

中文短句之情緒分類

Sentiment Classification of Short Chinese Sentences

孫瑛澤[§] 陳建良[†] 劉峻杰[↑] 劉昭麟[‡] 蘇豐文[↓]
Ying-Tse Sun Chien-Liang Chen Chun-Chieh Liu Chao-Lin Liu Von-Wun Soo
^{§†‡}國立政治大學資訊科學系
National Chengchi University, Taipei, Taiwan
[↑]國立清華大學資訊工程系
National Tsing Hua University, Hsinchu, Taiwan
{[§]s9538, [†]g9813, [‡]chaolin}@cs.nccu.edu.tw, [↓]soo@cs.nthu.edu.tw

摘要

本研究工作探索文件分類的一個極端情境，依照微網誌中具有表情符號的文字內容，來猜測該敘述的情緒類別。短句的字數有限且欠缺上下文的語境，是較傳統文件分類難處理的地方；另一方面，微網誌之短句常直接帶有一些提示情緒類別的字眼，因此有助於簡單的情緒分類。我們將微網誌的短句分為正面情緒、負面情緒兩種類別；應用詞彙與情緒標記的統計資訊，加上同義詞和情緒詞彙辭典等資訊的輔助，運用向量空間模型、向量支持機與機率模型等概念來進行分類工作。分類準確率依照實驗設計的程序不同有相當大的差異，最高約超過八成，最低則僅約五成。

Abstract

We explore an extreme case of text classification. The short statements in micro-blogs were collected, and were associated by a category based on the sentiment indicated by the associated icons. We evaluated different methods that assigned the categories with just the wordings in the short statements. Short statements in micro-blogs are harder to classify because of the shortage of context, yet it is not rare for the statements to include words that may be linked to sentiments directly. In this work, we considered two polarities of sentiments: negative and positive. We employed the statistical information about the word usage, a dictionary for Chinese synonyms, and an emotional phrases dictionary to convert short statements into vectors, and applied techniques of support vector machines and probabilistic modeling for the classification task. The results of classification varied with the classification methods and experimental setups. The best one exceeded 80%, but the lowest just made 55%.

關鍵詞：情緒分類、文件分類、支援向量機、naïve Bayes、特徵選擇

1. 緒論

Web 2.0 時代的來臨改變媒體傳播方式以及人類與資訊的互動關係，不再是由傳統的媒體扮演資訊生產者和資訊守門員的角色，而是使用者導向的互動式媒體，強調使用者自己生產資訊並選擇自己所要接收的資訊，使得網路上的數位內容越來越豐富，其中部落格、YouTube 是最佳的代表。時下許多的使用者透過 Web 2.0 網站以文字、圖片、影音

等媒介抒發自己的心情或發表對於非特定人事物的看法，這些數位內容有別於傳統媒體具有高度相似性的傳播內容或具有某種程度的客觀性的意見陳述，而是多元化的內容和意見的主觀性，也融入更多個人化的情緒和想法。近年來，更衍生出微網誌(micro-blog)的概念，最典型的例子像是推特(Twitter)和撲浪(Plurk)；微網誌沒有完整的文章結構，使用者必須在少數且有限的字數下表達使用者當下想傳達的內容和狀態，因為本身文句的結構非常簡短，使用者必須以更短的詞彙和語句表達，所以更有利於使用者間的即時互動和回應，也讓情緒表達變得更明顯和簡短，同時避免了冗長且複雜的情感陳述。

情緒偵測(emotion detection)的發展對於人類與機器的互動具有高度的應用價值，機器可以透過偵測人類的情緒提供具有附加價值的服務或做出適當的回應，扮演一個更值得信任角色[8][14]，包括依照使用者情緒推薦相符合的文章、音樂等商品。

本研究透過知名微網誌撲浪(Plurk)[3]的中文短句中的情緒詞彙進行情緒文件分類，有別於傳統文件分類。我們將文件分類技術應用於以文本為基礎的情緒偵測(text-based emotion detection)問題。本研究的語料專注在短篇文件，其中每一篇文件僅由一句短句所組成。因為短篇文件所包含的語境和詞彙通常比較不足夠，所以短篇文件的文件分類效果通常會比長篇的文件分類效果不佳。我們假設每一句短句至少包含一個情緒詞彙且每一篇可以被分類為正面情緒和負面情緒兩種分類。本論文將於第二節探討應用文件分類相關技術於情緒偵測的文獻，第三節介紹語料、斷詞方法與詞彙選擇、三種不同分類模型，第四節則是實驗結果和分析，最後第五節為結論與未來展望。

2. 相關背景

情緒相關的研究與定義在資訊科學蓬勃發展前，在心理學等領域已經是一個熱門且成熟的研究議題，人類在不同的情境底下會有不一樣的情緒反應。透過文件分類技術可以預測其他使用者發表文章的情緒。常見從文字中偵測情緒問題的方法，找出文字內容和經由人工標記的情緒類別之間的關聯性；當我們所蒐集的語料夠多，使得找出文字與情緒類別之間的關聯性具有顯著相關時，就可以利用這些關聯性，預測未知情緒類別的文章之可能情緒。

有許多情緒偵測方法所適用的範圍為句子層次的推論，也有段落層次和全篇文章層次的情緒偵測方法 [5][9]。雖然每一句短句在部落格文章中可能有不同的情緒，但整篇文章在語意上和語法上應該會有一致的傾向(polarity)，傾向於正面情緒或負面情緒的其中一面。因為文章由許多的語句所組成，句子層次的情緒偵測研究為全文層次情緒偵測研究的基礎，本篇研究專注在句子層次的情緒偵測研究。接下來將探討一些使用文件分類相關技術於情緒偵測的文獻，彼此最大的差異在於偵測方法上的差異。

2.1. 關鍵字偵測法(Keyword-Based Detection Methods)

關鍵字偵測法是最簡單也是最符合直覺的文字情緒偵測方法。情緒偵測的判斷僅透過事先選定的相關關鍵字是否出現在給定文件中。

關鍵字偵測法中可以增加分類準確性的關鍵處為前處理的部分。其中前處理的方法可以分為剖析器萃取關鍵字和建立情緒關鍵字辭典兩種方法：第一種方法，透過剖析器

(parser)萃取出情緒關鍵字，目前已經具有成熟的剖析器軟體，可以從給定的訓練語料中利用句子剖析的方法萃取出具有代表性的關鍵字；第二種方法，建立情緒關鍵字辭典，情緒關鍵字辭典除了包含情緒字彙外，還必須描述字彙之間的關聯性和所屬的情緒類別，爲了清楚描述語意相似字彙和語意相反字彙的關聯性和分類關係，情緒關鍵字辭典通常以同義詞辭典(thesaurus)或者本體論表示法(ontology representation)的形式存在。過去文獻[6][19]中，許多研究提出類似 WorldNet 辭典結構的半自動情緒關鍵字辭典和情緒類別標籤建立方法。

2.2. 學習偵測法(Learning-Based Detection Methods)

有別於關鍵字情緒偵測方法透過主觀的人工方法建立關鍵字辭典或透過半自動方法剖析出可能的關鍵字候選名單，學習偵測法透過客觀的統計式機器學習演算法訓練出的學習樣式(learned pattern)判斷給定的輸入文字所屬的情緒類別。過去文獻中使用支援向量機(support vector machines)[18]和條件隨機域(conditional random field)[20]統計式機器學習演算法幫助情緒類別的分類問題。

2.3. 混合方法(Hybrid Methods)

單純使用關鍵字偵測方法或統計式機器學習演算法無法得到令人滿意的分類效果，許多混合方法將情緒偵測結合已偵測的關鍵字、學習演算法得到的樣式(pattern)和其他補充資訊的調整(refinement)一起判斷，使得系統的分類效果明顯的改善。Wu 等學者[18]提出的混合方法，採用規則式方法(rule-based approach)萃取特定情緒相關的語意類別再利用中文詞彙本體論方法萃取出屬性(attributes)，這些語意類別和屬性以關聯規則(association rule)方式表示之間的情緒關聯性；接著以這些關聯規則取代上述的關鍵字當成統計式學習模型訓練過程中的特徵(feature)。

3. 研究方法

本節介紹三種不同分類模型的情緒偵測方法。首先 3.1 節介紹語料的來源微網誌撲浪(Plurk)的內容，3.2 節介紹使用者在訊息中加註的表情符號作爲情緒類別；接著開始介紹三種不同的分類模型：3.3 節 CKIP_SVM 法使用自行建構詞彙辭典搭配支援向量機(SVM)進行情緒分類，3.4 節 NTUSD_NB 法使用現成的 NTUSD 情緒詞彙辭典搭配 naïve Bayes 機率模型進行情緒分類，最後 3.5 節 Hybrid 法使用權重結合 CKIP_SVM 法和 NTUSD_NB 法達成情緒分類。

3.1. 中文短句：撲浪(Plurk)

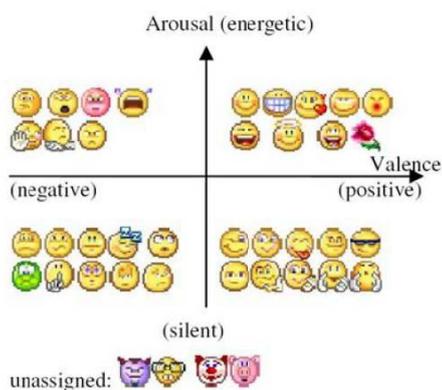
撲浪是一個免費的社交網站，它允許每一個人擁有類似自己的部落格的文字輸入介面，可以在任何時間點留下自己的心情狀態和當下的想法。微網誌強調比部落格更即時的紀錄，因此常被稱爲「以秒爲單位的日記」，撲浪限制每則文字的訊息長度在 140 個字內，有別於一般部落格的長篇文章，也造成大部分的使用者發的訊息經常以短句或甚至幾個簡單的字詞所組成。撲浪的使用者編輯介面有一個特別的設計，當使用者要發出訊息

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。

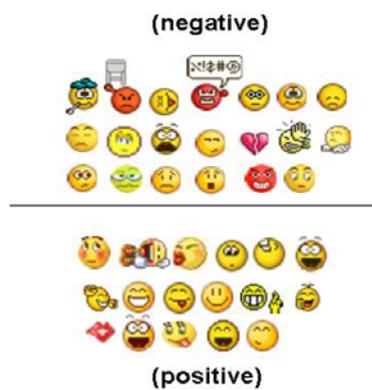
3.2. 情緒類別

Yang 等學者[20]提出的表情符號四象限研究方法，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(如圖一)，部分的表情符號與四個象限沒有任何關聯性分類為 unassigned。Yang 的方法是基於 Thayer 所提出的心理學模型，橫軸(東西向)衡量正面和負面的情緒程度(pleasantness 和 unpleasantness 兩極)，縱軸衡量活動程度的強與弱(arousal 和 sleepiness 兩極)[16]。

在情感分析(sentiment analysis)研究[14]中的 polarity 分類器將案例(instance)分成正面類別和負面類別兩種類別。本研究不考慮 energetic-silent 軸，將表情符號簡化為正面(positive)和負面(negative)兩大情緒類別(如圖二)。另外，Yang 所使用的語料是 Yahoo! Kimo blog 的四十個表情符號作為分類，而本研究採用撲浪的三十七個表情符號作為分類與 Yahoo!的表情符號不完全一樣。



圖一



圖二

3.3. 方法一：自行建構詞彙辭典搭配支援向量機(CKIP_SVM 法)

3.3.1. 前處理：CKIP 中文斷詞

詞是最小有意義且可以自由使用的語言單位，在中文的自然語言處理問題中必須先能分辨出文中的詞才能進行後續處理，因此本研究將以短句為單位的撲浪語料輸入中央研究院(CKIP)小組所開發的中文斷詞系統進行斷詞[1]。

根據語料中的表情符號和前節中所討論的情緒類別方法，我們可以在斷詞前就知道語料中的個別短句分別屬於哪一個情緒類別(正面情緒或負面情緒)，再將斷詞完的詞彙與詞彙出現次數依據所屬的表情符號歸類為正面情緒詞彙表和負面情緒詞彙表。

接著，將上述兩個詞彙表當作特徵選擇的候選詞彙集，希望藉由特徵選擇的方法降低分類問題中可能產生的 curse of dimension 的問題並提高分類準確性。

3.3.2. 特徵選擇(Feature Selection)

爲了縮減候選詞彙的數量並找出和情緒類別高度相關的詞彙集，本節介紹特徵選擇方法。在自然語言處理中，有許多常用的特徵選擇演算法，包括 point-wise mutual information、Pearson Chi-Square correlation 等。本研究利用 point-wise mutual information 方法，在經過中文斷詞處理的候選詞彙集中找出和情緒類別最相關的詞彙子集合，接著透過中文同義詞辭典(thesaurus)下降特徵維度。

3.3.3. Point-Wise Mutual Information

Point-wise mutual information(PMI)是一種衡量兩樣本之間相依程度關係的方法[12]，在自然語言處理領域中經常使用 PMI 處理詞彙共同出現關係(word collocation)和詞彙歧義解析(word sense disambiguation)。PMI 的數學定義如下：

$$PMI(w_1, w_2) = \log_2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)}, \quad (1)$$

其中 w 是前段所述的詞彙集中的某一詞彙且是一特徵， w_1 和 w_2 個別出現的機率 $P(w_1)$ 和 $P(w_2)$ 可以透過計算 w_1 和 w_2 個別出現的次數作爲機率估計值；而 w_1 和 w_2 共同出現的機率 $P(w_1, w_2)$ 代表 w_1 和 w_2 兩個字共同出現(co-occurrence)的機率，可以透過計算兩個字在特定距離內共同出現的次數作爲機率估計值。本實驗與 Yang[20]一樣採用變形的 point-wise mutual information。

$$Co(e, w) = C(e, w) \times \log_2 \frac{P(e, w)}{P(e)P(w)}, \quad (2)$$

其中 e 爲情緒的類別標籤，正面情緒或負面情緒； $c(e, w)$ 是情緒類別 e 和詞彙 w 共同出現在語料中的次數， $c(e)$ 和 $c(w)$ 是情緒類別 e 和詞彙 w 分別出現在語料中的次數； $P(e, w) = C(e, w)/N$ ， $P(e) = C(e)/N$ ， $P(word) = C(w)/N$ ，其中 N 爲所有詞彙出現的總數。每一句短句至少有一個詞彙可以計算 PMI 值。圖三是兩句未經過正規化的短句的正面情緒和負面情緒詞彙共同出現關係。爲了讓 PMI 值的可比較性增加，我們使用下列公式對 PMI 值進行正規化(normalization)使其落入固定的範圍內：

$$PMI \ score = \frac{x_i - co_{min}}{co_{max} - co_{min}}, \quad (3)$$

x_i 爲 $word_i$ 的 PMI 值， co_{min} 爲在語料中最小的 PMI 值， co_{max} 爲語料中最大的 PMI 值。將所有的情緒和詞彙共同出現的組合(emotion-words collocation pairs)的 PMI score 依照降冪排列，再從中選擇前 FN 個情緒和詞彙共同出現的組合作爲特徵子集合。設定門檻值 T 來刪除一些不具有鑑別性的詞彙，如果 $collocation_1$ 和 $collocation_2$ 兩組組合相減的差小於門檻值 T，則分別將兩組組合從特徵子集合刪除，兩組組合的差距很小代表相同的詞彙但是情緒類別卻不同。若要透過所有情緒和詞彙共同出現的組合分辨所屬的情緒相對來說比較困難，容易造成錯誤分類，因此刪除這些不具鑑別力的情緒和詞彙共同出現組合可以增加特徵子集合的品質。圖四是特徵選擇的演算法。下小節中透過中文同義詞辭典將相似詞義的詞彙集合轉換成同一個特徵維度，使得特徵維度的大小下降。

3.3.4. 利用中文同義詞辭典下降特徵維度

同義詞辭典是一種將概念相同之詞彙匯集在一起，不受詞彙表面上的形式所侷限，並竭盡所能網羅相類似的詞彙在同一個概念的分支下。本研究選用中文同義詞辭典，最初為梅家駒學者以人工方式編攥的『同義詞詞林』再經過哈爾濱工業大學整理與增加為擴展

範例：正面情緒	範例：負面情緒
<u>大家來回憶一下美好的童年吧</u>	<u>為什麼我不管中文名字還是英文都很菜市場</u>
Co(positive, 大家)= 2049.934	Co(negative, 為什麼)=492.0230
Co(positive, 來)= 1327.704	Co(negative, 我)= 350.096
Co(positive, 回味)= 11.400	Co(negative, 不管)= 6.729
Co(positive, 一下)= 331.540	Co(negative, 是)= 260.103
Co(positive, 美好)= 266.122	Co(negative, 中文)= -7.205
Co(positive, 童年)= 0.670	Co(negative, 名字)= -18.137
Co(positive, 吧)= 331.012	Co(negative, 還是)= 396.644
	Co(negative, 英文)= 25.030
	Co(negative, 都)= 1230.832
	Co(negative, 很)= 77.766
	Co(negative, 菜市場)= 5.201

圖三 正負面情緒和詞彙共同出現關係

演算法：PMI score 特徵選擇

參數定義：原始共 N 個特徵，目標特徵數量 FN，

情緒和詞彙共同出現組合 PMI 值的差之門檻值 T。

輸入：輸入 2N 組的候選集合{ $(w_1, e_1, PMI_1), (w_1, e_2, PMI_2), \dots, (w_k, e_1, PMI_{2k-1}), (w_k, e_2, PMI_{2k}), \dots, (w_N, e_1, PMI_{2N-1}), (w_N, e_2, PMI_{2N})$ }

初始化：將候選集合依 PMI 值大小排序，設定特徵計數器 $fc=0$ ，
特徵詞彙集合 $FWS=\{\}$ 。

處理程序：

```

Do for  $i = 1$  to  $2N$ 
  IF (FWS 內沒有包含  $w_i$ )
    將  $(w_i, e_i)$  加入 FWS，
     $fc$  加一。
  ELSE
    依據  $w_i$  從 FWS 中找出其 PMI 值  $PMI_j$ 。
    IF  $(PMI_j - PMI_i \leq T)$ 
      從 FWS 中將  $(w_j, e_j)$  移除，
       $fc$  減一。
    IF  $(fc > FN)$ 
      終止演算法。
  
```

輸出：特徵詞彙集合(FWS)

圖四 特徵選擇的演算法

版[2]，內容以現代漢語語詞為主，除了一般字彙、成語之外，亦收錄一些方言詞、古語詞，且除了同義詞的分類外，亦將同一詞義範疇之所有相關詞彙納入，其詞彙總數大約七萬多字。同義詞詞林內部分類共有四層，採階層式分類，最高層第一層為越抽象的類別，最低層第四層為越抽象的細類別，越底層的分類種類越多，越高層則越少，另外每一個同義詞有一個內部編碼代表其階層式分類。例如：「老師」、「學生」、「警察」這些具體的細類別為抽象類別「人類」的下層類別。我們將同義詞詞林第一層編碼相同的詞彙分類在同一類中且將對應至相同的特徵維度計算，依據上段所計算的 PMI score 標記每一個詞彙為正面情緒或負面情緒。若詞彙不存在，則標記為 null。

圖五中第一個例子，Eb37B 為詞彙「無力」的編碼，其中 E 代表第一層的編碼，Eb 為第二層的編碼。從圖中可以發現詞彙「無力」屬於第一層的 E 類別，又負面情緒的 PMI 大於正面情緒的 PMI，因此我們將「無力」歸為負面情緒且屬 E 類別(negative_E)；詞彙「涼爽」同屬於第一層的 E 類別，但正面情緒的 PMI 大於負面情緒的 PMI，因此我們將「無力」歸為正面情緒且屬 E 類別(positive_E)；詞彙「噁心」沒有出現在同義詞詞林中，所以為 null。

3.3.5. 特徵向量轉換

詞彙經過特徵選擇後，必須將經過上述特徵詞彙轉換成二元特徵向量以利於輸入支援向量機(SVM)[7]進行情緒分類。

同義詞詞林第一層共有十二個類別加上額外的 null 類別，再乘上情緒類別正面和負面兩類別，總共可以得到 26 個特徵維度。如圖六所示，其中「1: P_A」代表第一個特徵維度為同義詞詞林的第一層 A 類且屬於正面的詞彙，只要短句中出現 A 類且屬於正面的詞彙則特徵向量的值為一，否則為零。以「今天辦公室有超多點心」為例，經過前段的處理後，會被轉換成「今天:P_C」、「辦公室:N_B」、「有:P_E」、「超多:P_null」和「點心:P_B」，根據圖六有出現的會被標記為 1，沒有出現的標記為 0，則此例會被轉換成向量[0,0,1,1,1,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0]。為了符合 LIBSVM[10]的輸入格式，向量呈現方式為「-1 3:1 4:1 5:1 9:1 25:1」，最前面的-1 代表是負面情緒，若是 1 則代表正面情緒；3,4,5,9,25 則代表 1 出現的位置。

3.4. 方法二：NTUSD 辭典搭配 naïve Bayes 模型(NTUSD_NB 法)

Example 1: 無力(Eb37B)	Example 2: 涼爽(Eb27A)	Example 3: 噁心(null)
Co(negative, 無力)=0.0057	Co(negative, 涼爽)=0.0009	Co(negative, 噁心)=0.0061
Co(positive, 無力)=0.00053	Co(positive, 涼爽)=0.0027	Co(positive, 噁心)=0.0003
Result: 無力(negative_Eb)	Result: 涼爽(positive_Eb)	Result: 噁心(negative_null)

圖五 同義詞詞林

1:P_A	5:P_C	9:P_E	13:P_G	17:P_I	21:P_K	25:P_null
2:N_A	6:N_C	10:N_E	14:N_G	18:N_I	22:N_K	26:N_null
3:P_B	7:P_D	11:P_F	15:P_H	19:P_J	23:P_L	
4:N_B	8:N_D	12:N_F	16:N_H	20:N_J	24:N_L	

圖六 向量中的特徵維度

3.4.1. NTUSD 情緒詞彙辭典

NTUSD [13]是台灣大學自然語言處理實驗室所建立的語意辭典，裡面包含了正面以及負面情緒的詞彙集，正面情緒大約有 2800 筆詞彙，負面情緒大約有 8000 筆詞彙，總共大約 10000 多筆詞彙。NTUSD_NB 法將撲浪上獲得的語料直接利用 NTUSD 辭典進行斷詞，擷取其中的情緒詞彙當作情緒分類的特徵，亦即僅考慮出現在 NTUSD 的詞彙。以下的 naïve Bayes 機率模型就是利用 NTUSD 來當作基礎所做的情緒判斷的方法。

3.4.2. Naïve Bayes 機率模型

Naïve Bayes 機率模型是以貝氏定理(Bayes' theorem)為基礎，透過事前機率(prior probability)、事後機率(a posterior probability)將未知類別的測試資料指定給最大類別機率的類別。依據貝氏定理中最大事後機率(Maximum a posterior probability, MAP)可以得到下列分類的法則：

$$e_{MAP} = \arg \max_{e \in E} P(e | d) = \arg \max_{e \in E} \frac{P(d | e)P(e)}{P(d)} = \arg \max_{e \in E} P(d | e)P(e) \quad (4)$$

e_{MAP} 為最大類別機率的類別， E 為所有類別的集合 {正面情緒，負面情緒}， e 為類別的值， d 為未知類別的測試文件。透過上述公式可以由計算 $P(d | e)P(e)$ 得到推論最有可能的機率類別 e_{MAP} 。

Manning[11]依照隨機變數的定義方式將 Naïve Bayes 區分成 Multinomial Naïve Bayes 和 Bernoulli Naïve Bayes 兩種模型。兩個模型最大的不同在於產生給定類別 e 下的文件 d 條件機率 $P(d | e)$ 的方式不同。在 Multinomial 模型中，我們以文字集合 $X = \langle t_1, \dots, t_k, \dots, t_n \rangle$ 作為文件 d 的表示方式，其中 $\langle t_1, \dots, t_k, \dots, t_n \rangle$ 為在文件 d 出現過的詞彙去除沒有出現在 NTUSD 所有的 n 個詞彙中；在 Bernoulli 模型中，透過 n 維二元陣列 $B = \langle b_1, \dots, b_k, \dots, b_n \rangle$ 表示第 k 個出現在 NTUSD 的詞彙是否出現在文件 d 中，有出現 b_k 為 1、沒有出現 b_k 為 0。透過條件獨立假設(conditional independence assumption)和位置獨立假設(positional independence assumption)使得下列條件機率關係成立：

Multinomial :

$$P(d | e) = P(\langle t_1, \dots, t_k, \dots, t_n \rangle | e) = \prod_{k=1}^n P(t_k | e) \quad (5)$$

Bernoulli :

$$P(d | e) = P(\langle b_1, \dots, b_k, \dots, b_n \rangle | e) = \prod_{k=1}^n P(b_k | e) \quad (6)$$

由上述公式(4)和(5)可以得到 Multinomial 的分類法則：

$$e_{MAP} = \arg \max_{e \in E} P(d | e)P(e) = \arg \max_{e \in E} \prod_{k=1}^n P(t_k | e)P(e) \quad (7)$$

同樣從公式(4)和(6) 可以得到 Bernoulli 的分類法則：

$$e_{MAP} = \arg \max_{e \in E} P(d | e) P(e) = \arg \max_{e \in E} \prod_{k=1}^n P(b_k | e) P(e) \quad (8)$$

上式我們使用最大概似估計法(Maximum likelihood estimation, MLE)分別估計事前機率 $P(e)$ 和事後機率 $P(t_k | e)$ 。另外，對條件機率 $P(t_k | e)$ 和 $P(b_k | e)$ 採取 Laplace smoothing 以避免資料過於稀疏產生機率為零的情況。

3.4.3. 機率分數與類別轉換

給定測試語料中某篇文件經過 Bernoulli naïve Bayes 跟 Multinomial naïve Bayes 模型計算之後，正面情緒與負面情緒的兩個類別各有一個機率數值(實數值)，再依兩種類別相對的機率大小對此篇文件進行分類。此處將所算出的機率數值稱作「機率分數」，為落在 0~1 區間的實數值。「類別分數」為 0 和 1 所組成的二元離散類別分數，若依據機率數值做出分類決定，結果是正面情緒，令其「類別分數」為 1；反之結果是負面情緒，令其「類別分數」為 0。此處的「類別分數」和「機率分數」為混合模型的不同輸入值，將在下一節詳細敘述不同分數對混合模型的影響。

3.5. 方法三：混合模型 (Hybrid 法)

我們提出混合模型(hybrid model)，融合兩種模型可以互補前兩項方法的缺點以及優點，得到更加的分類效果。利用權重的概念將兩種模型整合在同一混合模型：

$$\text{Hybrid (SVM, NB)} = \lambda \times \text{SVM(CKIP)} + (1 - \lambda) \times \text{NB(NTUSD)} \quad (9)$$

其中SVM(CKIP)是採用CKIP斷詞系統搭配同義詞詞林自行建構詞彙再以 SVM 模型對短句情緒分類的結果，若結果為正面情緒則 SVM(CKIP)=1，負面情緒則為 0；NB(NTUSD)使用 NTUSD 情緒詞彙辭典搭配 naïve Bayes 機率模型所做的短句情緒分類結果，分成兩種情形，一是將NB(NTUSD)代入「類別分數」，二是代入「機率分數」； λ 和 $(1 - \lambda)$ 為權重的概念控制SVM(CKIP)和NB(NTUSD)兩個方法在混合方法的重要程度。

$$\begin{aligned} \min_{\lambda} \sum_{\text{all } i} \text{error}_i &= \min_{\lambda} \sum_{\text{all } i} (\text{Reality}_i - \text{prediction}_i)^2 \\ &= \min_{\lambda} \sum_{\text{all } i} (\text{Reality}_i - \text{Hybrid}(\text{SVM}, \text{NTUSD})_i)^2 \end{aligned} \quad (10)$$

混合系統中常用權重的方式控制各種方法在整個系統中貢獻比例，常用最佳化方法找出最佳權重。Hybrid 法中權重 λ 的大小，代表SVM(CKIP)和NB(NTUSD)應該各自在混合方法中貢獻多少比例，最適合的 λ 可以使得混合方法有最佳的分類效果。我們利用一群訓練資料估計最佳的權重值。給定方程式(10)為最小平方誤差的總和(minimum sum squared error)，透過偏微分求解最佳的權重 λ 。其中 error_i 代表訓練資料中第 i 筆資料的平方誤差； Reality_i 代表第 i 筆資料的真實類別值，0 代表負面情緒類別而 1 代表正面情緒類別； prediction_i 為第 i 筆資料透過混合方法Hybrid(SVM, NTUSD) $_i$ 所算出的預測值。

混合模型輸出的預測值為實數值，為了方便後續準確率、精確率和召回率的計算，

透過門檻值 T' 的設定將實數值轉換為二元離散數值(0 和 1)。若預測值大於門檻值，則判斷為正面情緒；若預測值小於門檻值，則判斷為負面情緒。為了使混合模型的分類效果最佳，我們在 0 到 1 區間內搜尋使得分類準確率最高的最佳門檻值 T' 。

4. 實驗分析

4.1. 語料與衡量方法

表一 語料數量與比例

本研究中的實驗語料選用撲浪的限制：句長為小於 25 個字元的繁體中文所組成的短句，並且以所對應的表情符號作為正面情緒和負面情緒的類別依據。另

	正面情緒		負面情緒		總計
	筆數	比例	筆數	比例	
模型訓練語料	4649 筆	46.7%	5351 筆	53.5%	10000 筆
參數調整語料	2281 筆	45.6%	2719 筆	54.4%	5000 筆
測試語料	1903 筆	44.6%	2364 筆	55.4%	4267 筆
總計	8833 筆	45.8%	10434 筆	54.2%	19267 筆

外，為了達成監督式分類的目標，選用的短句必須同時包含文字和表情符號，否則一律不採用。我們從撲浪共收集 19267 筆短句作為實驗語料分析，其中 10000 筆當作訓練語料，再選擇 5000 筆語料作為混合模型參數調整語料，最後保留其他 4267 筆作為測試語料，分佈狀況如表一 語料數量與比例。我們使用準確率(Accuracy)、精確率(Precision)、召回率(Recall)和 F1 分數衡量實驗效果。定義如下，準確率為正確分類成正面情緒和負面情緒的案例個數除以所有案例個數；精確率為正確分類成正面情緒(或負面情緒)的案例個數除以所有分類模型預測為正面情緒(或負面情緒)的案例個數；召回率為正確分類成正面情緒(或負面情緒)的案例個數除以所有正面情緒(或負面情緒)的案例個數；F1 分數則是結合了精確率和召回率如下：

$$F1 \text{ 分數} = \frac{2 \times \text{召回率} \times \text{精確率}}{\text{召回率} + \text{精確率}} \quad (11)$$

4.2. 實驗結果

4.2.1. CKIP_SVM 法

在 CKIP_SVM 法中，10000 筆訓練語料經過 CKIP 斷詞系統可以斷出 80556 組詞彙。接著為了下降特徵的維度，再將斷詞後的候選詞彙集合輸入 PMI score 特徵選擇演算法中得到 10535 組特徵詞彙集合(FWS)，其中演算法中 PMI 的門檻值設定為 0.62，最後依照同義詞詞林的第一層將相似詞義的詞彙集合轉換成同一個特徵維度得到 26 個特徵維度。我們選用 LIBSVM[10]中預設的核心函數 Radial Basis Function (RBF)作為情緒分類模型。RBF 核心函數需要選定兩個參數， C 和 γ ，且參數的好壞往往影響分類效果的優劣，因此本研究使用 LIBSVM 的工具套件中的 Grid 搜尋工具從 10000 筆訓練語料中找出最佳的 C 和 γ 分別為 1 和 0.25，使得在訓練語料中準確率的最大值約 78.39%。

表三為利用 4267 筆測試資料所得到的分類效果評比。從簡單平均分類效果可以發現精確率、召回率、F1 分數和準確率都落在 52%到 56%間；其中正面情緒的召回率遠

表二 CKIP_SVM 法 Confusion matrix

SVM (案例數)		真實類別 正面情緒	真實類別 負面情緒
預測	正面情緒	350	282
	負面情緒	1589	2046

表三 CKIP_SVM 法的分類效果

SVM	精確率	召回率	F1 分數	準確率
正面情緒	55.38%	18.05%	27.23%	56.15%
負面情緒	56.29%	87.89%	68.62%	
簡單平均	55.83%	52.97%	54.36%	

遠低於負面情緒的召回率的原因，可以從表二中正面情緒分類為正面情緒僅 350 筆，遠小於正面情緒分類為負面情緒的 1589 筆，使得正面情緒召回率極低；除了原始語料負面情緒的比例較高外，SVM 分類模型預測結果為負面情緒的比例也過高造成 false negative 數量較大。表四中上述的五句例句利用空白隔開表示經過 CKIP 斷詞後的結果，<討厭>表示候選的特徵詞彙；P 代表正面情緒類別，N 代表負面情緒類別。例句 1 中<討厭>這個特徵詞彙與負面情緒類別的共同出現程度較高，導致將正面情緒錯誤分類為負面情緒；例句 2 和 3 中雖然都有<懶散>這個特徵詞彙，但前者的「懶散」是描述意見持有者(opinion holder)的負面精神狀態，而後者的「懶散」是形容「睡到自然醒」這個動作也暗示意見持有者心中是正面情緒的，不過對於分類模型而言「懶散」是一個負面傾向比較高的特徵詞彙，因此容易造成錯誤分類成負面情緒；同樣的例句 4 和 5 都出現一樣的特徵詞彙組合，但是後者是一個包含詞彙「不」的否定句，但目前的斷詞系統會將「不」單獨斷詞造成分類不易。包含有否定詞彙的例句通常會造成語意相反，然而否定句不一定代表例句為負面情緒，必須同時考量否定詞彙修飾的詞彙，例如：「不太順利」是負面的，「不壞」是正面的。另外，否定詞彙可能不是緊鄰修飾詞彙，這是單獨靠斷詞系統選擇特徵詞彙的研究限制。

表四 CKIP_SVM 法的分類結果分析

經過斷詞後的撲浪語料		真實	預測
1	最 <討厭> 的 <英文> 終於 <考完> 囉	P	N
2	剛 放完 假 卻 有 <疲倦> <懶散> 的 <感覺>	N	N
3	<今天> 又是 <懶散> 的 在家 一天 睡 到 <自然> 醒	P	N
4	<今天> <考試> 真的 很 <順利>	P	P
5	<今天> <考試> 不太 <順利>	N	P

4.2.2. NTUSD_NB 法

在 NTUSD_NB 法中，我們利用 10000 筆訓練資料分別依照前面章節所陳述訓練。Bernoulli naïve Bayes 和 Multinomial naïve Bayes 兩種分類模型以 4267 筆測試資料得到分類結果如表五。因為撲浪有許多的未知字詞，若一筆語料中所有斷詞結果的詞彙都沒有出現在 NTUSD 的辭典中，我們不對此語料進行分類，並列入拒絕率的計算中。由表六我們可以發現負面情緒的分類精確率和召回率都顯著高於正面情緒，原因除了語料的正面情緒比例小於負面情緒比例外，最重要的因素是 NTUSD 辭典中的正面情緒詞彙集大約有 2800 筆詞彙，負面情緒詞彙集有大約 8000 筆詞彙，比例上差距懸殊。在相同測試語料和拒絕率情況下，Bernoulli 模型的正面情緒召回率明顯優於 Multinomial 模型的正面情緒召回率，然而 Bernoulli 模型的負面情緒召回率卻明顯低於 Multinomial 模型的

表五 NTUSD_NB 法 Bernoulli 和 Multinomial 模型的 Confusion matrix

Bernoulli (案例數)		真實類別 正面情緒	真實類別 負面情緒	Multinomial (案例數)		真實類別 正面情緒	真實類別 負面情緒
預測 類別	正面情緒	663	164	預測	正面情緒	594	100
	負面情緒	223	1134	類別	負面情緒	292	1198

表六 NTUSD_NB 法 Bernoulli 和 Multinomial 模型的分類效果

Bernoulli	精確率	召回率	F1 分數	準確率	拒絕率
正面情緒	80.17%	74.83%	77.41%	82.28%	48.82%
負面情緒	83.57%	87.37%	85.42%		
簡單平均	81.87%	81.10%	81.48%		
Multinomial	精確率	召回率	F1 分數	準確率	拒絕率
正面情緒	85.59%	67.04%	75.19%	82.05%	48.82%
負面情緒	80.40%	92.30%	85.94%		
簡單平均	83.00%	79.67%	81.30%		

表七 NTUSD_NB 法 Bernoulli 和 Multinomial 模型的分類結果分析

撲浪語料		真實類別	Bernoulli	Multinomial
1	今天被<稱讚>非常<開心>	P	P	P
2	要下班囉好累明天繼續	N	M	M
3	快要被冷<死了>	N	N	N
4	在 give 板贈物真是亂<有效率>，屢試<不爽>	P	N	N
5	陰陰的天氣<適合>睡覺，<不適>合上班!	N	P	N

負面情緒召回率；整體而言 Bernoulli 模型和 Multinomial 模型的平均分類準確率和平均 F1 分數沒有顯著差異，差距不超過 1%。

表七是分別擷取 Bernoulli 模型和 Multinomial 模型的部分分類結果，<開心>表示此中文詞彙「開心」出現在 NTUSD 辭典中，P 代表正面情緒類別，N 代表負面情緒類別，而 M 代表此測試語料中的所有詞彙皆沒有出現在 NTUSD 辭典中，使得分類無法進行因此標記為 M，並列入拒絕率的計算中。例句 1 中，「稱讚」和「開心」都出現在 NTUSD 辭典中且出現在正面情緒的語料中的機率較高，使得兩種機率模型都正確分類為正面；例句 2 中，所有的詞彙都沒有出現在 NTUSD 辭典中，造成此案例沒有特徵無法分類標記為 M，若情緒詞彙辭典中包含「好累」，就可以增加分類為負面情緒的可能性。例句 3 雖然正確分類，但透過辭典的斷詞不夠準確，正確的斷詞應該為「冷死了」而不是「死了」，造成語意不正確但分類正確的情形。例句 4 是錯誤分類的情況，原本可能的特徵「屢試不爽」屬於正面詞彙，但辭典中沒有「屢試不爽」這個詞彙卻有負面情緒詞彙「不爽」，造成機率模型分類成負面情緒。例句 5 中，因為 NTUSD 辭典沒有「不適合」這個詞彙，只有「不適」這個詞彙出現在負面詞彙集中，使得斷詞的結果為<不適>；另外，因為下列數學式(12)和(13)中的條件機率 $P(t_k | e)$ 和 $P(b_k | e)$ 的估計值不同造成 Bernoulli 模型分類為正面情緒而 Multinomial 模型分類為負面情緒的不一致情況。

$$\frac{P(e=\text{正面})P(b_k=\text{適合}|e=\text{正面})P(b_k=\text{不適}|e=\text{正面})}{P(e=\text{負面})P(b_k=\text{適合}|e=\text{負面})P(b_k=\text{不適}|e=\text{負面})} > 1 \quad (12)$$

$$\frac{P(e=\text{正面})P(t_k=\text{適合}|e=\text{正面})P(t_k=\text{不適}|e=\text{正面})}{P(e=\text{負面})P(t_k=\text{適合}|e=\text{負面})P(t_k=\text{不適}|e=\text{負面})} < 1 \quad (13)$$

4.2.3. Hybrid 法

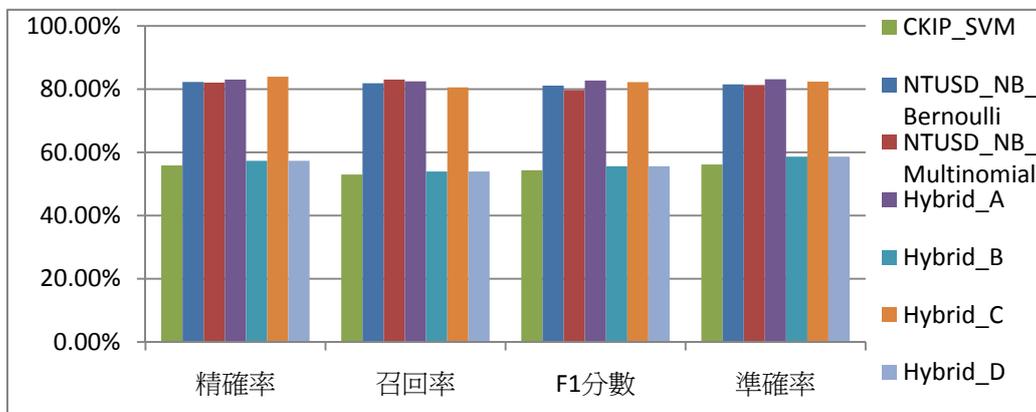
在 NTUSD 情緒詞彙辭典搭配 naïve Bayes 模型中，分別實作了 Bernoulli 和 Multinomial 兩種機率模型，且輸出的預測值又分為機率分數(實數值)和類別分數(離散數值 0 或 1)，因此在混合模型中需要估計的最佳權重 λ 共有四種排列組合。

表八混合模型中，類別分數的準確率的分類效果明顯優於機率分數，原因在於類別分數的預測值為離散的數值 0 或 1，類別的界線模糊性相對較小，使得分類的效果較好。

表八 混合模型的分類效果

名稱	機率模型	分數種類	門檻T'	權重 λ	精確率	召回率	F1 分數	準確率
Hybrid_A	Bernoulli 模型	類別分數	0.5	0.4179	83.00%	82.44%	82.72%	83.13%
Hybrid_B		機率分數	0.3	0.4231	57.33%	53.93%	55.58%	58.65%
Hybrid_C	Multinomial 模型	類別分數	0.5	0.3738	83.92%	80.55%	82.20%	82.34%
Hybrid_D		機率分數	0.3	0.4231	57.33%	53.93%	55.58%	58.65%

註：上表的拒絕率皆為 48.82%；精確率、召回率、F1 分數和準確率皆為簡單平均結果。



圖七 不同模型分類準確度比較

因為 CKIP_SVM 法無拒絕率而 NTUSD_NB 法有拒絕率 48.82%，所以表八的四組 Hybrid 法的拒絕率與 NTUSD_NB 法相同。圖七可以發現以分類準確率而言，NTUSD_NB 法中的 Bernoulli 模型的 82.28%、Multinomial 模型的 82.05% 稍微低於 Hybrid 法中的類別分數的 Bernoulli 模型的 83.13%、Multinomial 模型的 82.34%；除了召回率外，Hybrid_A 和 Hybrid_C 的分類精確率、F1 分數和準確率都些微優於 NTUSD_NB 法。另外，因為拒絕率不同的關係，CKIP_SVM 法和 NTUSD_NB 法無法直接進行比較。CKIP_SVM 法的分類準確性為 56.15%，此實驗可以再更加深入地探討實驗參數對於 CKIP_SVM 法實驗結果的影響，更加詳細的實驗請參考[3]。

5. 結論

本研究透過撲浪語料(Plurk)探討網路上中文短句的情緒分類問題試圖偵測使用者的情緒狀態屬於正面或負面，並且採用具有明顯正面情緒和負面情緒的表情符號作為情緒分類的標記。情緒詞彙的剖析上，分別使用中研院的斷詞系統搭配 PMI 特徵選擇和 NTUSD 情緒詞彙辭典兩種方法作為情緒詞彙特徵，再分別使用支援向量機(SVM)模型和 naïve Bayes 機率模型進行情緒分類；另外，賦予 SVM 模型和 naïve Bayes 機率模型兩種分類方法各自的權重組成混合情緒分類模型。本研究有效率的降低特徵詞彙的數量，並且能在語境有限的微網誌的短句利用三種不同的分類模型精準的分類正面和負面情緒。

從實驗數據和 NTUSD 辭典的詞彙可以發現使用者在撲浪上發表短文、短句時，使用負面情緒的詞彙的次數和多元性明顯比較高，使得負面情緒案例數較高且造成負面情緒的分類較準確。另外平均而言，圖七中 Hybrid_A 和 Hybrid_C 都有不錯的分類效果，但拒絕率高達 48.82%影響了本實驗的實質分類效果；高拒絕率的因素來自於 NTUSD 辭典的詞彙數量不足夠，造成把短句轉換為特徵向量時，成為一個空的向量，迫使 NTUSD_NB 法 Bernoulli 和 Multinomial 模型無法分類，因此使用情緒詞彙辭典進行斷詞並作為文件分類案例的特徵(feature)的分類效果的優與劣和辭典中的詞彙數量和品質有密切的關係，如同表七例句 2 的說明；解決高拒絕率的方法，可以透過在未來的實驗增加訓練語料量的大小並擴充辭典詞彙數量或者透過處理零機率的 smoothing 相關技術降低零向量出現的機會。為增進 CKIP_SVM 法中分類的整體效果，未來可以嘗試 point-wise mutual information 法以外的 co-occurrence relation 技術改善特徵詞彙和情緒類別之間配對的正確性，以及精進中文斷詞方法找出更好的情緒詞彙邊界。

中文詞彙和網路用語的演化與發展，造成許多的新詞彙的出現也是分類準確率的挑戰，同時如何自動化維持詞庫和辭典也是未來的研究問題。在本研究中，無法正確分類擁有不同語意的相似詞彙短句，未來可以透過文法剖析器或本體論的應用來降低短句的語意模糊性問題。

致謝

我們感謝三位匿名評審的寶貴意見，雖然我們一時無法在有限的頁數之內回應所有意見，但是仍將在未來工作中，努力進行評審所提出的建議工作。承蒙國科會計畫案 NSC-97-2221-004-007、NSC-98-2815-C-004-004-E 與 NSC-99-2221-004-007 的補助，謹此致謝。

參考文獻

- [1] 中央研究院中文斷詞系統，<http://ckipsvr.iis.sinica.edu.tw/> [2010/06/25]
- [2] 哈爾濱工業大學訊息檢索實驗室同義詞詞林擴充版，
http://www.nlp.org.cn/docs/doclist.php?cat_id=9&type=7 [2010/06/20]
- [3] 劉峻杰，*應用支援向量機於以文字為基礎的情緒分類*，國立清華大學資訊工程系，碩士論文，2009。
- [4] 撲浪，<http://www.plurk.com/> [Online]
- [5] J. C. Ang, *Prosodic Cues for Emotion Recognition in Communicator Dialogs*. M.S. thesis, University of California Berkeley, 2002.

- [6] D. B. Bracewell, "Semi-Automatic Creation of an Emotion Dictionary Using WordNet and its Evaluation," *Proceedings of the IEEE conference on Cybernetics and Intelligent Systems*, pp. 21-24, 2008.
- [7] C. Cortes and V. Vapnik, "Support-Vector Network," *Machine Learning*, pp. 273-297, 1995.
- [8] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion Recognition in Human-Computer Interaction," *IEEE Signal Processing Magazine*, 18(1), pp. 32-80, 2001.
- [9] L. Devillers, I. Vasilescu, and L. Lamel, "Annotation and Detection of Emotion in a Task-Oriented Human-Human Dialog Corpus," *Proceedings of the ISLE Workshop on Dialogue Tagging for Multi-Modal Human-Computer Interaction*, 2002.
- [10] H. T. Lin and C. J. Lin, *A Study on Sigmoid Kernels for SVM and the Training of Non-PSD Kernels by SMO-Type Methods*, Technical report, Department of Computer Science & Information Engineering, National Taiwan University, 2003.
- [11] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2009.
- [12] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, The MIT Press, 1999.
- [13] NTUSD (National Taiwan University Semantic Dictionary)
<http://nlg18.csie.ntu.edu.tw:8080/opinion/pub1.html> [2010/06/20]
- [14] R. W. Picard, E. Vyzas, and J. Healey, "Toward Machine Emotional Intelligence: Analysis of Affective Physiological State," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10), pp. 1175-1191, 2001.
- [15] J. Read, "Using Emoticons to Reduce Dependency in Machine Learning Techniques for Sentiment Classification," *Proceedings of the ACL Student Research Workshop*, The Forty-Third Annual Meeting of the Association for Computational Linguistics, 2005.
- [16] J. A. Russel, "A Circumplex Model of Affect," *Journal of Personality and Social Psychology*, 39(6), pp. 1161-1178, 1980.
- [17] Z. Teng, F. Ren, and S. Kuroiwa, "Recognition of Emotion with SVMs," *International Journal of Innovative Computing, Information and Control*, pp. 701-710, 2006.
- [18] C.-H. Wu, Z.-J. Chuang, and Y.-C. Lin, "Emotion Recognition from Text Using Semantic Labels and Separable Mixture Models," *ACM Transactions on Asian Language Information Processing*, 5(2), 2006.
- [19] J. Yang, D. B. Bracewell, F. Ren, and S. Kuroiwa, "The Creation of a Chinese Emotion Ontology Based on HowNet," *Engineering Letters*, 16(1), pp. 166-171, 2008.
- [20] 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.