

# 基於社交媒體的政治情感分析：美媒於總統選舉前後 對中態度調查

呂宣誼 (B07303084)<sup>†</sup>, 邱威諭 (B07303037)<sup>†‡</sup>, 楊尚霖 (B07303116)<sup>†</sup>,  
赫謙 (B07303118)<sup>†</sup>, and 郭宇杰 (B07611039)<sup>‡</sup>

<sup>†</sup> 台大經濟系

<sup>‡</sup> 台大資管系

## 摘要

本文爬梳了美國媒體於推特上與中國有關的報導，並分析於選舉前後媒體對待中國的態度是否有所差異。實驗結果顯示，選舉前後媒體報導對中國態度有所差異，且各間媒體呈現相似的趨勢：選後對中國的態度較為平緩；選舉前後媒體所報導與中國有關的主題也不盡相同，如選舉前「香港」是常見的報導主題，選後則較少出現於報導中，而「台灣」則較常見於選後的報導。

**Keywords**— 美中關係、隱藏語意分析、情感分析、社群媒體分析

## 1 專案動機

美國與中國已然是當今世界上最強勢的兩大國家，兩大強權之間的關係皆是各國矚目的焦點。自二十世紀開始，中美之間的關係從戰爭同盟國、從支持中華民國政府轉為與中華人民共和國政府建交，到現今彼此防範彼此，美中關係的變化受到其領導人的風格與國家路線影響甚鉅，媒體的態度能於一定程度上反映出國家對於某項議題的想法與關注度，尤其是針對政治議題，例如當局政府相當親中時，媒體可能就會傾向於報導美中關係的建立；若當局政府相當厭中，則媒體可能會產出較多兩國衝突的報導。在資訊產生及傳播相

當快速的當代社會，資訊往往係透過媒體的反芻消化後才進到閱聽人的眼裡耳中，媒體對待某個議題的態度對於資訊便顯得十分重要。因此我們好奇，美國媒體在選舉前後，面對兩個對待中國態度截然不同的候選人與政府，針對中國的態度是否會有所不同。

本文的架構如下：段落 2 爬梳了社交媒體與政治情感分析領的相關研究；段落 3 介紹我們如何設定資料範圍、資料爬取方法、資料的前處理；段落 4 整理了我們所做的研究與實驗，包含研究方法、實驗結果等。段落 5 在文末總結這份研究。

## 2 文獻回顧

針對社群媒體與情感分析有許多的研究和文獻著墨於上，例如最早 [1] 首次使用 Twitter 平台進行政治情感分析，並發現 Twitter 確實可以用來反映選舉結果，進而預測選舉。隨著機器學習技術的革新以及計算資源的突破，有更多的方法能夠使用在這類型的分析上。[2] 設計了一種結合 Twitter 流量和情感分析的模型進行選舉結果預測，並使用文檔中的單詞作為向量，訓練支持向量機 (Support Vector Machine, SVM) 分類器對文本進行情感分析；[3] 採用遞歸神經網路 (Recurrent Neural Networks, RNN) 與長短期記憶模型 (Long Short-Term Memory, LSTM) 的方法進行政治情感分析與分類；而 [4] 則將

詞嵌入 (Word Embeddings) 融合卷積神經網路 (Convolution Neural Networks) 對 Twitter 的資料進行選舉預測分類。多數政治情感分析的研究都是著墨於選舉結果，[5] 的研究則是揭露了政治傾向、正負字詞與候選人間的關聯性認知影響，啟發了我們的研究方向。除了分析方法的不同外，[6] 分析了美國的政治光譜與主流媒體的傾向，[7] 調查美國民眾對中的好感度，我們也參考 [6, 7] 所得出的結果分析美國媒體的對中態度。

## 3 資料蒐集與評估

### 3.1 資料蒐集

為了預估民眾態度可能會轉變的時間點，我們首先整理西元 2019 至 2021 年之間發生的中美事件，如 2020 年 3 月 2 號時，美國國務院宣布將對五家中國官方媒體機構實施雇員；2021 年 2 月 11 日時拜登上任總統後首次和中國國家主席習近平通話等。基於上述時間段，我們將關鍵字設為「China」，抓取 2020 年至 2021 年（美國總統大選前後各 10 個月）美國人主流新聞媒體推特上的推文，最後選定 ABC、NBCnews、CNN、CBSNews、nytimes、WSJ、FoxNews 七間媒體進行分析。

### 3.2 資料前處理

我們使用推特開放的高級搜索 API<sup>1</sup> 訪問目標推特帳號的貼文並抓取目標媒體的帳號 ID、推文內容、日期與網址。我們將資料的不同欄位進行若干前處理，如將 API 爬下來後的預設日期轉為格式化日期、推文內容去除對分析沒有用的標點符號（如 [\*] \$ @ % +=）等等。

## 4 實驗設計

我們首先根據 [7] 的結果針對資料進行假設：

**假設 1** 美國媒體於選前選後的對中態度會有所差異。

<sup>1</sup><https://twitter.com/search-advanced>

**假設 2** 美國媒體於選前選後關注的對中議題並不相同。

另外，我們也想知道我們分析出來的媒體態度是否符合 [6] 的調查，因此我們設定了另一個假設：

**假設 3** 各家不同立場之媒體的報導風格與主題皆不同。

基於這三個假設，我們設計了一系列的實驗來分析資料是否符合我們的假設。

### 4.1 對中態度的情感分析

#### 4.1.1 模型與套件

挑選適合的模型與套件對於資料分析而言是相當重要的前置作業。我們分析 TextBlob<sup>2</sup>、flair<sup>3</sup>、twitter-roberta-base-sentiment<sup>4</sup> 三個常用的情感分析套件，發現 TextBlob 太過粗糙，將不在語料庫內的字都視為沒有情感；flair 又太過敏感，結果幾乎都是落在光譜的兩端，最後我們選定使用支援表情符號，且將表情符號納入分析語料中的 twitter-roberta-base-sentiment 套件。而處理 Word Embeddings 的工具，我們則是使用較為靈敏與準確的 bert-base-multilingual-uncased-sentiment<sup>5</sup>，而非上課提及的 Keras BERT<sup>6</sup>。

#### 4.1.2 統計圖表與討論

透過圖表反應資料趨勢是常見的資料視覺化手法，我們製作不同媒體報導資料的情感分數與頻率之間的直方圖，取 NBCNews 的結果為例，呈現如圖 1。其中上半部是選舉前的頻率分佈直方圖，下半部則是選舉後的頻率分佈直方圖。

由於七間媒體選前選後的情感分數頻率趨勢相似，選前情感分數低的推文頻率較高、選後則分佈較為平均，因此我們有以下推論：

<sup>2</sup><https://textblob.readthedocs.io/en/dev/>

<sup>3</sup><https://github.com/flairNLP/flair>

<sup>4</sup><https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment>

<sup>5</sup><https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment>

<sup>6</sup><https://github.com/CyberZHG/keras-bert>

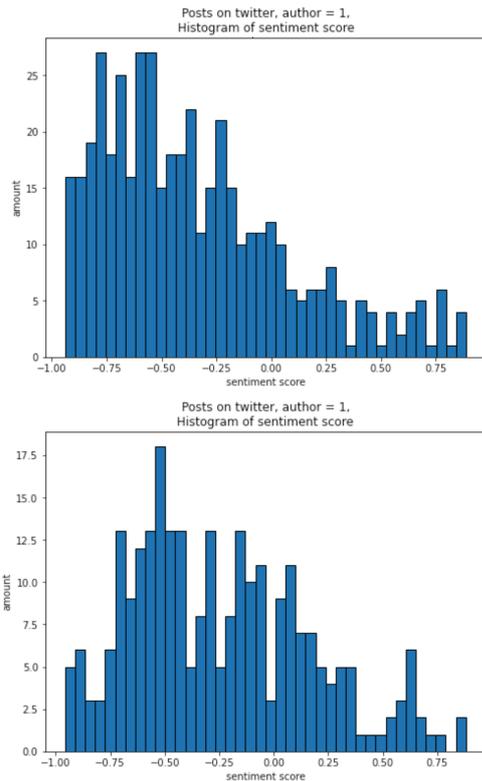


圖 1: NBCNews 情感分數頻率的直方圖

**推論 1** 選舉前後媒體對中態度有影響，且選後態度較為中立。

緊接著，我們也製作了隨著時間推進，不同媒體報導推文情感分數的點狀圖，如圖 2。圖 2 的 T-value 揭示了結果的顯著性，而根據其選舉前後的情感分數趨勢也論證了推論 1。而針對這般結果，我們試圖給予合理的解釋。

**解釋 1** 我們認為，由於民眾普遍不喜歡總統候選人川普，因此並不會特別關注其所支持的反中相關議題。且 2020 與 2021 年間適逢疫情，選舉期間所關注的中國議題應有部分是醫療相關。

## 4.2 不同媒體的報導風格與對中態度

選舉前後媒體對中態度是否不同，段落 4.1 已經給出了肯定的答案，而我們接著想繼續探討選舉前後，不同媒體的報導主題是否一致，且是否如假設 3 所述，不同立場媒體間的風格與主題會有所不同。我們採用分群演算法的方式，企圖將所有的報導依據媒體、對中態度等分到不同的群集，

惟效果非常不好，幾乎所有的資料都被分到同一個群集中。

**解釋 2** 由於蒐集到的資料本身就有偏誤，資料本身過於相似，分群演算法很容易將所有資料直接視為同一群。

仔細思考後能發現，這個結果其實並不驚奇，因為我們的資料是在推特上蒐集與中國有關的報導，資料本身可能就已經存在偏誤，因此透過分群演算法在嘗試分群時，由於資料的相似度太高，所有的資料很容易會被視為同一群。

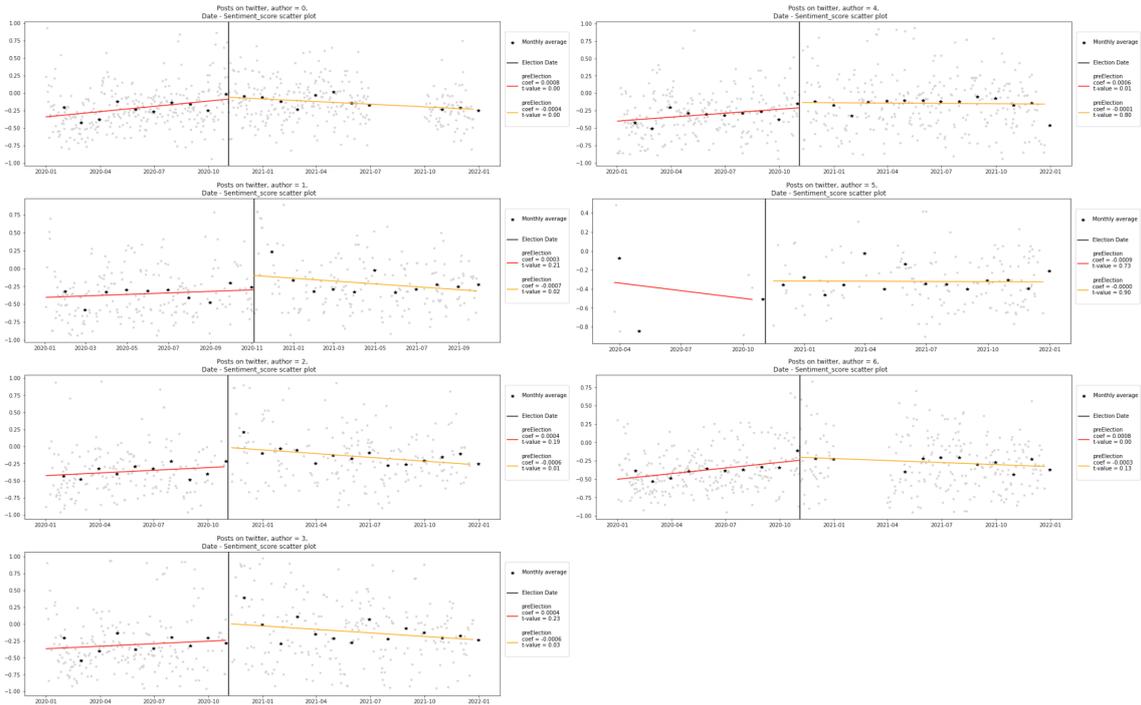
## 4.3 選舉前後的文字意涵

在得到段落 4.2 的結果後，我們嘗試從其他角度切入，判斷媒體於選舉前後討論的議題是否真的不相同。一個簡單的分析方式是透過 TFIDF 計算選舉前後資料的 Cosine 相似度，以此判斷資料間的關係。惟我們的資料在使用 TFIDF 時，向量中容易有大量的元素賦值為 0，因此得到的相似度也無法給出進一步的答案。

之後我們改為使用 Word Embeddings，針對 {選舉前、選舉前}、{選舉前、選舉後}、{選舉後、選舉後} 三對資料進行 Cosine 相似度分析，得到的 Cosine 相似度在各間媒體間都有超過 0.7 的結果。

**推論 2** 不論是選前選後，同一間媒體報導的用詞用字，以及發文習慣、風格等並沒有太大的差異。

推論 2 給了我們繼續延伸研究的靈感，倘若選舉前後並不會影響同一間媒體報導的習慣用字與風格，那麼實作一個分類模型，並將選前的資料作為訓練集訓練，給予選後的報導語料，模型應該要能根據寫作習慣與風格等，判斷出該篇報導係來自哪間媒體。我們將選前的報導的 80% 作為訓練集、剩下的 20% 作為驗證集，訓練一個深層神經網絡 (Deep Neural Networks) 分類模型，並將選後的報導資料作為測試集，以此驗證推論 2。圖 3 顯示了分類的準確度。然而，最終訓練出來的模型準確度不如我們想像的高。



此圖為時間自 2020 至 2021 年間各間媒體推文報導的情感分數散佈圖，以黑線標示美國總統選舉日期，黑線以左代表選舉前，以右則為選舉後。紅線代表選舉前的推文報導情感分數趨勢，黃線則代表選舉後的推文報導情感分數趨勢。

圖 2: 報導推文情感分數之時間序列點狀圖

Pre-election		Train   80%		Valid   20%	
Validation (on preElection) information is shown below:					
	precision	recall	f1-score	support	
0	0.80	0.83	0.81	277	
1	0.62	0.70	0.66	89	
2	0.85	0.75	0.79	67	
3	0.75	0.66	0.70	120	
4	0.65	0.68	0.66	183	
5	0.00	0.00	0.00	4	
6	0.60	0.59	0.60	158	
accuracy			0.71	898	
macro avg	0.61	0.60	0.60	898	
weighted avg	0.71	0.71	0.71	898	

Post-election		Test   100%			
Testing (on postElection) information is shown below:					
	precision	recall	f1-score	support	
0	0.72	0.74	0.73	825	
1	0.39	0.36	0.38	263	
2	0.35	0.46	0.40	265	
3	0.40	0.43	0.41	324	
4	0.52	0.44	0.48	503	
5	1.00	0.03	0.06	211	
6	0.49	0.67	0.57	447	
accuracy			0.53	2838	
macro avg	0.55	0.45	0.43	2838	
weighted avg	0.57	0.53	0.51	2838	

在驗證集時，模型分類的準確度可以達到 0.71，而使用測試集測試時，分類的準確度則掉落至 0.53。

圖 3: 深層神經網絡分類模型的結果

**推論 3** 使用選前資料訓練的模型無法準確地分類選後的資料，說明選前選後，媒體的報導之間仍有某些語意或報導主題的落差。

為了找出使選前選後媒體分類不準確的肇因，我們萃取出選前與選後的語料中最常見的幾項主題。前後對照後發現，選前與選後的報導中，僅有約 43%，不到一半的主題重疊。舉例來說，選前常見的主題有「WuHan」、「Hong Kong」等，而選後這些主題出現的頻率下降，轉而「Taiwan」、「Biden」等主題較多出現在報導內（見圖 4）。

```
importantPre = findTopN_more_freq(TFIDF_vectorizer.get_feature_names(), TFIDF_vec_news_pre, topN)
importantPre

array(['china', 'coronavirus', 'new', 'outbreak', 'chinas', 'people',
       'cases', 'wuhan', 'trump', 'hong', 'kong', 'chinese', 'president',
       'virus', 'health', 'said', 'world', 'pandemic', 'trade', 'death',
       'says', 'officials', 'country', 'spread', 'latest', 'countries',
       'state', 'beijing', 'reported', 'global'], dtype=<U19>)

importantPost = findTopN_more_freq(TFIDF_vectorizer.get_feature_names(), TFIDF_vec_news_post, topN)
importantPost

array(['china', 'biden', 'chinas', 'covid', 'chinese', 'new', 'says',
       'country', 'beijing', '19', 'president', 'said', 'world', 'taiwan',
       'year', 'xi', 'russia', 'space', 'climate', 'global', 'people',
       'government', 'years', 'military', 'administration', 'jinping',
       'state', 'foreign', 'analysis', 'origins'], dtype=<U19>)
```

圖 4: 選舉前與選舉後常見的報導主題

**解釋 3** 由於選前選後資料的主題性重疊性太低，因此就算使用深層神經網路，訓練出來的模型依舊無法根據媒體風格等準確分類選後的報導。

自此，我們的實驗結果說明了在選舉前與選舉後，媒體的報導行為的確是有差異。

## 5 結論

本文的研究重點旨在探討美國總統候選人對中態度的不同是否會影響選舉前後美國媒體報導對中國的態度。我們透過情感分析的模型計算出媒體報導的情感分數，並找出時間趨勢，證明選舉前後美國媒體對中國的態度的確有所差異；而我們也透過計算文本相似度、萃取出選前選後的報導主題異同等方法說明了選舉前後美國媒體與中國有關的報導主題並不相同。基於我們研究的觀察，未來可以繼續分析台灣選舉前後台灣媒體對中國的態度是否與美國媒體有類似的行為。

## 參考資料

- [1] Andranik Tumasjan, Timm Sprenger, Philipp Sandner, and Isabell Welp. Predicting elections with twitter: What 140 characters reveal about political sentiment. volume 10, 01 2010.
- [2] Adam Bermingham and Alan Smeaton. On using Twitter to monitor political sentiment and predict election results. In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*, pages 2–10, Chiang Mai, Thailand, November 2011. Asian Federation of Natural Language Processing.
- [3] Saurabh Dorle and Nitin Pise. Political sentiment analysis through social media. In *2018 Second International Conference on Computing Methodologies and Communication (IC-CMC)*, pages 869–873, 2018.
- [4] Xiao Yang, Craig Macdonald, and Iadh Ounis. Using word embeddings in twitter election classification, 2017.
- [5] Zirun Qi, Veda Storey, and Wael Jabr. Sentiment analysis meets semantic analysis: Constructing insight knowledge bases. 12 2015.
- [6] 伍治堅. 美國媒體的政治傾向. <https://kknews.cc/zh-tw/media/lknpkz9.html>, 2018.
- [7] Reagan National Defense Survey. Reagan national defense survey. [https://www.reaganfoundation.org/media/358085/rndf\\_survey\\_booklet.pdf](https://www.reaganfoundation.org/media/358085/rndf_survey_booklet.pdf), December 2021.