能，就能讓使用者可以輕易找到他想要的資訊並以其正負評價當參考依據，藉由擴張此系統的應用範圍來滿足使用者各式各樣的需求。

6. Member's Workload

施怡如-Plurk 情緒分析

陳瀅如-斷詞、去 Stopwords、分類

郭岱茵-系統介面、蒐集 Plurk DATA

7. REFERENCES

[1] 林卓彥，自動分類方法之比較，中正大學資訊工程研究所，2006

[2] 周嵩能，以微網誌語料進行情緒分析之研究，台南大學資訊工程學系碩士論文，July, 2009

[3] 楊昌樺，網路日誌空間情緒分析方法之研究，台灣大學電機工程學系博士論文，July, 2009

[4] 楊鼎，楊愛民，一種基於情感辭典和 Naïve Bayes 的中文文本情感分類方法，計算機應用研究，Vol.27, No.10, Oct. 2010

[5] Bing Liu, Minqing Hu, Junsheng Cheng, Opinion Observer: Analyzing and Comparing Opinions, International World Wide Web Conference Committee, 2005

[6] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features", Proceedings of the European Conference on Machine Learning Springer, 1998

[7] Gilad Mishne, Experiments with Mood Classification in Blog Posts, Stylistic Analysis Of Text For Information Access, 2005

[8] R. Thayer, The Biopsychology of Mood and Arousal: Oxford University Press, USA, 1989

[9] C.-H. Yang, H.-Y. Lin, and H.-H. Chen, "Emotion Classification Using Web Blog Corpora," Proceedings of the IEEE/WIC/ACM International Conference on Web *Intelligence*, IEEE Computer Society, pp. 275-278, 2007.

[10] Yiming Yang and Xin Lin, "A Re-Examination of Text Categorization Methods", Proceedings of the 22nd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval ,1999, Pages 42-2