

社群網站上之政治活動分析

以台灣 2016 年總統選舉為例

蒲郁文

國立交通大學資訊工程學系

106A 學士班專題競賽

摘要

本研究嘗試使用關鍵字擷取（keyword extraction）、情感分析（sentiment analysis）、社會網絡分析（social network analysis）等方法，分析 2016 年台灣總統選舉競選期間臉書各大政治性粉絲專頁之公開資料，探討使用者於社群網站上之政治活動的模式。本研究發現此次選舉有 26.63% 的社群媒體貼文是攻擊型競選，與台灣、中國或中國國民黨有關的議題是人們最關注的焦點，且特定族群會傾向於分享特定來源的資訊。同時，本研究也提供了後續研究者一套有效地自動化分析線上政治活動的方法。

關鍵字 — 社群媒體、自動化內容分析、社會網絡分析、數位行動主義、政治參與

研究問題

- ① 各政黨／候選人／意見領袖分別發佈了多少篇貼文？
每日的貼文數量隨著時間有什麼樣的改變？
- ② 各政黨／候選人／意見領袖最關心的議題是什麼？
人們關注的焦點隨著時間有什麼樣的改變？
本次選舉人們最關心什麼？
- ③ 哪些政黨／候選人／意見領袖最常使用攻擊型競選？
什麼時候人們最常使用攻擊型競選？
本次選舉有多少比例的貼文是攻擊型競選？
- ④ 各發文者與其分享的資訊的來源有什麼樣的關係？
特定族群是否會傾向於分享特定來源的資訊？

資料蒐集

- 資料集：2016 年台灣總統選舉競選期間臉書各大政治性粉絲專頁之公開資料
- 時間範圍：投票日前三個月至投票日止
- 資料總數：18,967 筆臉書貼文，來自 135 個台灣熱門的政治性粉絲專頁
- 蒉集方法：
 - 使用 Facebook Graph API
 - 參考 Socialbakers 及 Likeboy 提供之社群媒體統計數據，輔以我個人的自身經驗，決定欲蒐集成的粉絲專頁清單
 - 查詢這些粉絲專頁最新的按讚數，只保留擁有超過一萬個讚的粉絲專頁

- 採用中央研究院資訊科學研究所中文詞知識庫小組所研發之中文斷詞系統
 - 對台灣本地的文本有較好的表現
 - 具有良好的未知詞辨識能力
- 斷詞前先對貼文進行預處理
- 使用物件關係對映與資料庫存放資料

研究問題

- ① 各政黨／候選人／意見領袖分別發佈了多少篇貼文？
每日的貼文數量隨著時間有什麼樣的改變？
- ② 各政黨／候選人／意見領袖最關心的議題是什麼？
人們關注的焦點隨著時間有什麼樣的改變？
本次選舉人們最關心什麼？
- ③ 哪些政黨／候選人／意見領袖最常使用攻擊型競選？
什麼時候人們最常使用攻擊型競選？
本次選舉有多少比例的貼文是攻擊型競選？
- ④ 各發文者與其分享的資訊的來源有什麼樣的關係？
特定族群是否會傾向於分享特定來源的資訊？

總體分佈

- 使用 pandas、Matplotlib、Plotly 等函式庫
- 繪製貼文數量對時間的關係圖
- 繪製累計貼文數量對時間的關係圖
- 統計各粉絲專頁分別有多少筆貼文、佔全部貼文的百分比

總體分佈



Figure 1: 貼文數量對時間的關係圖

貼文數量呈現週期性變化，每逢假日貼文數量就會顯著降低。隨著投票日的接近，各政黨／候選人／意見領袖也愈來愈躍躍地在社群網站上發佈貼文。

總體分佈

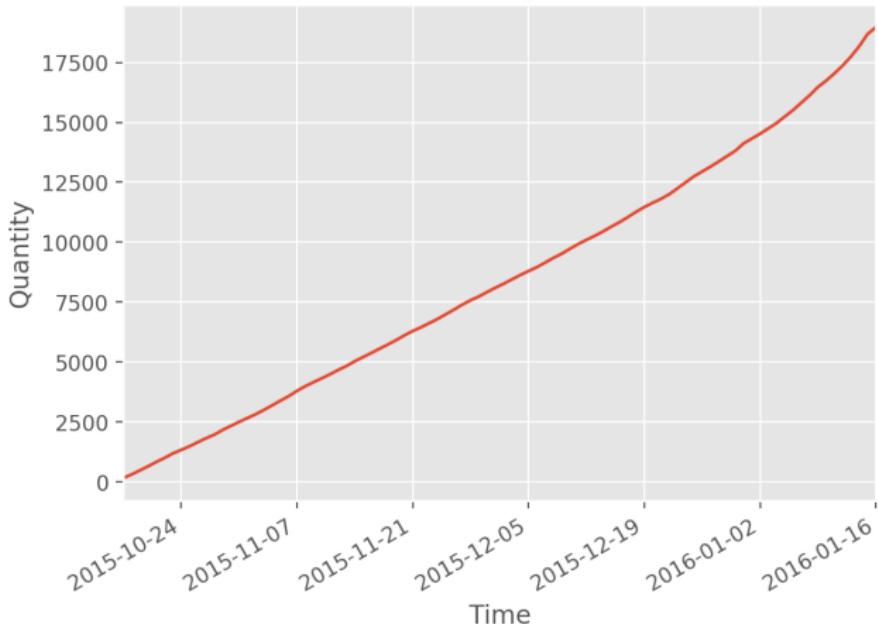


Figure 2: 累計貼文數量對時間的關係圖

Table 1: 各粉絲專頁的貼文數量（節錄）

粉絲專頁名稱	貼文數量	百分比
柯建銘	485	2.56%
Taiwan Fugue	361	1.90%
台灣新力量 / 羅致政...	337	1.78%
宋楚瑜找朋友	334	1.76%
反馬英九聯盟	330	1.74%
林昶佐 Freddy Lim	320	1.69%
基進黨（基進側翼）	319	1.68%
BillyPan 潘建志醫師	306	1.61%
宋楚瑜	301	1.59%
綠黨	297	1.57%
民主進步黨	293	1.54%

研究問題

- ① 各政黨／候選人／意見領袖分別發佈了多少篇貼文？
每日的貼文數量隨著時間有什麼樣的改變？
- ② 各政黨／候選人／意見領袖最關心的議題是什麼？
人們關注的焦點隨著時間有什麼樣的改變？
本次選舉人們最關心什麼？
- ③ 哪些政黨／候選人／意見領袖最常使用攻擊型競選？
什麼時候人們最常使用攻擊型競選？
本次選舉有多少比例的貼文是攻擊型競選？
- ④ 各發文者與其分享的資訊的來源有什麼樣的關係？
特定族群是否會傾向於分享特定來源的資訊？

關鍵字擷取

- 使用 SnowNLP 函式庫實作的 TextRank 演算法
 - 衍生自 PageRank 演算法
 - 不需輸入全部文本 (cf. tf-idf)
- 改良 SnowNLP 函式庫提供的停用字清單
- 合併範圍內的貼文進行分析
- 分析各粉絲專頁發佈的貼文的關鍵字
- 分析各時間區間發佈的貼文的關鍵字
- 分析整個競選期間所有人發佈的貼文的關鍵字

關鍵字擷取

Table 2: 依作者劃分（節錄）

粉絲專頁名稱	#1	#2
朱立倫	台灣	新
大安推范雲	范雲	台灣
王丹网站 Wa...	中國	中
國民黨青年團	青年	台灣
呂欣潔 Jennif...	政治	欣潔
管碧玲 (kuan...	管媽	市長
段宜康	朱立倫	國民黨
鄉民實業坊	黑子	種
我是台灣人 (...	台灣	綠黨
林智堅	新竹市	新竹
BillyPan 潘...	醫師	台灣

從此表我們可以看出各政黨／候選人／意見領袖重視的議題。例如，中國國民黨關注朱立倫以及經濟、兩岸等事務；王金平關注國會、改革等事務；綠黨關注勞工、環境等議題；而馬躍·比吼則關注原住民的文化和土地等等。

關鍵字擷取

Table 3: 依時間劃分（節錄）

時間區間	#1	#3
11/15 ~ 11/22	台灣	政府
11/08 ~ 11/15	台灣	中國
11/01 ~ 11/08	台灣	中國
10/25 ~ 11/01	台灣	國民黨
10/17 ~ 10/25	台灣	新

透過此表，我們可以發現人們關注的焦點皆圍繞在台灣、中國、國民黨上，唯有在 11 月 1 日至 11 月 8 日間，由於舉行兩岸領導人會面，馬英九、馬習會也躍升為社群網站上的討論焦點。

關鍵字擷取

Table 4: 合併所有貼文（節錄）

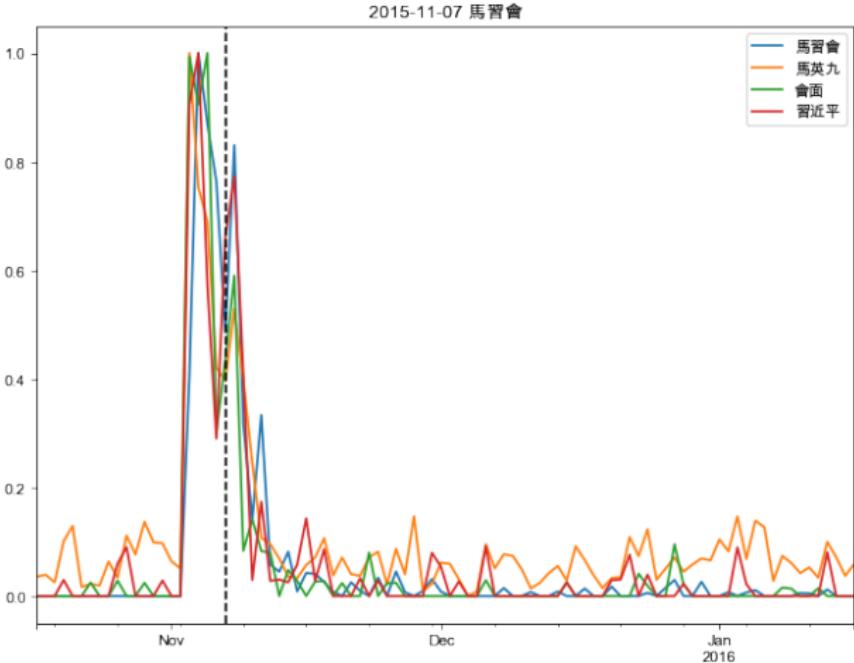
#1	#3	#7
台灣	國民黨	中國

此表告訴我們本次選舉人們最關心跟台灣和中國有關的議題，而國民黨也是人們討論的焦點。值得一提的是，此次選舉的另一大黨，民主進步黨，並未出現在前十大關鍵字之中。

關鍵字擷取

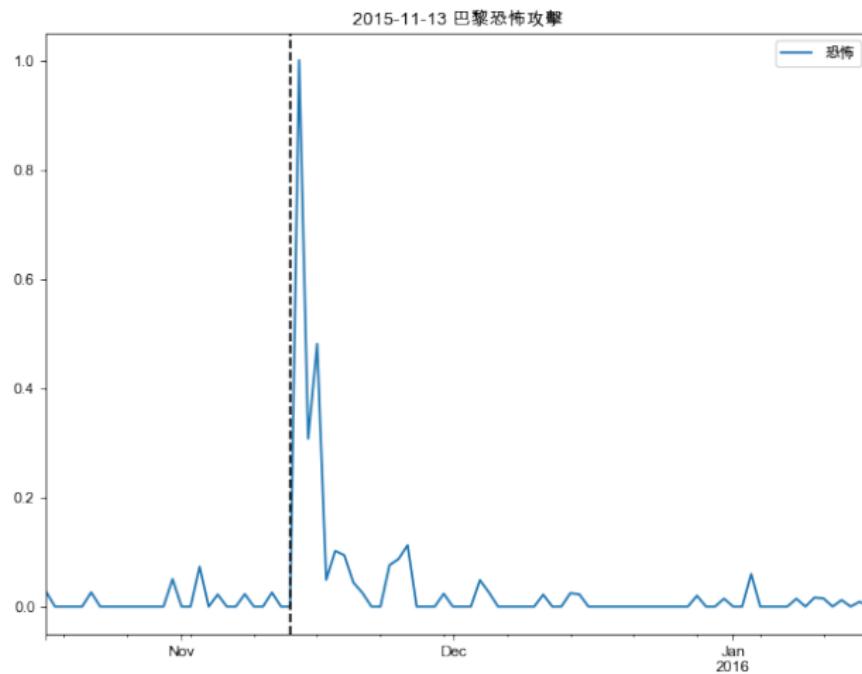
- 另一種分析各時間點之熱門議題的方法（謝明恩等人，2016）
- 尋找出現次數有明顯高峰的單詞
 - 找出所有出現超過一百次的單詞
 - 統計每個單詞每日的出現率
 - 以「平均值 + 2 × 標準差」為門檻，找出連續三日出現率高於門檻的單詞
- 找出在相近時間點出現高峰的單詞，組成詞組
- 從詞組反推事件
- 將事件中的每個單詞的出現率正規化，並繪圖

關鍵字擷取

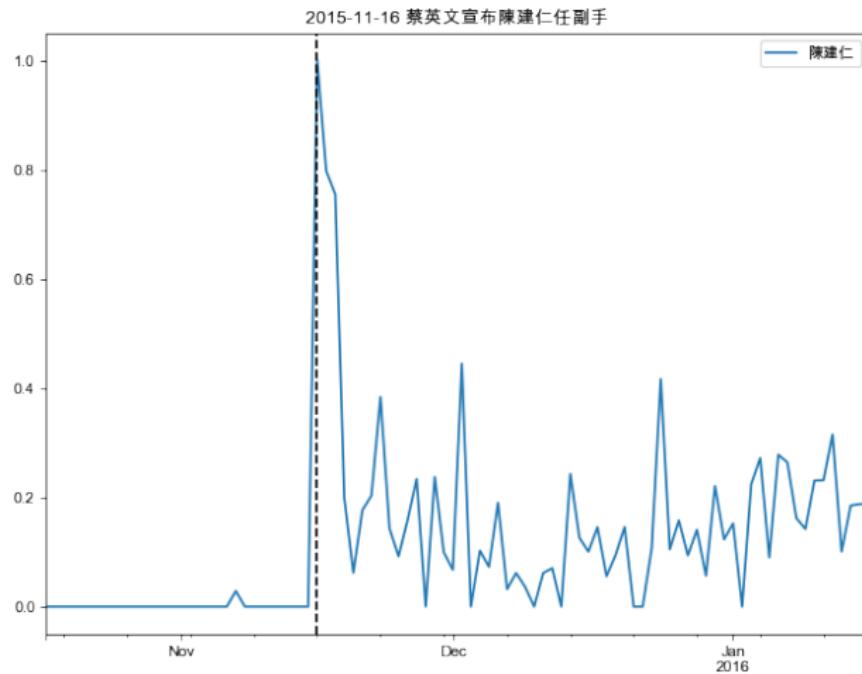


竄升單詞：領導人
總統／監督／國際
傷害／現狀／中國
抗議／憲政／報告
交流／簽署／尊嚴
會面／見面／協議
鬥陣／舉行／前往
安排／交換／地位
表／黑箱／馬習會
馬／馬習／馬英九
屬／聲明／習近平
握／停止／新加坡

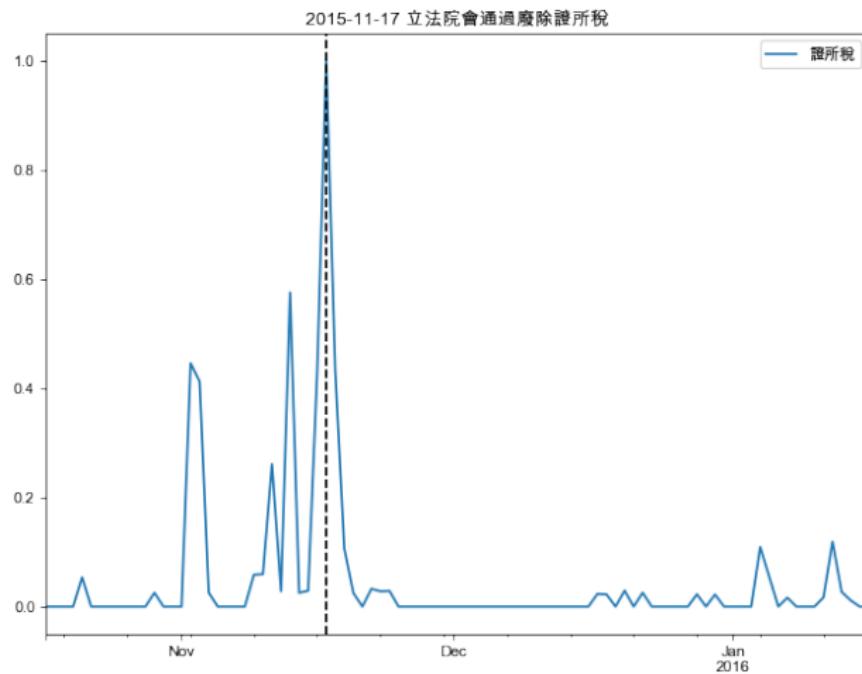
關鍵字擷取



關鍵字擷取

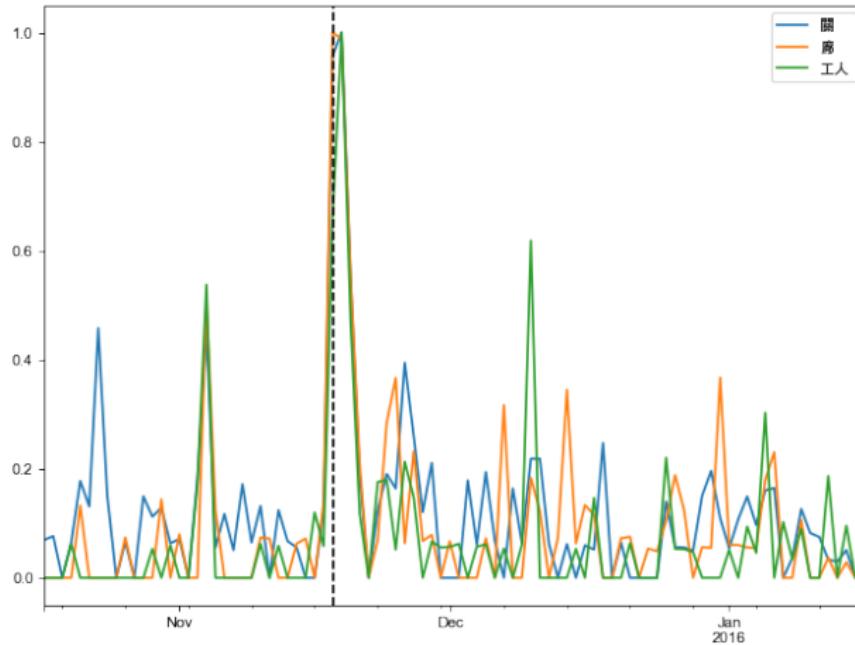


關鍵字擷取

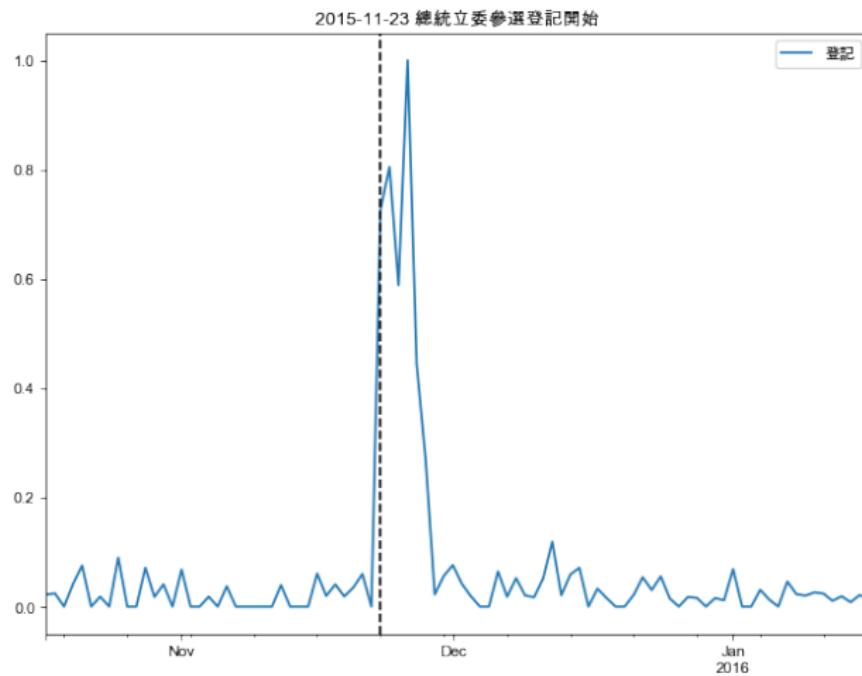


關鍵字擷取

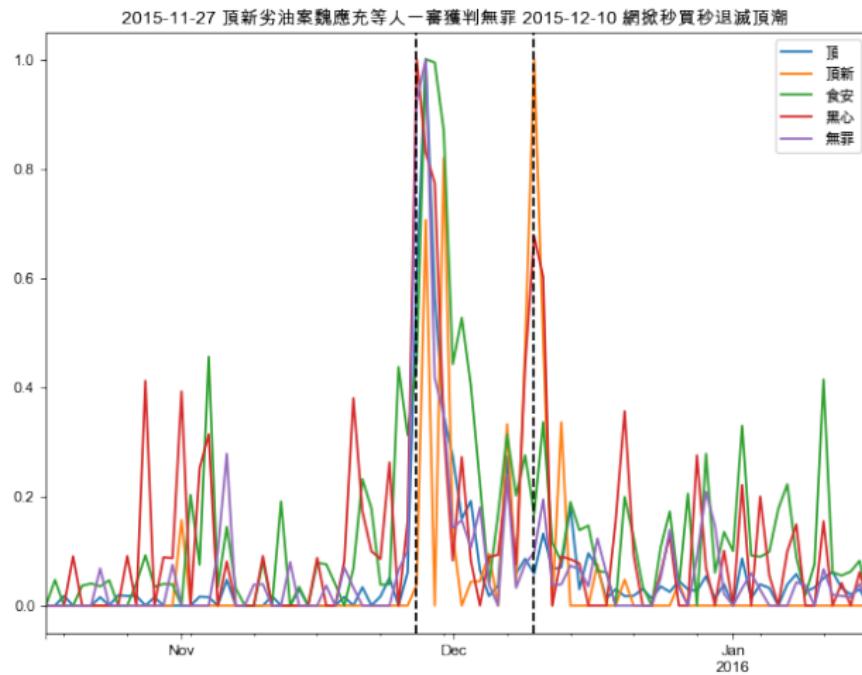
2015-11-18 朱立倫宣布王如玄任副手，關廠工人案爭議延燒



關鍵字擷取

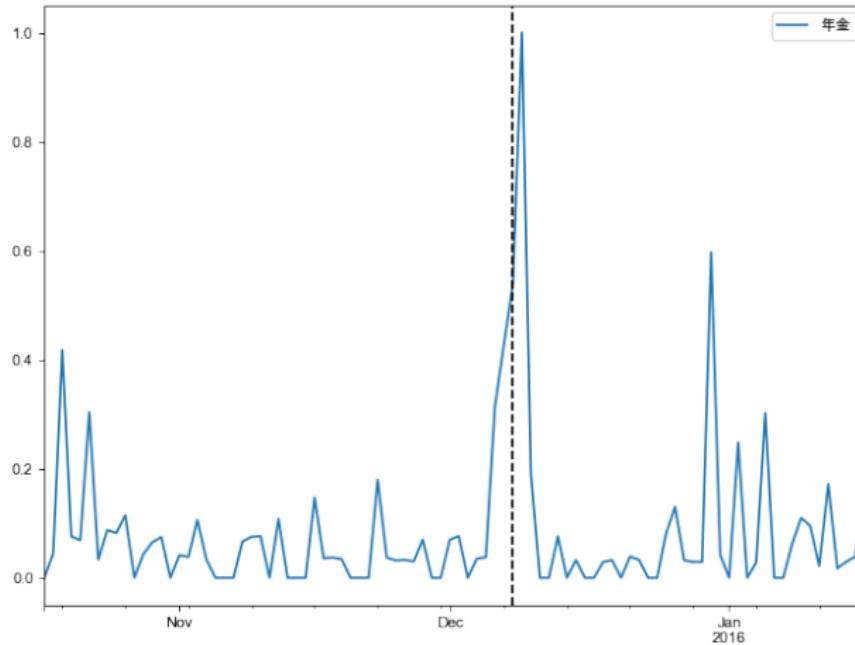


關鍵字擷取

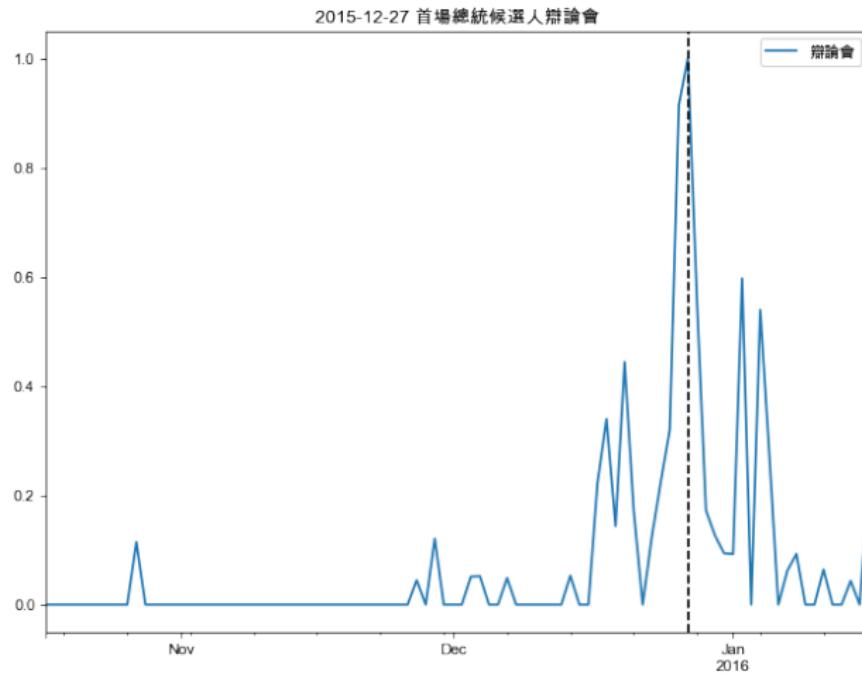


關鍵字擷取

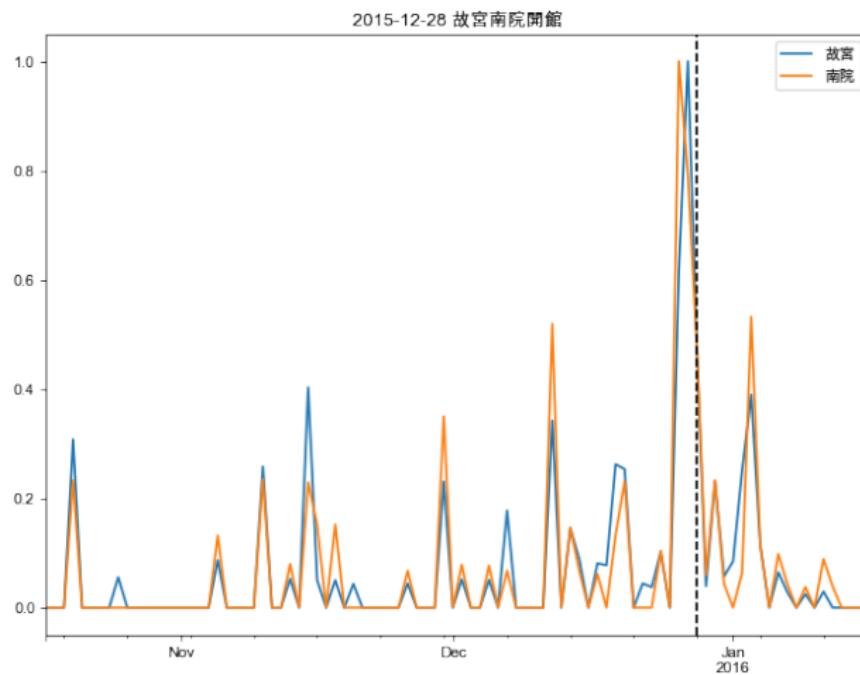
2015-12-08 三黨候選人出席公務員座談，談年金改革



關鍵字擷取

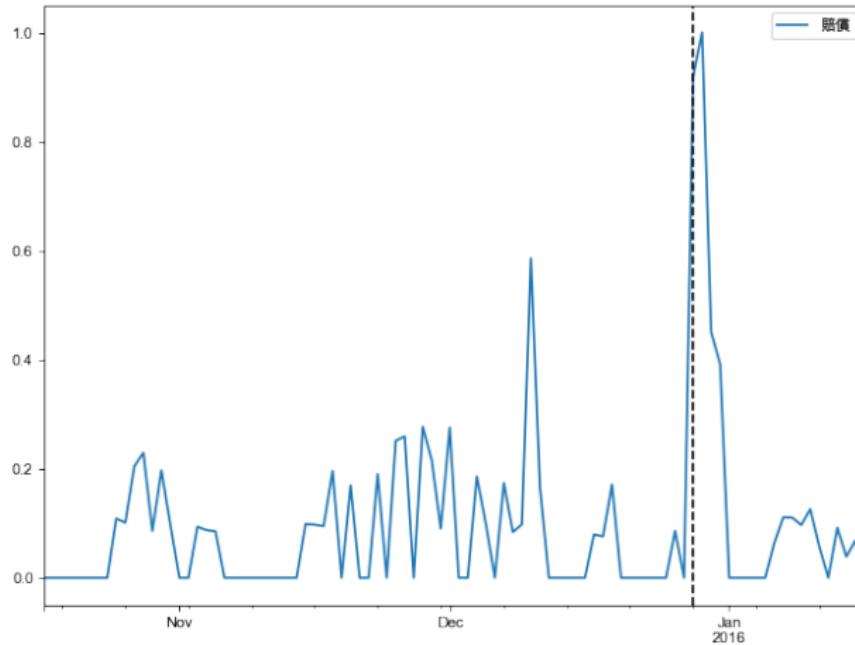


關鍵字擷取



關鍵字擷取

2015-12-28 日韓就慰安婦議題達成協議，日方賠償



研究問題

- ① 各政黨／候選人／意見領袖分別發佈了多少篇貼文？
每日的貼文數量隨著時間有什麼樣的改變？
- ② 各政黨／候選人／意見領袖最關心的議題是什麼？
人們關注的焦點隨著時間有什麼樣的改變？
本次選舉人們最關心什麼？
- ③ 哪些政黨／候選人／意見領袖最常使用攻擊型競選？
什麼時候人們最常使用攻擊型競選？
本次選舉有多少比例的貼文是攻擊型競選？
- ④ 各發文者與其分享的資訊的來源有什麼樣的關係？
特定族群是否會傾向於分享特定來源的資訊？

- 將全部貼文分為兩個類別
 - 攻擊型競選：貼文內容是以批評反對陣營為主、貼文重點是在批評反對陣營
 - 說服型競選：非屬攻擊型競選的貼文（包含中立貼文）
- 分析每筆貼文的情感，再進行統計
- 自行實作單純貝氏分類器
 - 二元化多項單純貝氏（binarized multinomial naive Bayes）模型
 - 只管一個單詞在一篇文章中有沒有出現，不管該單詞在文中究竟出現了多少遍
- 均勻地選取 948 筆（約 5%）貼文作為訓練與測試資料，人工標記情感類型
- 使用十等分交叉測試來評估模型的表現

情感分析

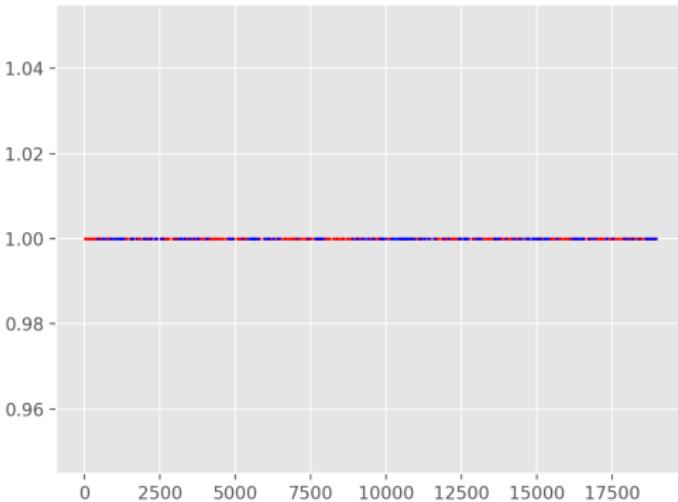


Figure 3: 訓練／測試貼文的分佈圖

圖中的每個點皆代表一筆訓練／測試貼文，點的水平位置代表它在所有貼文中的位置。紅點代表說服型競選貼文，一共有 651 筆；藍點代表攻擊型競選貼文，一共有 297 筆。

將攻擊型競選定義為陽性，本情感分類模型的準確度 (accuracy) 為 86.82%，查準率 (precision) 為 85.25%，查全率 (recall) 為 70.29%，F 度量 (F-measure) 為 76.75%。

情感分析

Table 5: 貼文情感分析（節錄）

粉絲專頁名稱	01/10	01/03	...	10/17	所有時間區間
			...		
	01/17	01/10		10/25	
朱立倫	1/30	1/18	...	1/5	10/210
大安推范雲	2/35	0/27	...	1/13	14/239
王丹網站 Wa...	3/14	6/11	...	5/12	37/137
國民黨青年團	1/4	0/4	...	0/0	4/51
:	:	:	:	:	:
宋楚瑜	7/57	4/51	...	1/22	32/294
特急件小周...	19/25	13/20	...	4/6	99/142
所有粉絲專頁	509/2355	446/1868	...	336/1275	4827/18128

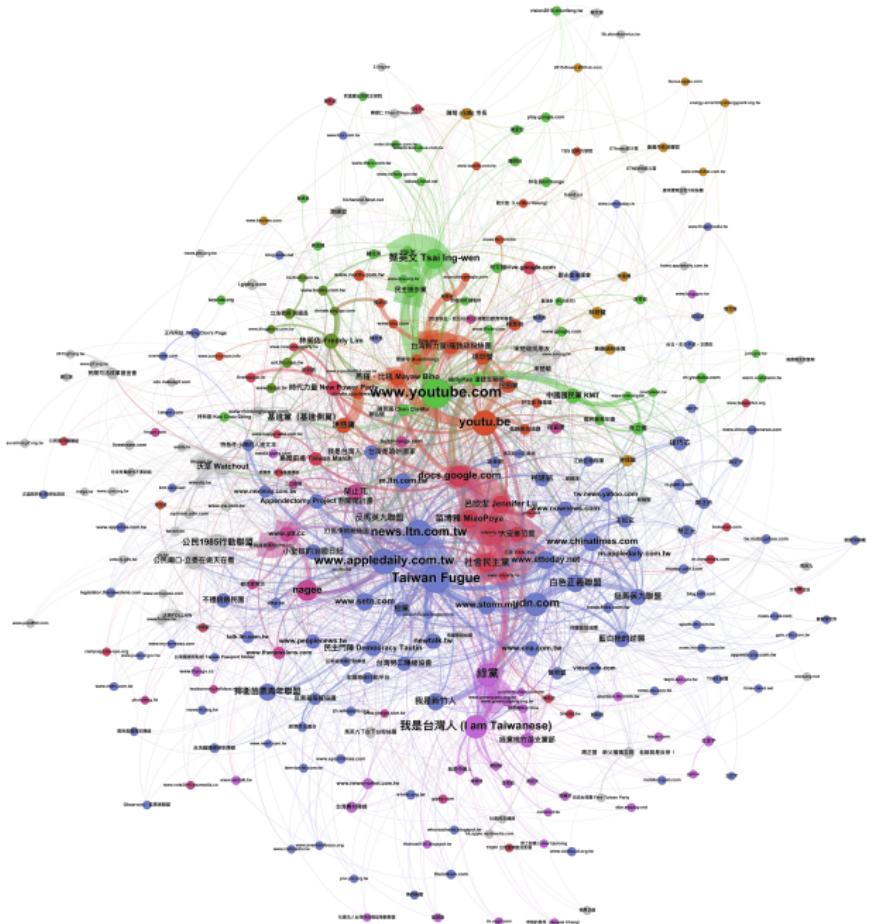
此次選舉只有 26.63% 的社群媒體貼文是攻擊型競選。許多政治人物都不常使用攻擊型競選，然而，具鮮明政治立場，或以抨擊反對陣營為目的的意見領袖，則較常使用攻擊型競選。Taiwan Fugue、反馬英九聯盟、不禮貌鄉民團、藍白拖的逆襲等粉絲專頁攻擊型競選貼文的比例皆超過八成。這也有可能是政治人物為了營造正面競選的形象所進行的分工。

使用攻擊型競選的比例並未隨著時間而有太大的變化，唯有在 11 月 1 日至 11 月 8 日間，或許是因為舉行兩岸領導人會面的緣故，攻擊型競選貼文的比例比平時高出了一成左右。另外，在投票日前一週，使用攻擊型競選的比例則比平時低了 5% 左右。

研究問題

- ① 各政黨／候選人／意見領袖分別發佈了多少篇貼文？
每日的貼文數量隨著時間有什麼樣的改變？
- ② 各政黨／候選人／意見領袖最關心的議題是什麼？
人們關注的焦點隨著時間有什麼樣的改變？
本次選舉人們最關心什麼？
- ③ 哪些政黨／候選人／意見領袖最常使用攻擊型競選？
什麼時候人們最常使用攻擊型競選？
本次選舉有多少比例的貼文是攻擊型競選？
- ④ 各發文者與其分享的資訊的來源有什麼樣的關係？
特定族群是否會傾向於分享特定來源的資訊？

- 探討社群網站上的「分享」網絡
- 資料預處理
 - 找出所有文本中的連結
 - 還原短網址
 - 找出臉書連結導向的粉絲專頁
 - 捨棄「自己分享自己的貼文」的紀錄
- 繪製各發文者與其分享的貼文／連結的來源之網絡圖
 - 採用 Gephi 網絡視覺化軟體
 - Force Atlas 佈局演算法
 - 模組性 (modularity) 分析
 - HITS 網絡連結分析
 - 只保留分支度不小於 3 的節點



本研究從全部貼文中，一共整理出 12,428 筆分享紀錄。上圖為貼文／連結的「分享」網絡圖。從此圖我們可以發現一些規則，例如，nagee 較常分享蘋果日報及批踢踢的消息；Taiwan Fugue 較常分享自由時報及蘋果日報的消息；白色正義聯盟較常分享聯合新聞網的消息；我是台灣人 (I am Taiwanese) 較常分享綠黨的消息等等。

另外，此圖也呈現出各政黨／候選人／意見領袖之間的關係。例如苗博雅、呂欣潔和范雲皆屬於同一個政黨，因此他們在圖中也屬於同一個社群。同樣地，白色正義聯盟、挺馬英九聯盟和藍白拖的逆襲彼此有相近的政治立場，因此他們在圖中的位置也比較相近。

Table 6: 權威粉絲專頁排名

Label	Authority
Taiwan Fugue	0.245268
我是台灣人 (I ...	0.194027
白色正義聯盟	0.172589
nagee	0.170947
綠黨	0.166116
挺馬英九聯盟	0.164113
呂欣潔 Jennif...	0.161191
民主鬥陣 Dem...	0.158304
藍白拖的逆襲	0.155271
社會民主黨	0.154146
反馬英九聯盟	0.151556

Table 7: 匯集網站排名

Label	Hub
www.youtube.com	0.324420
news.ltn.com.tw	0.288759
www.appledaily...	0.288031
udn.com	0.276527
youtu.be	0.263457
docs.google.com	0.199894
www.setn.com	0.196341
www.storm.mg	0.194338
www.ettoday.net	0.188077
www.chinatimes...	0.187592
m.ltn.com.tw	0.162871

從 HITS 的分析結果，或從圖中圓點的大小，我們還可以看出哪些消息來源最常被人們引用。在此次選舉中最常被人們引用的消息來源是 YouTube、自由時報、蘋果日報和聯合新聞網等等。從這裡我們或許也能看出各消息來源在使用者心中的可信度。

重要參考文獻

-  A. O. Larsson and H. Moe, "Studying political microblogging: Twitter users in the 2010 swedish election campaign," *New Media & Society*, vol. 14, no. 5, pp. 729–747, 2012.
-  W. Zhang, T. J. Johnson, T. Seltzer, and S. L. Bichard, "The revolution will be networked," *Social Science Computer Review*, vol. 28, no. 1, pp. 75–92, 2010.
-  B. Hosch-Dayican, C. Amrit, K. Aarts, and A. Dassen, "How do online citizens persuade fellow voters? Using twitter during the 2012 dutch parliamentary election campaign," *Social Science Computer Review*, vol. 34, no. 2, pp. 135–152, 2016.
-  謝明恩, 林立秦, 張逸群, and 蔡明娟, "P T T 軍情室," *2016 Data Mining Final Project Slides*, no. 24, 2016.

閱讀全文：

<http://ywpu.me/APASNS.pdf>