國立臺北大學電機工程學系專題報告

建構情緒分析系統-以車電產業為例

Leveraging Text mining and Sentiment Analysis to Increase Vehicle Sales

組員：曾柏諺

指導老師：林嘉淦 老師

執行期間：2014 年 9 月至 2015 年 6 月

1. 摘要

每一樣產品都會有優缺點，因此消費者的心中會因為產品的優缺點產生評價。企業為了要獲取更大的利潤，通常要透過收集這些評價資訊，進而使商品可以做進一步的改善。在傳統時代，人們購買商品的評價，大多透過鄰居傳達到整個社區，企業難以透過有效且便利的方式取得消費者對於該產品的評價，企業因此失去改善產品的機會。近年來，由於網路科技發達，使得消費者習慣更容易從各種不同管道抒發自己對於產品的看法，如撰寫在線上網站，或是透過填寫商品的回饋單，甚至是透過客服信箱進行客訴等等。其中企業最難以掌握則是部落格、臉書、論壇等這些散落在網路上的資訊，這些資訊又間接影響到消費者是否購買產品的意願。其中，在消費者的眾多評論中，有大多數的消費者會從中留下各種情緒句子，像是喜歡、討厭等形容詞。在上述的情形中，每一個句子都有隱含正面、中立、負面評價的情緒字眼。在過去的研究中，大多探討如何提升斷字斷句的精準度，少有從系統架構的角度進行研究。因此在本篇研究中，將結合文字探勘與情緒分析的做法，建構一套情緒分析系統，設法在大量的中文字中以正負情緒作為主要分析方向，並透過資料視覺化進行呈現。並以知名車輛品牌為例，進行品牌、型號、銷售量評價三個角度分析，作為未來企業了解消費者產品評價參考。

關鍵字：評價、情緒句子、情緒字眼、文字探勘、情緒分析

2. 簡介

從 2004 開始，學者 Hu 和 Liu (2004)就發明了五元素(quintuple)的基本方法，將一句話分成五個特點。具體、特徵、觀點、發言人、時間這五大要素，可做成概要的分類。五元素這樣的方式，Liu 和 Ch (2010)也寫在 NLP(Neuro-Linguistic Programming 神經語言規劃)的書中。

每個情緒字眼都有正負面及中立去代表那句話，但是用人工方式在大量資料下是非常辛苦的。假如能讓電腦做判斷，那必定可以省出很多時間跟人力。因此有學者使用非監督式和監督式學習為最主要兩種方式。此兩種最主要差別為：非監督式讓機器透過演算法，使看似雜亂無章的數據資料，能變成有明確意義的資訊；而監督式學習會建立一個模型，之後再以其他資料去比對找到其結果。再來說明學者 Pang et al(2002)在論文中提出監督式學習方法，這是一個機器學習的方法，可以由訓練資料中學到或建立一個模型，並依此模式推測新的實例。因此訓練資料便是最重要的步驟之一。在收集夠多樣本的時候，便可以將之後拿到的數據直接做推測。監督式方法有 Naive Bayes、Maximum entropy law 以及 Support vector machine。除了這兩種方法，還有半監督式學習。半監督學習則是介於兩者之間的學習技術，它同時利用有標記樣本與無標記樣本進行學習，此種學習方式 PU learning。

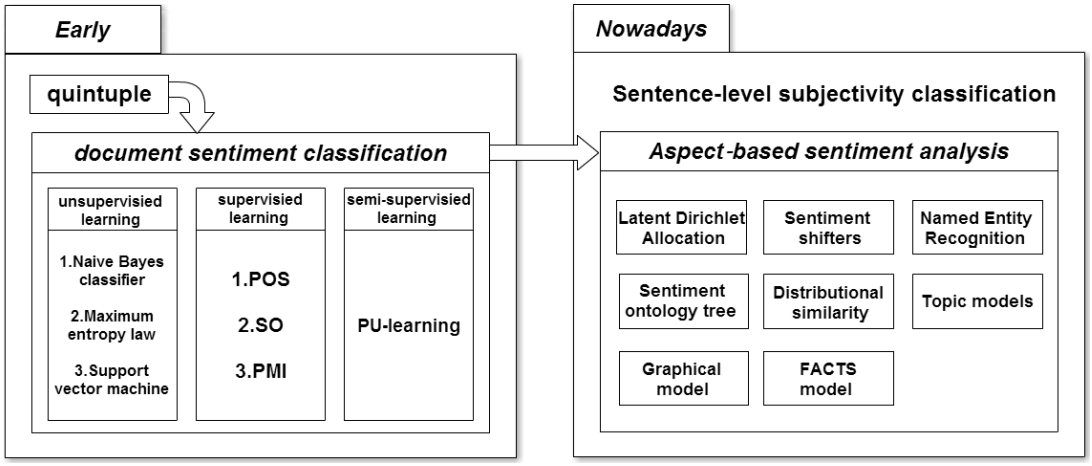
從上述都方法都可以歸類為文本情緒分類(document-level sentiment classification)。然而這樣的作法太過粗糙簡略，因此學者在近年來發展句子情緒分類(sentence-level sentiment classification)，將句子分析的更透徹。

學者認為將一句話能夠分辨正負情緒是不夠的，這並不能詳細描述商品哪邊的缺失，因此從五元素發展屬性基礎情緒解析法(Aspect-based sentiment

analysis)，從物品中的特徵去探討細節的部份。從屬性基礎情緒解析所遇到的問題就是網路上抓取留言作分析。就一般留言來說，我們很容易去分析人們對商品好壞的留言，然而如果是在部落格以及論壇的話，常常會出現不相關的訊息還有廣告來妨礙去分析。因此學者用 Latent Dirichlet Allocation (Blei, Ng & Jordan, 2003;Griffiths & Steyvers, 2003; 2004)、Sentiment shifters (Polanyi and Zaenen 2004)和 Named Entity Recognition (David Nadeau, 2007)這三種方法，使的分析留言後容易對情緒詞彙辨別。除此之外以上三種方法外，還有 Sentiment ontology tree (Wei and Gulla, 2010)、Distributional similarity (Li et al 2010, Zhang and Liu, 2011)、Topic models (Sauper, Haghighi and Barzilay, 2011)也是由學者者提出來的。甚至在 2011 年的時候，也出現了 Graphical model 和 FACTS model.

從 2008 至近幾年，又有研究者提出建立情緒詞彙基礎方法(Lexicon-based sentiment analysis) 以及提出一些方法來匯整意見(Opinion summarization)。建立情緒辭彙有下列四種方式，分別是 Lexicon-based approach (Ding, Liu and Yu, 2008)、Double Propagation (Qiu et al 2009)、Corpus-based approaches (Hazivassiloglou and McKeown, 1997; Turney, 2002; Yu and Hazivassiloglou, 2003; Kanayama and Nasukawa, 2006; Ding, Liu and Yu, 2008)、Dictionary-based methods (Hu and Liu, 2004; Kim and Hovy, 2004;Kamps et al 2004)。

另外關於匯整意見的部分，有 Informativeness and Readability (資訊量和可讀性)和 Summarization using an ontology (Lu et al. Coling-2010)。最後，其他議題像是意見比較(Comparative Opinions)、評論預測(Review rating prediction)和意見垃圾偵測(Opinion spam detection)都是近十年來最熱門的議題之一。

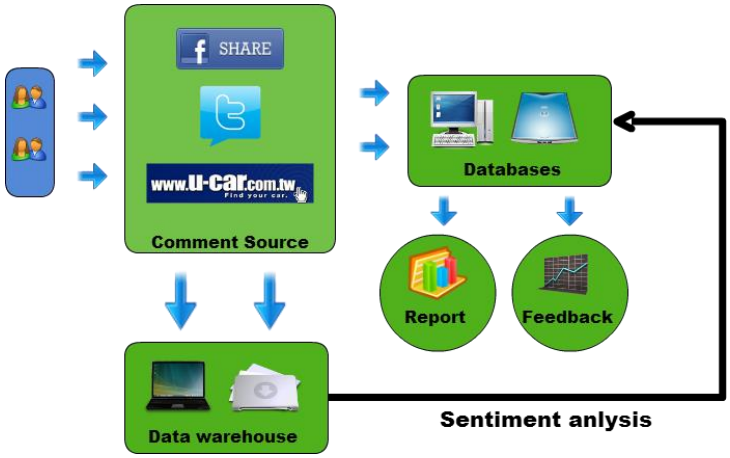


圖一、情緒分析之方法演進過程

3. 專題進行方式

步驟 1： 消費者將自身使用過商品的經驗寫在臉書、推特、論壇等。透過搜尋引擎搜尋關鍵字，並且找出網頁連結將文章、評論等一切重要資訊儲存在資料倉儲。

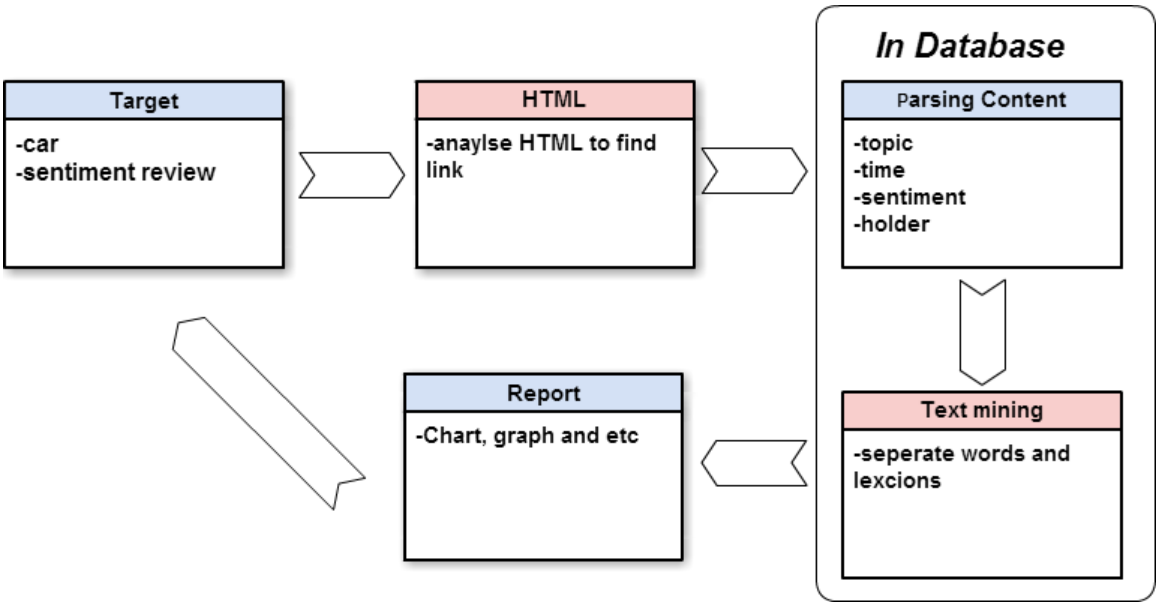
步驟 2：透過程式幫忙將句子切開並與取出情緒字詞和時間綁定，把這些句子拆成字或詞的來比對。其中有用到 Hidden Markov Model 的概念去執行。如果沒有字或詞彙，可以到目錄資料夾去新增。當完成這些工作，開始尋找情緒性的字眼，並且將他們把時間綁定，再來拿這些資料做統計和分析，最後儲存在資料庫當中，做成報告和回饋來反映目前情況。



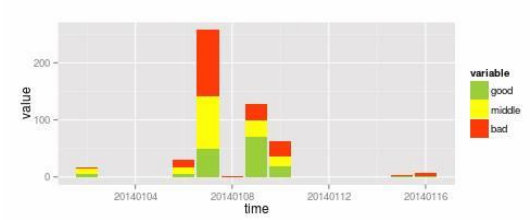
圖二、操作程序之流程圖

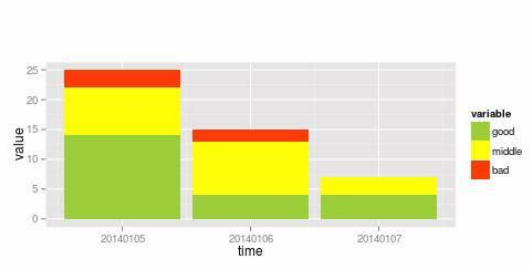
(Hidden Markov Model 是統計模型，它用來描述一個含有隱含未知參數的馬爾可夫過程。其難點是從可觀察的參數中確定該過程的隱含參數。然後利用這些參數來作進一步的分析，例如模式識別。)

4. 主要成果與評估



圖三、本篇報告支架構圖



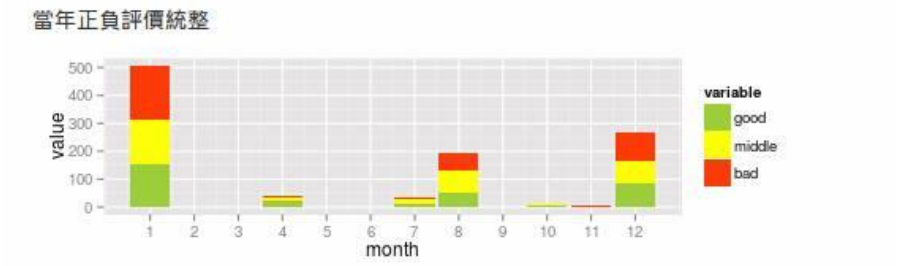


圖四　　圖五

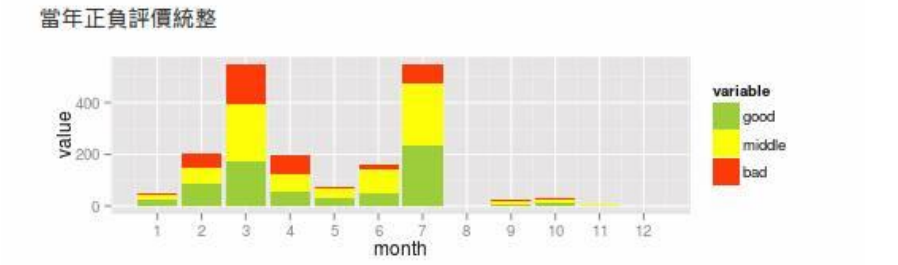
2014 年 1 月份關於納智捷文章分析(正負評價堆疊圖)(圖四)

2014 年 1 月份關於豐田文章分析(正負評價堆疊圖)(圖五)

X 軸為月份，Y 軸為數量，紅色為負面、黃色為中立、綠色為正面評價。這是由 1 月份某幾天消費者有評論的資料，可以看出納智捷在 1 月 7 日這天負面詞彙數量大於正面，但在 1 月 9 日這天正面又大於負面。而豐田的部分在 1 月 5 日、6 日、7 日都是正面都大於負面評價較多的。



圖六

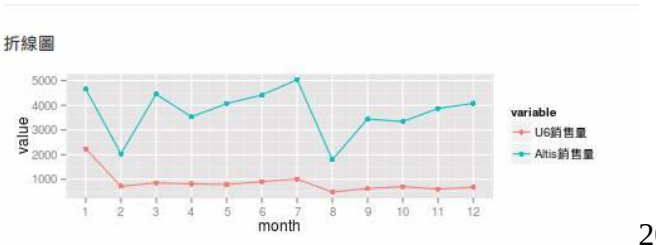


圖七

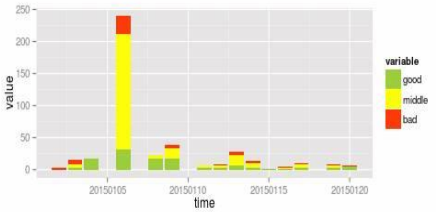
2014 年關於納智捷文章分析(整年正負評價堆疊圖)(圖六)

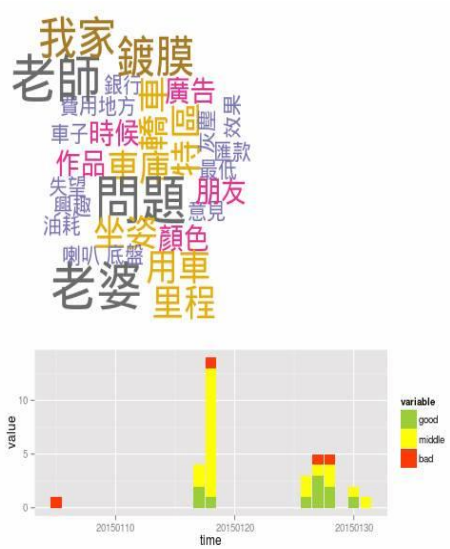
2014 年關於豐田文章分析(整年正負評價堆疊圖)(圖七)

X 軸為月份，Y 軸為數量。由圖可知，納智捷在 1 月及 12 月份的評論比較多，三個區塊其實差不多，但負面稍微偏多；再來豐田的評論集中在 1 月到 7 月，由 3 月到 7 月來看，明顯消費者是正面評價居多。



2014 整年車型銷售量折線圖(圖十)





圖十一　　圖十二

2015 年 1 月分關於 Altis 文章分析(文字雲+堆疊圖)(圖十一)

2015 年 1 月分關於 U6 文章分析(文字雲+堆疊圖)(圖十二)

由左圖可以看到文字雲中，油耗的字顯得特別大，可推得消費者對於 Altis 的油耗不滿意。再來是右圖中，老師、老婆、我家、鍍膜等幾個詞彙顯得比較大，只能猜測老師或者是有家庭主婦想買車的想法，還有關於鍍膜、顏色等耐用及美觀問題，也是考量之一。如果廠商參考此圖，可以在當月的時間推出教師或主婦優惠，使的消費者更容易購買，也可以對於舊車重新鍍膜或是烤漆給予一翻優惠來吸引消費者上門。

5. 結語與展望

從以上研究成果中，可以發現到這些方法的重點都是要嘗試從大量的文字中，辨識出正面與負面等情緒，好讓這些結果產生出商業價值，這也是口碑評價對於現在公司來說非常重要的一環，公司該在消費者評價這一部份的管理層面上多出一份力。

事實上，還是有許多問題需要被解決，包含像是特殊的嘲諷句子，如：「這手機很棒，我從來都無法開機。」以及像是中文文法、日文文法等等問題，因此做的精準做得好，還需要投入大量的資源與成本。更重要的是跨語言領域，包含找中文系、英文系等語言專家，加入一些修辭學的概念等等，甚至是加入語音、圖像等多元資料，因為在大數據的時代講究的是跨領域的整合，從文字上解析還不夠，還需要更多元的資料，才能做得更好。

6. 銘謝

感謝林嘉泠教授給我跟康富城一起到華創公司學習的機會，並且還認識了吳柏翰學長。在教授及學長的教導下，每星期都能接觸更新的資訊，覺得自己又學到的更多的東西和知識。生長在大數據時代，嘗試處理這些資料是一件非常新鮮的事情，也在寫程式當中，康富城幫助我非常多來完成這次的專題，真的非常謝謝你們。

7. 參考文獻

[1]　Bing Liu 著，2012，” Sentiment Analysis and Opinion Mining”。

[2]　J. Li, A. Najmi, R. M. Gray, Image classification by a two dimensional hidden Markov model, IEEE Transactions on Signal Processing, 48(2):517-33, February 2000.