



社群媒體分析

第三次讀書會報告_第十組

N124320007 藍筱琦

N124320012 鄭義璋

N124320029 黃靖紋

N124320019 吳邦齊

N124320020 陳軍弦

N124320024 何牧

N124320011 林紀吟

N124320017 郭良益

指導教授：黃三益 博士

中華民國 114 年 5 月

目錄

目錄.....	viii
圖次.....	vii
第一章 緒論	8
第一節 研究背景	8
第二節 資料清洗	8
第三節 LDA 模型視覺化	10
第四節 GuidedLDA 主題模型視覺化.....	12
第五節 Bertopic 主題模型.....	14
第六節 字詞網路圖.....	15
第七節 單中心網路圖	15
第八節 總結	17

圖次

圖 1 PTT 爬蟲方式	8
圖 2 停用詞處理	9
圖 3 LDA 主題模型參數設定.....	10
圖 4 主題一視覺圖	11
圖 5 主題二視覺圖	11
圖 6 主題三視覺圖	12
圖 7 GuidedLDA 主題模型參數設定	12
圖 8 主題一視覺圖	13
圖 9 主題二視覺圖	13
圖 10 主題三視覺圖	14
圖 11 主題三視覺圖	15
圖 12 字詞網路圖	15
圖 12 第七節 單中心網路圖	16

第一章 緒論

第一節 研究背景

根據 PTT 爬蟲資料，分析該論壇三個版別中，關鍵字為中國之社群討論關係；本次以八卦、政黑、外匯三個版別進行探討，這三個版別之特性列點如下：

1. 八卦 - 最多人討論的板別
2. 政黑 - 發表對政治不滿的板別
3. 外匯 - 與金融相關之版別，由於此次時間恰好列於關稅相關討論期間，故一同列入。

時間設定 2025/04/13 至 2025/5/9，關鍵字僅設定中國，合計獲取 6373 筆資料。PTT 爬蟲方式如圖一所示：



The screenshot displays the 'PTT 爬蟲 (200)' application window. It features a '參數設定' (Parameter Setting) tab and a '任務結果' (Task Result) tab. Under '參數設定', there is a '選擇看板' (Select Board) dropdown menu with options: Finance(金融業), Food(美食), ForeignEX(外匯), forsale(二手買賣), gay(甲), and GetMarry(結婚). The 'ForeignEX(外匯)' option is currently selected. Below this, there are input fields for '搜尋起始日期' (Search Start Date) set to '2025/04/13' and '搜尋結束日期' (Search End Date) set to '2025/05/09'. To the right, there is a '搜尋關鍵字' (Search Keyword) field containing '中國' and a '排除關鍵字' (Exclude Keyword) field which is empty. A green bar at the bottom contains the text '儲存更改' (Save Changes).

圖 1 PTT 爬蟲方式

第二節 資料清洗

資料清洗處由於該論壇多為中文討論者，相關細節一一整理如下，字詞加權：

1. 言論自由 500
2. 國家安全 500
3. 國民黨立委 500
4. 民進黨立委 500
5. 民眾黨立委 500

停用詞處理處如圖 2 所示，另外圖片限制，自定義停止詞羅列如下：

📌、💡、💥、🔴、●、◆、☀、▪、太、號、嘅、路線、係、元、林道、唔、店、歎、着、現在、一直、一堆、一下、網址、心得、附註、內容、標題、新聞、完整、嘖嘖、表示、來源。

清除停用詞 (12)

參數設定	Input - 8	任務結果
語言 *	Chinese	
是否清除單字元 ⓘ	是	
清除英文字母 *	是	
清除換行符號 *	是	
清除html tag *	是	
使用預設停止詞	是	
是否轉為小寫英文	是	
清除數字 *	是	
清除特殊標點符號 *	是	
自定義停止詞	<div>📌 💡 💥 🔴 ● ◆ ☀ ▪</div>	

儲存更改

圖 2 停用詞處理

轉為 DTM 後，計算其字詞、單中心網路圖及關聯式文字雲，由於有先於 ngram 先進行計算排除停用詞，故 DTM 處不多加調整：

☰ 轉為DTM  (66) ✕

參數設定

Input - 12

任務結果


保留詞彙 ⓘ
以換行符號區隔，e.g.
國立中山大學
西子灣
壽山...

最多篩選詞彙數量 ⓘ
200

儲存更改

第三節 LDA 模型視覺化

設定主題數為 3，詞彙頻率下限為 40，alpha、chucksize 使用預設，迭代次數為 50 次、主題保留關鍵字數量 20、詞彙頻率上限 0.5、update_every 為 1，詳如圖三所示：

☰ LDA主題模型  (98) ✕

參數設定

Input - 12

任務結果

目標欄位 *
result

主題數 *
3

詞彙頻率下限 ⓘ
40

alpha
預設為主題數/50

chucksize ⓘ
預設為2000

是否輸出字典
是

迭代次數
50

主題保留關鍵字數量
20

詞彙頻率上限 ⓘ
0.5

Beta
預設為0.1

update_every ⓘ
1

儲存更改

圖 3 LDA 主題模型參數設定

而三大主題結果可以鑑別出以美國及關稅相關為主如圖 4，台灣政治內部討論如圖 5，中國影響討論如圖 6：

LDA Vis

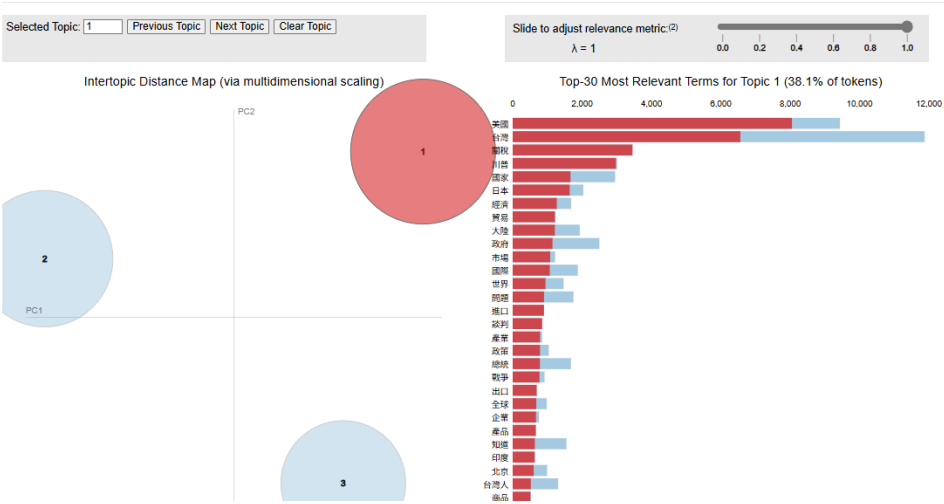


圖 4 主題一視覺圖

LDA Vis

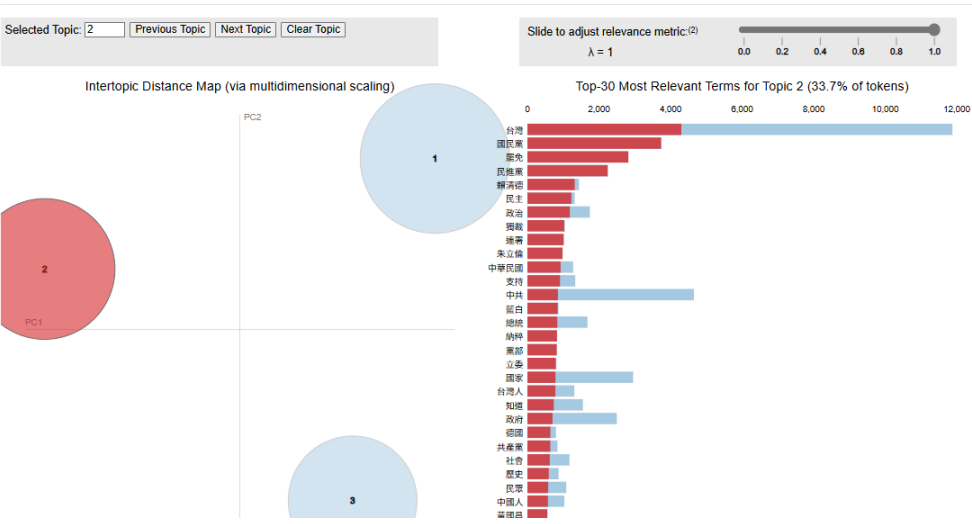


圖 5 主題二視覺圖

LDA Vis

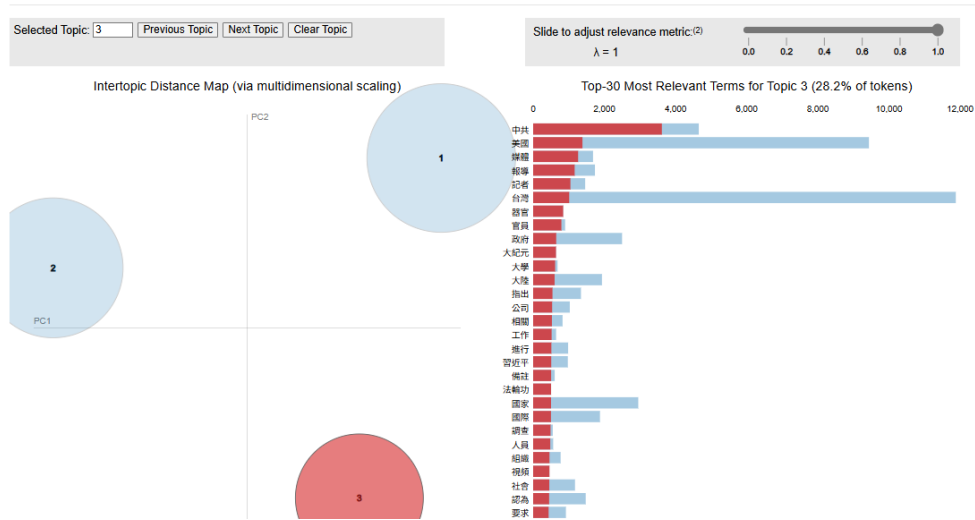


圖 6 主題三視覺圖

第四節 GuidedLDA 主題模型視覺化

設定主題數為 3，詞彙頻率下限為 40，alpha、chucksize 使用預設，迭代次數為 50 次、主題保留關鍵字數量 20、詞彙頻率上限 0.5，細節如圖七所示：

GuidedLDA 主題模型 (100)

參數設定 | Input - 12 | 任務結果

目標欄位 *
result

主題數 *
3

詞彙頻率下限 ⓘ
40

alpha
預設為主題數/50

主題種子字 ⓘ
美國 關稅 川普
台灣 國民黨 民進黨
中共 大陸 政治

迭代次數
50

主題保留關鍵字數量
20

詞彙頻率上限 ⓘ
0.5

Beta
預設為 0.1

是否輸出字典
是

圖 7 GuidedLDA 主題模型參數設定

而三大主題結果可以鑑別出以美國及關稅相關為主如圖 8，台灣政治內部討論如圖 9，但第三項中國影響更偏向是外交相關議題，另詳如圖 10：

GuidedLDA Vis

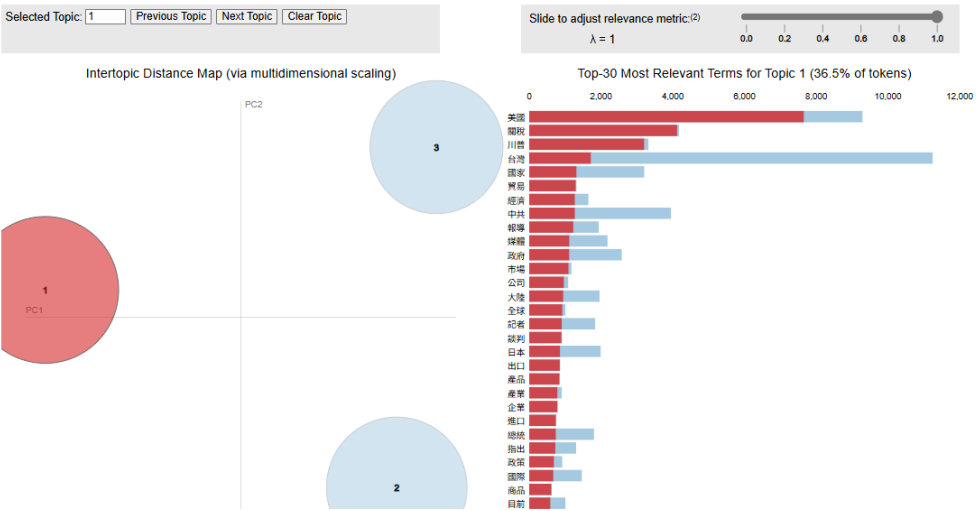


圖 8 主題一視覺圖

GuidedLDA Vis

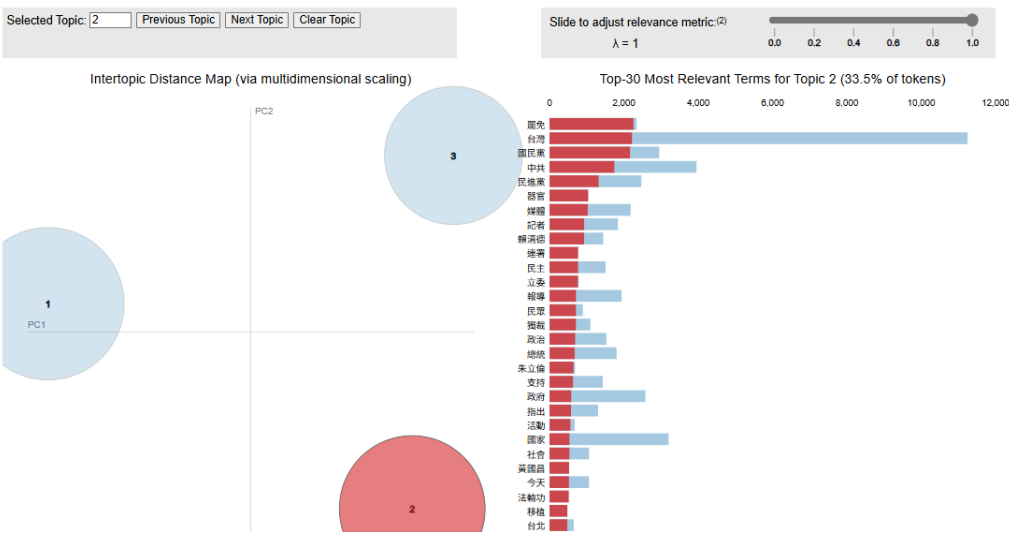


圖 9 主題二視覺圖

GuidedLDA Vis

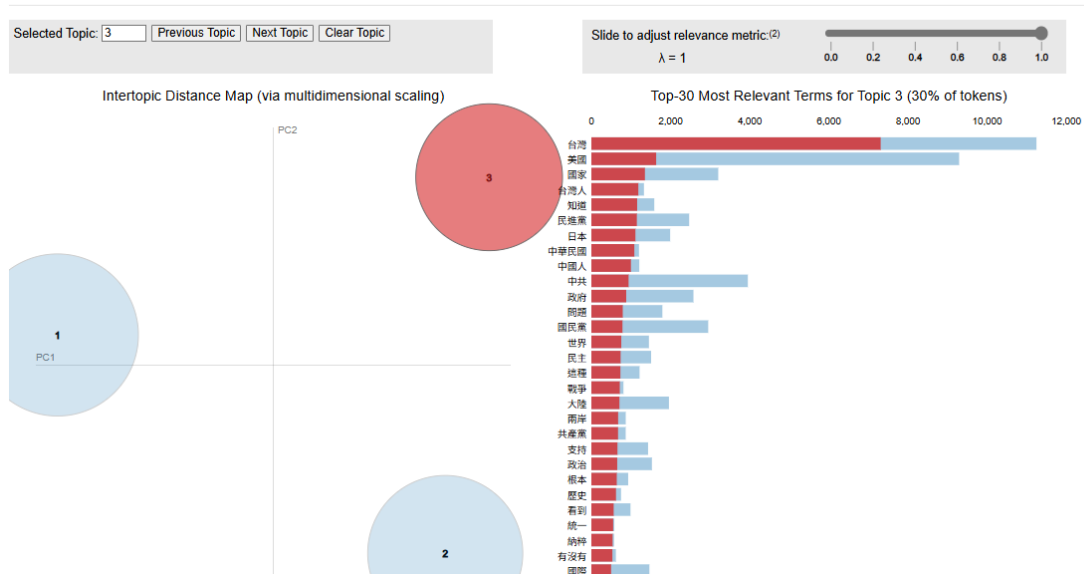


圖 10 主題三視覺圖

第五節 Bertopic 主題模型

首先使用 bert-base-chinese 訓練中文 BERT 模型，訓練完成後 BERTopic 參數設定為多語言、文章向量壓縮維度為 8、主題數為 5、主題保留關鍵字數量為 20，可以看到主題分為 0、4 探討著台灣相關內容，主題 2、3、4 討論著外交相關內容，詳如圖 11 所示。

BERTopic Vis



圖 11 主題三視覺圖

第六節 字詞網路圖

接續轉化 DTM 後進行字詞網路圖分析，主要概念群組如下：

1. 政治相關：「中華民國」、「立法院」、「反共」、「法輪功」等。
2. 地理位置：「台北」、「亞洲」、「歐洲」、「越南」等。
3. 經濟相關：「企業」、「貿易」、「獨裁」等。
4. 社會文化：「教育」、「研究」、「網路」等。

從圖中可以觀察到「中華民國」是一個較大的節點，與多個概念相連，可能是分析文本中的核心主題。其他明顯的關聯包括政治、地區、社會和文化等多個維度的詞彙間的連結。詳如圖 12 所示。

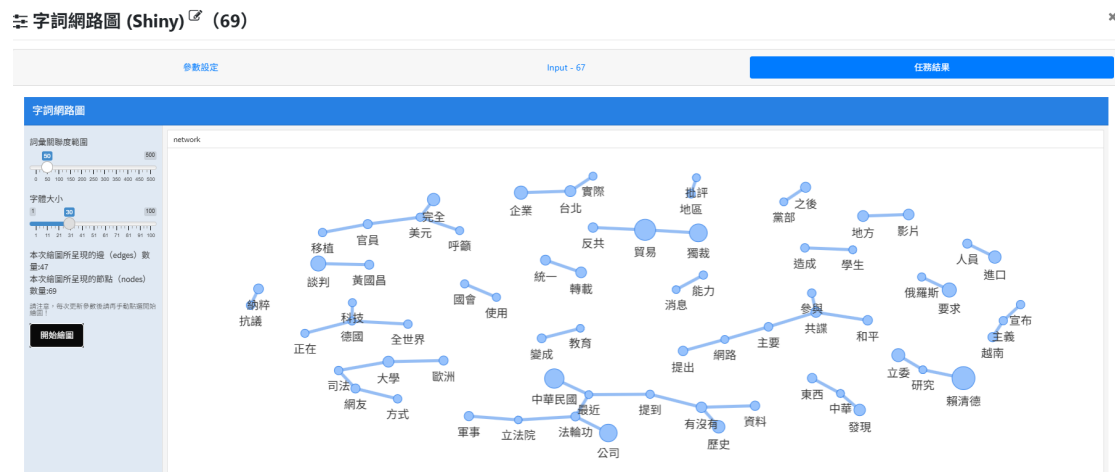


圖 12 字詞網路圖

第七節 單中心網路圖

從圖中可以觀察到某些節點較大，如「中國」可能是核心概念或高頻詞彙。詳如圖 13 所示。

圖 13 第七節 單中心網路圖

將停用字清除後，使用 Word2Vec 將文章轉換為 100 空間維度向量，並篩選最低詞頻 5 及參考字詞數 5。

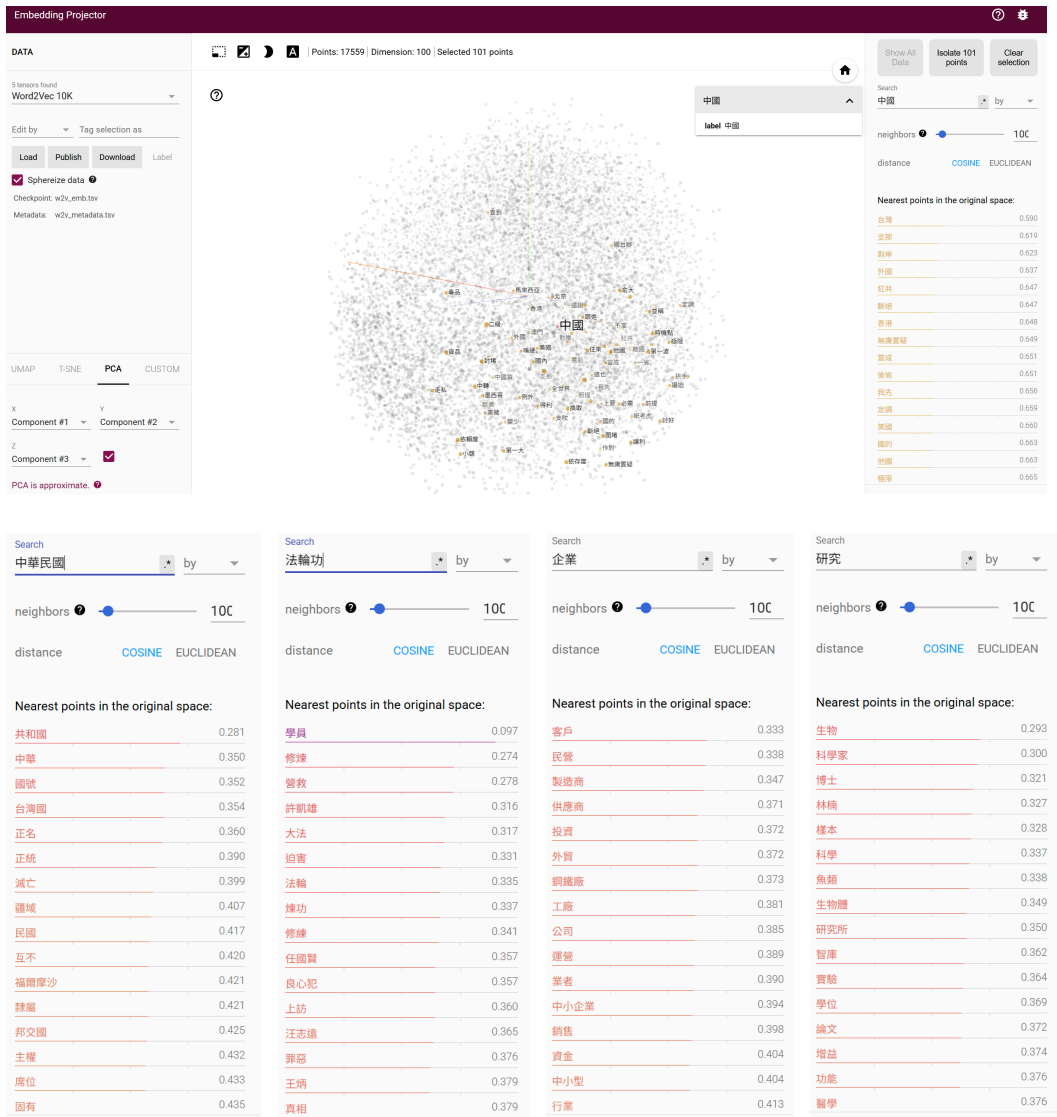
得出之 csv 檔案轉換為 tsv 後，丟入 <https://projector.tensorflow.org/> 投影，確認字詞關係圖，觀察各字詞網路圖之關鍵字之相關參考字詞。

(2)以法輪功為關鍵字，可觀察大多為中國對於法輪功的一些新聞，偏向負

面的。

(3)以企業為關鍵字，可觀察多為製造或供應商等上下游關係等，說明台灣與中國之間於企業端合作密切。

(4)以研究做為關鍵字，可觀察對於中國之生物、科學之討論較多。



第九節 總結

本研究透過 PTT 爬蟲技術，針對八卦、政黑及外匯三個版面進行「中國」關鍵字的社群討論關係分析，蒐集時間從 2025 年 4 月 13 日至 5 月 9 日，共獲

取 6,373 筆資料。研究採用多元主題模型方法進行分析，以探索網路社群對中國議題的討論脈絡與焦點。

一、主要研究發現

三種主題模型的分析結果顯示出網路社群討論的幾個明確方向：

5. LDA 模型分析呈現三大主題：

- 美國與關稅相關討論
- 台灣內部政治討論
- 中國影響相關討論 1

6. GuidedLDA 模型分析結果大致相似，但第三主題更專注於外交議題層面，顯示引導式主題模型在特定議題上能有更明確的聚焦。

7. Bertopic 模型使用 bert-base-chinese 進行訓練，設定 5 個主題，將討論內容分為：

- 台灣相關內容 (主題 0、4)
- 外交相關內容 (主題 2、3、4)

二、模型比較與應用價值

三種主題模型各有特色，但均能有效識別網路社群對中國議題的主要討論向度。特別是 Bertopic 模型透過預訓練的中文 BERT 模型，能更精確地捕捉中文語境下的語義關係。

三、字詞網路圖分析

字詞網路圖是透過文本分析後將相關詞彙間的連結可視化呈現，從附件中可見此分析確實識別出四個主要概念群組：

- 除了「中華民國」、「立法院」和「反共」外，還包含「法輪功」等政治議題相關詞彙

- 包含「台北」、「亞洲」、「歐洲」外，還有「越南」等區域詞彙 1
- 這表明討論範圍涵蓋了國內及國際地理視野
- 地理詞彙的出現反映了網路社群在討論中國相關議題時常採用國

際比較視角

- 「企業」、「貿易」、「獨裁」等詞彙表明了經濟議題討論與政治制度的連結

四、單中心網路圖分析

單中心網路圖與字詞網路圖不同，專注於展示單一核心詞彙與其他詞彙的連結關係，而「中國」在此網路圖中是一個較大的節點，可作為搜尋關鍵字且確實是整個分析文本的核心詞彙

研究結果顯示，PTT 論壇上關於中國的討論主要集中在美中關係（特別是關稅議題）、台灣政治與外交議題上，反映了當前社會對這些議題的高度關注。