

# 機器學「習」：以文字探勘法探索 習近平時期之大外宣戰略\*

邵 軒 磊

(國立臺灣師範大學東亞學系副教授)

## 摘要

因為中共政經地位的轉變及其「銳實力」的影響，國際上之「中國形象」在近年有相當變化。尤以習近平提出「大外宣」戰略後，上述發展更為顯著，同時其細節也尚待研究。對於研究者而言，因為各種資料的質與量的快速成長，使得以個人經驗與智識判斷為主的研究方式受到挑戰。相對於此，使用數位方法的優點在於得以檢驗並更有效累積；觀察資料與模型，也能適應未來變化而增添調整。因此，本文作者試圖將數位技術應用至此一題目，比如文字探勘、機器學習、主題分析模型等，在龐大現代政治論述文本中建立主題模型，嘗試尋找政治領袖在演講中所透露之政治訊息以及政治價值，亦即指出習近平時期演講在各個主題的概括樣貌。初步的解答是：以習近平自身講話做主題分析後，確實發現其對外與對內用語不同，機器能分辨並歸類其用語特色；也能看出在領域上則「外交、經濟、生態」類文本之主題與「黨建、政治、國防」之主題不同。本研究蒐集了中國國家主席習近平的發言作為語料庫，並使用數位方法初探中文政治文獻，期待藉此關注中共大外宣與銳實力。

**關鍵詞：**文字探勘、機器學習、大外宣、習近平

\* \* \*

---

註\* 本文為「從知識系譜到知識地圖：大數據與機器學習下的『中國研究』（107-2410-H-003 -058 -MY3）」與「AI的能與不能：人文社會的脈絡、價值與效應」（108-2634-F-003-005）之研究成果。

## 壹、研究背景

「對外宣傳大布局」（Grand External Propaganda Strategy，以下簡稱「大外宣」），是近年研究者觀察主要對於中共促進在外影響力的主要核心議題。對中共黨理論而言，「宣傳」是用以執行中共之「統一戰線」（united front）政策，爭取友人孤立敵人之工具。因此，對外宣傳在傳統黨組織任務分配中，亦具有對內對外之戰略意義。

就習近平時期中共之外宣戰略，可以下列講話為代表：

…（宣傳思想工作之任務是）宣傳闡釋中國特色，要講清楚每個國家和民族的歷史傳統、文化積澱、基本國情不同，其發展道路必然有著自己的特色；講清楚中華文化積澱著中華民族最深沉的精神追求，是中華民族生生不息、發展壯大的豐厚滋養；講清楚中華優秀傳統文化是中華民族的突出優勢，是我們最深厚的文化軟實力；講清楚中國特色社會主義植根於中華文化沃土、反映中國人民意願、適應中國和時代發展進步要求，有著深厚歷史淵源和廣泛現實基礎。對我國傳統文化，對國外的東西，要堅持古為今用、洋為中用，去粗取精、去偽存真，經過科學的揚棄後使之為我所用。要精心做好對外宣傳工作，創新對外宣傳方式，著力打造融通中外的新概念新範疇新表述，講好中國故事，傳播好中國聲音（新華網 2013）。

上述講話強調「講清楚、講好中國故事」但方法是要「融通中外的新概念新範疇新表述，講好中國故事」。一般而言，學者解讀將「融通中外」解釋為要「以外國文化能接受的方式強調中國故事」，強調「國情論再包裝」（新華網 2013）。亦有西方學者注意到習時期外宣的特色之一是在外國也使用內部宣傳主題與手法（Klimeš and Marinelli 2018, 313），這可能是我們理解此一戰略的線索。詳細的文獻觀點，本文將在「文獻回顧」一節中詳述。

另一方面，西方國家以銳實力（sharp power）來理解習時期的大外宣方針，如 C. Walker 與 J. Ludwig（2017）提出的，中共的外宣貌似軟實力（soft power），但與軟實力不同，其背後就是威權政府對民主世界增加影響力。軟實力與銳實力的不同在於「由下而上」與「由上而下」，就像好萊塢與孔子學院的差異。其中以軟實

力的提倡者奈伊（J. Nye, 2018），甚至認為銳實力會是軟實力的威脅，為文「銳實力是怎麼威脅軟實力」。在奈伊的思維中，銳實力代表中俄，軟實力代表美國：「銳」對上「軟」的鮮明形象，使得銳實力一躍成為美國外交研究界理解中國外宣方針的詞彙。

然而具體的操作，在早期的文章中主要是指官方控制的智庫媒體與宣傳機構，具有「權力單極化、由上而下的控制、言論管制與收買」（monopoly on power, top-down control, censorship, and coerced or purchased loyalty）等特色（Walker and Ludwig 2017）。而到了2019年才逐漸有學者指出「（中共）使用資訊控制管理對政治與媒體領域的征服」（political and media spheres of other countries by conducting information operations）才是銳實力的主要目的，主要體現為錯誤資訊、假新聞、資訊操作（misinformation, fake news and manipulation of information）等手段（Leonova 2019, 21-22）。

因此，本研究基於上述兩層次的意識，首先在政策關心上試圖回答習時期的大外宣戰略是什麼內涵？體現為具體討論「融通中外與內外有別」之宣傳議題，以及其具體外宣用詞。其次，在研究方法上，筆者也意識到對於在外宣中的具體操作，如使用巨量人造資訊流（information flow）的關心。這樣的資訊流數量過於龐大，而較難以人工一一解析，在未來必須使用數位工具。因此，本文特意使用數位研究法，如文字探勘（text mining）與機器學習（machine learning）來探索此一課題，這樣的工具能夠超越人力所能閱讀的部分，雖然也有相當的研究限制（請參考結論節），在詮釋或立論時應當配合使用者自身的主體知識（domain knowledge）或是專家同儕的意見。

總言之，本文將使用習近平演說語料做為資料核心，使用數位工具來研究這些語料，並得到若干數位實證做為其外宣策略的回饋，用以提供研究者檢視。其中本文有兩個預設前提，其一、習近平演說語料體現了外宣策略；其二、數位研究法中所做之關鍵詞或主題分析能夠從語料中擷取出論述策略。關於前者，由於外宣策略並未有一種公諸於世的具體文件，以往學術文獻多從克林姆林學（Kremlinology）角度詮釋中共行為（Yahuda 1972, 32）。因此使用習近平演說語料做為探索資料，是不亞於傳統作法的策略。而後者，同樣的某些主題或詞彙能夠成為外宣策略，也是延伸自傳統中共研究的詮釋學途徑，認為某些詞彙能脫離原有文意脈絡，成為中共某些特定政治立場之載體，如文革時期使用「毛語錄」來表現政治主張，因此熟練的學者可能從「政治宣言之使用詞彙」來斷定中共的路線與政治風向。易言之，本文在這個議題的研究方法論基礎是使用傳統中共研究的觀測語彙理念，但在研究

器械上使用了數位工具。這麼做的好處在於，相對於全套傳統的研究方式，本文之數位工具本身是可以傳遞的（相對於資深研究者的心法經驗），而觀察資料也有可能增添調整（比如社交軟體通訊數據或研究者之論文）。

本文僅使用數百篇習近平演說語料，可以歸類於探索性研究（Exploratory Research）作品，旨在釐清從「資料蒐集、資料清洗、模型建制、結果詮釋」之整體流程，並做學術前沿探索；所以本文並不宣稱自己為對中共外宣研究之最適結果，毋寧說是做為其他既有研究之補充與開拓未來研究之可能性，若有學者使用更大規模的語料（如微博語料），則結果應能更為令人期待。

## 貳、文獻回顧

在中國大陸研究之領域，研究官方文稿中的詞語與口號，一直是相當重要的課題。在過去，研究者們總是試圖從領導人用語中讀出「深意」，比如說特別在研究習近平時期對外宣傳政策時，多數人都注意到了習特別提出了「融通中外」這樣的用詞，在中文文獻中有下列討論，如：「融通中外取代內外有別」（唐潤華 2015, 30）、「融通中外是戰術、內外有別是戰略」（李彥冰、荊學民 2010, 33）等說法。但這些說法多數強調「政策詮釋」，多以描述中國現今社會狀況與國際環境為主要立論基礎，因此其結論往往限於「對字面的理解」。

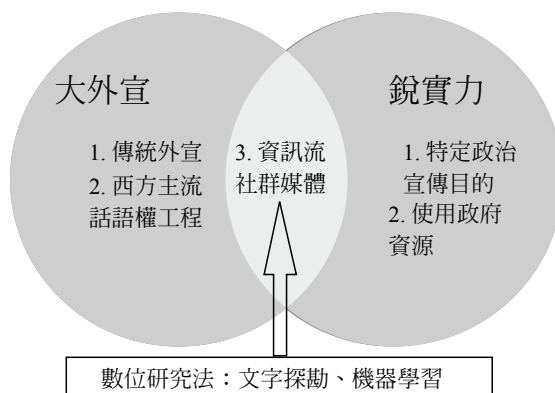
若延續前節研究背景中所描述，吾人可以發現對中共外宣政策的研究，其理解典範正在轉移。比如說「大外宣」有三種組成：1. 超越針對華人社群「傳統外宣」（海外中文媒體、華人社團與中文學校）；2. 針對西方主流（播放中國國家形象片，甚至使用資源收買媒體），爭取話語權的「全球戰略的文化工具」（何清漣 2019）；3. 針對新媒體、社交平臺的資訊流。（Nip and Sun 2018, 98）其中又以第三項理解外宣政策資訊流值得詳述。

另外，關於「使用數位方法」研究的論文，亦有討論中國政府推特帳號（twitters）之推文與回文（Nip and Sun 2018），確證了中國官方常常藉由有意的操作某些話題（在 Nip and Sun 的研究案例中是以「南中國海」為關鍵字），來推升其話題性；其中一個特色是幾乎不與使用者互動，「單向」的傳遞官方立場。這篇文章甚至藉由數據分析能夠指出，在這些重複推文（repeating tweets）的行為背後，存在很多虛假帳號（Nip and Sun 2018, 118）。這是這篇文章很重要的貢獻與發現，因為這種方法能夠由下而上的指出某些「外宣」行為，而非依照學者經驗猜

測或藉由政策文件來說明。

這與本文所討論的另一個議題「銳實力」，若合符節。在研究背景中已經有若干描述，茲簡述「銳實力」之特色如下：1. 內容帶有特定意識型態與政治目的（Nye 2018）；2. 在資源上由上而下使用政府資源（Walker and Ludwig, 2017）；3. 除了傳統媒體以外，更著力使用社交媒體、網路平臺（Leonova 2019）。在近期銳實力的討論中，不只是研究影響策略，也需要理解其影響手段。中共提出「銳實力」以一帶一路倡議等等結合，在經濟社會方面影響「政治」，已經引起歐美諸國的關心。臺灣學者也意識到此一議題，相關的研究除前述學者陳至潔之外（Chen and Hsu 2018），學者王韻研究習近平之統戰戰略，也呼應了對這個主題的關心（Wang and Groot 2018）。因此，本研究試圖結合上述文獻對「大外宣 / 銳實力」的關心，以數位方法對中共領導人文本分類做出得以相互參照的研究。茲將「大外宣 / 銳實力 / 與數位研究法」圖示如下（圖 1）。若是「大外宣與銳實力」兩者的共同交集，就是使用「巨量資訊流去『攻占』社群媒體」，因此本文使用數位工具，能應對探索此一領域。

圖 1 本文重要觀點圖示



資料來源：由作者自製

近年將文字作為一種「數據資料」，再利用機器學習甚至人工智慧（artificial intelligence，簡稱 AI）來分析巨量文字已逐漸成為趨勢（Wilkerson and Casas 2017）。有論者提出「Text as Data」的觀念，利用各種文字處理與統計分析方法進行分類或歸類的運算（可統稱為文字探勘技術），能進行自動化資訊組織與主題

分析 (Grimmer and Stewart 2017)。

不過雖然機器學習發展甚早，但至今在政治學領域，使用機器學習探索仍屬前沿文章，以往即使學者試圖從事相關研究，也多由人力閱讀（邵軒磊 2012）。近期有 Margaret E. Roberts 挑戰了使用結構式主題模型（structural topic model, STM），來解決文本結構化問題（Margaret et al. 2014）。Bjorn Burscher 則是成功加強機器訓練文本資料的研究模型（Burscher, Vliegthart, and Vreese 2015）。Benedikt Boecking 使用了 LDA 演算法（Latent Dirichlet Allocation, LDA），精細的預測了社會狀態 (Boecking, Hall, and Schneider 2015)。John Wilkerson 以機器學習對「文本」，甚至大數據文本做演算（Wilkerson and Casas 2017）。特別在「中國研究」議題上，國際上較為傑出的論文，有學者陳至潔從大數據文字資料中，研究中國政府對人權之「概念」，已經能做大量與深入的中文文字探勘技術，給予後人相當的啟發（Chen and Hsu 2018），以及在臺灣學界研究「中國大陸研究論文之主題分析」，也試著做了「數百篇中國研究論文」的遠程閱讀（邵軒磊、曾元顯 2018）。

簡言之，研究者需要傳統的政策詮釋能力用以能夠理解政策方向，而面臨巨大的偽造／人造資訊流（falsification of information flow），研究者也必須有相應的知識工具。因此，本文在此一階段作為探索性研究，試圖開展此一領域。

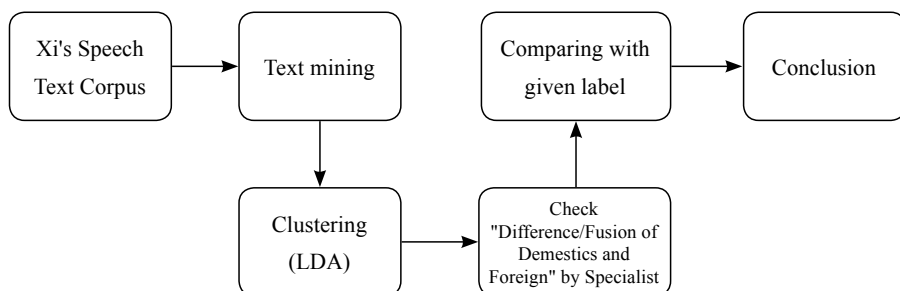
## 參、研究設計

### 一、研究流程

本研究設計流程如下：

1. 蒐集研究資料，以習近平演講文本建制語料庫。
  2. 使用文字探勘（text mining），含分詞、關鍵字探索、製作文本矩陣（Document Term Matrix, DTM）等。
  3. 文本分群，依照 DTM 將文本分群，並找出分群關鍵字。
  4. 對照分群結果，與原始標籤做對照。
  5. 評價外宣策略是否出現「內外有別／融通中外」情形。
- 總體而言，本文的研究流程理念圖示如下（圖 2）：

圖 2 研究流程設計



資料來源：由作者自製

本研究之主要思維是：將「原始研究資料」以「非監督式機器學習」分類後，再與人工原有標籤（既有給定分類）比較，用以論證是否在習近平演講中用詞或主題是否「內外有別」。其中，就「研究資料」與「研究方法」（文字探勘、文本分群）上再深入說明之。

## 二、資料來源

由於在初期研究中，限於既有的時間成本，並無法做到蒐集「所有外宣文章」作為語料庫。因此，本篇論文使用「習近平系列重要講話資料庫」（人民網 2018），其為中共官方媒體新華社設計架設，具有原始標籤八種分別為：「經濟、政治、文化、社會、生態、黨建、國防、外交」。由於採計節點與資料清洗的時間問題，本文採計其資料庫建制至 2017 年 9 月之 263 篇。茲將其「原始分類、篇數、樣本編號」整理如表 1。

表 1 原始分類數量表

原始分類	經濟	政治	文化	社會
篇數	35	43	43	47
樣本編號	1-35	36-78	79-121	122-168
原始分類	生態	黨建	國防	外交
篇數	16	28	14	37
樣本編號	169-184	185-212	213-226	227-263

資料來源：由作者自製

並建制可用資料欄位如圖 3。在數量上，數百篇量級的文本是否得以做出恰當的文字探勘素材？在其他論文中又是怎麼處理數量級的問題，如前述，Titus C. Chen 和 Chia-hao Hsu（2018）使用巨量與歷時的通訊文本，其信度較為卓著。而在某些狀況下或初期探索中，使用一定數量文本如葉淑慧、蔡明月（2018）或邵軒磊、曾元顯（2018），也是可茲接受的。

圖 3 原始習近平語料庫示意圖

id	type	date	title	content
1	經濟	2017/7/2	習近平在慶祝香港回歸祖國二十周年大會暨香港特別行政區第五屆政府	同胞們，朋友們：今天，我們懷著喜悅的心情，在這裡隆重集會，慶祝香港回歸祖
2	經濟	2017/5/25	習近平致信祝賀全國臺灣同胞投資企業聯誼會成立十周年	人民網北京5月24日電（記者徐雋）在全國臺灣同胞投資企業聯誼會成立10周年之際
3	經濟	2017/5/16	習近平在“一帶一路”國際合作高峰論壇開幕式上的開幕辭	各位國家元首、政府首腦、國際組織負責人：我宣佈，“一帶一路”國際合作高峰
4	經濟	2017/5/16	習近平在“一帶一路”國際合作高峰論壇開幕式上的開幕辭	各位同事：剛才，我們完成了最後一個階段會議的討論，“一帶一路”國際合作高
5	經濟	2017/5/15	習近平在“一帶一路”國際合作高峰論壇開幕式上的演講	尊敬的各位國家元首、政府首腦，各位國際組織負責人，女士們，先生們，朋友！
6	經濟	2017/1/18	習近平在世界經濟論壇2017年年會開幕式上的主旨演講	尊敬的各伊特爾德主席和豪森先生，尊敬的各國元首、政府首腦、副元首和夫人，
7	經濟	2016/11/22	習近平在亞太經合組織第二十四次領導人非正式會議第一階段會議上的	尊敬的庫琴斯基總統，各位同事：很高興同大家相聚在美國的利馬，共謀亞太合
8	經濟	2016/11/21	習近平在亞太經合組織工商領導人峰會上的主旨演講	布斯塔曼特主席先生，亞太工商界各位代表，女士們，先生們，朋友們：很高興
9	經濟	2016/11/17	習近平：集思廣益增進共識加強合作 讓互聯網更好造福人類	11月16日，第三屆世界互聯網大會在浙江省烏鎮開幕。國家主席習近平在開幕式上
10	經濟	2016/10/16	習近平致信祝賀第一百二十屆中國進出口商品交易會開幕	人民網廣州10月15日電（U+00AD）第120屆中國進出口商品交易會15日在廣州開幕。
11	經濟	2016/9/6	習近平在二十國集團領導人杭州峰會上的開幕辭	各位同事：我們用了一天半的時間，圍繞會議主題和重點議題進行了熱烈而富有啟
12	經濟	2016/9/5	習近平：中國發展新起點 全球增長新藍圖	女士們，先生們，朋友們：下午好！很高興同大家在杭州相聚。明天，二十國集團
13	經濟	2016/9/5	習近平：構建創新、活力、聯動、包容的世界經濟	各位同事：我宣佈，二十國集團領導人杭州峰會開幕！很高興同大家相聚杭州。
14	經濟	2016/6/25	習近平在上海合作組織成員國元首理事會第十六次會議上的講話	尊敬的卡裡莫夫總統，尊敬的各位同事：很高興來到美麗的塔什干，出席上海合
15	經濟	2016/1/16	習近平在亞洲基礎設施投資銀行開業儀式上的致辭	在亞洲基礎設施投資銀行開業儀式上的致辭 中華人民共和國主席 習近平 尊敬的各

資料來源：由作者自製

將上述資料作為「語料庫」，經由文字探勘所處理，如處理停用詞（stopword）、分詞（segmentation）、標準化（normalization）之後，可以獲得之文本矩陣（document term matrix, DTM），再接續下述文本分群研究。

### 三、研究方法

在研究方法上，本文使用非監督式文本分類。主要使用「潛在狄立克雷分配」（即 Latent Dirichlet Allocation, 以下稱為 LDA）演算法，使用詞袋模型（Bag-of-words model）的概念，計算某一「詞彙組合」的最大可能之貝式定理組合，從而找到「主題」（Blei, Ng and Jordan 2003; Blei and Lafferty 2007）。LDA 主要的核心概念為，自然直觀的現實世界中，吾人在閱讀時能將文章視為「字詞→文章」的原子化組成。但可以列出一種模型為「字詞→主題→文章」的三層模型。這樣的模型可以使用貝式機率（Bayesian probability）來表示。在特定有限文章（語料庫）中，主題數量、字詞範圍也確定的前提下，上述機率總和之極值就能夠成為一個數學可解的問題，亦即某形式分布的主題最有「解釋力」，遂由此得以成立「最佳分群」。需要說明的是：電腦並不知道「詞彙組合是否有意義」，他只能知道「某些組合機率更大」，因此對於詞彙組合（主題）是需要藉由研究者詮釋的。



## 肆、研究成果

因此，本文使用之 LDA 可以大規模自動化分出語料庫中的主題。如「研究流程圖」所示，本文蒐集了大量習近平演說作為語料，然後藉由非監督式分類。如果某些文本「被機器分為同類」，吾人得以觀察這些是否擁有某些特性，與其他不同類的文本得以比較。

### 一、習近平語料庫高頻關鍵字

總體而言，在最初分詞下，所有文本可達 21454 個「詞」。再進一步經過詞性、詞頻、出現次數等等方式，縮小範圍至 1966 個特徵詞。本文將藉由這些詞來作為每個文本主題分類的「依據」。依照順位呈現分詞結果如表 2：

由詞頻可以看出若干線索。最多的詞是「人民」，多來自「人民共和國」、「人民網」；其次「國家」則多來自「國家主席」一詞；「同志」則是發語詞。但接下來的「經濟、社會主義、世界、社會、歷史」則各自代表某些主題，不言而喻。另外，吾人也可以從對於中共用語的熟悉程度，找到一些線索，如「代表」多來自「三個代表」（江澤民的意識形態）、「科學」多來自「科學發展觀」（江澤民的意識形態），「小康社會」雖然比較普遍使用，但在中共黨用語中，是鄧小平時期的慣用語。「社會主義、馬克思主義」則如同字面上，都是與中共黨意識形態有關係。讀者也可以依照自身的關心，或依照下文複數的語料庫的方式來觀察。

### 二、內外有別之二分群檢視

在將前述詞頻統計的基礎之下，本研究檢視如果使用 LDA 演算法做出兩分類，這樣是否能夠體現「內外有別」的用詞狀態。我們選取兩分類，並列出其中兩列關鍵字的部分。表 3 展示在前六十位足以「代表」兩種分群的關鍵字：

表 2 習近平語料庫之前 100 高頻詞

rank	Var1	Freq	rank	Var1	Freq	rank	Var1	Freq	rank	Var1	Freq
1	人民	3572	26	關係	545	51	政策	314	76	中心	224
2	國家	2181	27	水準	532	52	理念	309	77	重點	221
3	經濟	1927	28	理論	529	53	交流	308	78	作風	219
4	社會主義	1818	29	馬克思主義	527	54	民主	305	79	法治	219
5	世界	1699	30	全球	526	55	共用	291	80	條件	216
6	社會	1566	31	體系	503	56	經濟社會	286	81	動力	215
7	同志	1287	32	能力	496	57	政府	283	82	優勢	214
8	歷史	1212	33	時代	487	58	金磚	282	83	體制	212
9	全面	1155	34	朋友	480	59	一帶	275	84	全黨	211
10	國際	1142	35	科學	478	60	責任	274	85	情況	211
11	政治	995	36	力量	472	61	中央	265	86	規律	207
12	特色	987	37	全國	446	62	祖國	265	87	生產力	203
13	制度	955	38	成果	441	63	小康社會	264	88	地位	203
14	領導	879	39	哲學	433	64	黨章	263	89	人才	200
15	事業	741	40	目標	431	65	命運	262	90	關鍵	200
16	精神	733	41	基礎	426	66	原則	256	91	黨員	200
17	道路	730	42	時期	406	67	方向	255	92	大會	198
18	民族	715	43	市場	381	68	風險	255	93	黨和人民	198
19	幹部	671	44	貢獻	375	69	同胞	237	94	工程	197
20	文化	654	45	利益	367	70	經驗	234	95	機遇	197
21	人類	646	46	核心	365	71	方式	232	96	環境	197
22	戰略	596	47	地區	353	72	技術	232	97	地方	194
23	科技	583	48	生態	325	73	主席	227	98	形勢	189
24	思想	563	49	代表	324	74	會議	226	99	活力	187
25	領域	546	50	機制	315	75	網路	226	100	企業	186

資料來源：由作者自製

表 3 習近平語料庫二分類之前 60 特徵詞

rank	Topic 1	Topic 2	rank	Topic 1	Topic 2	rank	Topic 1	Topic 2
1	價值	倡議	21	理論體系	成員	41	銘記	短板
2	方針	能源	22	藝術	權力	42	實踐證明	資源配置
3	紀念	對話	23	共產黨人	議程	43	作品	公約
4	真理	區域	24	條例	協同	44	優良傳統	狀態
5	人生	兩國人民	25	領導集體	機會	45	行政區	主權
6	工作者	植物	26	分子	生產總值	46	紅軍	案件
7	國防	模式	27	創業	地球	47	違紀	經濟體
8	全黨同志	貧困地區	28	職業	非公有制	48	作出貢獻	資金
9	強軍	夥伴	29	侵略者	專家學者	49	投身	資料
10	信念	人文	30	軍國主義	差距	50	國家主權	共商
11	問候	農村	31	文藝工作者	項目	51	內地	城鄉
12	實事求是	氣候變化	32	同胞們	潛力	52	初心	負責人
13	軍事	機構	33	治黨	恐怖主義	53	書寫	家園
14	家庭	人員	34	民主黨派	記者	54	廣大黨員	核心技術
15	衷心	同事	35	工人階級	貧困人口	55	戰鬥力	總理
16	信仰	司法	36	征程	科學技術	56	孩子	人權
17	共產主義	利用	37	強國	保護主義	57	高級幹部	名義
18	教師	管道	38	功勳	產業革命	58	文選	機關
19	時刻	框架	39	兩岸關係	精準	59	素質	人民幣
20	臺灣同胞	資本	40	基本法	行政	60	智庫	速度

資料來源：由作者自製

綜合來看，Topic 1 比較體現出「對內講話的特點」，其中包含「全黨同志、同胞們」等等用語；而 Topic 2 比較體現出「倡議、兩國人民」等等用語。若專家大致以人力閱讀也能若干可以理解其關鍵字之取向，當然，也有可能只從單一用語缺少上下文比較難以分辨的部分，吾人可進一步如下處理。

若試圖深入關於 Topic2 的相關詞彙，可以 KWIC (Key Word In Context Index, 關鍵字索引) 來找出在什麼樣的語境中使用這些詞彙。首先選取 Topic2 前

十位關鍵詞「倡議、能源、對話、區域、兩國人民、植物、模式、貧困地區、夥伴、人文」。然後先觀察在語料中的相鄰詞並計算相關詞，爲了呈現關鍵用詞，我們將高頻組合列出。

表 4 Topic 2 高頻關鍵字關聯詞表

Rank	Topic2 Top10	neighbor word	count	Rank	Topic2 Top10	neighbor word	count	rank	Topic2 Top10	neighbor word	count
1	夥伴	關係	159	16	區域	民族	15	31	植物	人民	9
2	夥伴	戰略	62	17	對話	交流	15	32	植物	國家	9
3	倡議	一路	28	18	模式	世界	15	33	夥伴	金磚	9
4	植物	科學	27	19	夥伴	全面	13	34	區域	性	8
5	植物	國際	27	20	夥伴	經濟	13	35	對話	國家	8
6	夥伴	交易	25	21	能源	能源	12	36	倡議	戰略	7
7	區域	合作	24	22	植物	世界	12	37	能源	全球	7
8	植物	學	24	23	植物	著	12	38	能源	領域	7
9	區域	經濟	20	24	對話	協商	12	39	區域	化	7
10	能源	技術	19	25	對話	機制	12	40	夥伴	朋友	7
11	植物	全球	18	26	能源	新	11	41	對話	問題	7
12	夥伴	精神	17	27	能源	資源	10	42	模式	制度	7
13	夥伴	國家	16	28	模式	經濟	10	以下多數並列 43 位			
14	區域	制度	15	29	倡議	全球	9				
15	區域	城鄉	15	30	倡議	國際	9				

資料來源：由作者自製

從表 4 可以看出，Topic 2 用詞中「夥伴關係」是外交核心用詞，幾乎所有外交文本都強調這個用語，次數高達 159 次，某些文本中甚至提及數次。其次爲「戰略夥伴」、「一帶一路倡議」，這也是常提及的用詞，但「一帶一路倡議」比設想中要少很多，可能推估是近期政策宣傳上有低調跡象。「植物」相關用語則是相當多（植物 - 科學、植物 - 國際、植物 - 全球），這是非常特殊而常常被研究者忽略的。「植物」是生態分類中的重要用詞，常常出現在環保相關講話中。筆者推估做爲中共宣傳材料以及國際合作檯桿中，較諸其他領域有「過於強調」（over-

emphasized) 之可能，猜測這是在外宣中試圖強調低層政治（生態），塑照和平崛起形象，不過尚待進一步分析。

若回到本文主軸再進一步，本文可以回頭將所有文本做分類，將其歸類於 topic 1/ topic 2，從各個原有標記中檢視，可以發現「所有」（被原有標記為）外交的文本，都被歸類到 Topic 2，而另外一端如「黨建、國防」都大部分被歸類於 Topic 1。若吾人將「外交（原始標記）/topic 2（機器學習分類）」作為「偏向『外』」分類而「黨建、國防」作為「偏向『內』」分類，我們可以得到某種分類文本中，偏向「內／外」的比率。使用兩分類可以告訴我們，某一種領域的分類中，哪些領域文本用詞較為「接近」，以及其使用「內／外」用語的差異。

表 5 習近平語料庫二分類與原始分類對照表

原有標記	文化	外交	生態	社會	政治	國防	經濟	黨建
Topic 1	30	0	2	33	33	12	7	24
Topic 2	13	37	14	14	10	2	28	4
Topic 1/ total Topic	69.77%	0.00%	12.50%	70.21%	76.74%	85.71%	20.00%	85.71%

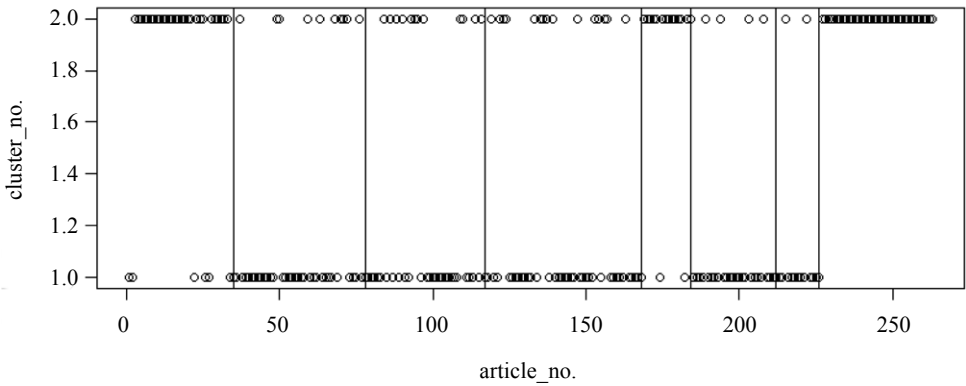
資料來源：由作者自製

雖然每個領域皆有若干「融通中外」的文本。但如表 5 所示，「外交、生態、經濟」使用「外用語」較多，文本也多被分類為「Topic 2」。而另一方面「黨建、國防、政治、社會、文化」則還是更多使用「內用語」，因此文本也更多機率被歸類為「Topic 1」。如果依照分類與編號排列，可以看出原始分類中哪些分類更為「內外有別」，那些更為「融通中外」。在圖 4 中每個點位置代表一篇演講，各領域分類以縱線分隔，縱軸表示主題分布，可以看出二分類的細節效果。

### 三、內外有別之多分群檢視

更進一步依照原始標籤數量，設定 LDA 分群數量為八種。因為原始標籤如前所述共有八種分別為「經濟、政治、文化、社會、生態、黨建、國防、外交」。因此，本文更進一步依照八種分類，呈現八種關鍵字（表 6）：

圖 4 原始標記與機器分群比較圖（二分類）



說明：原始分類從左到右依序為：經濟、政治、文化、社會、生態、黨建、國防、外交

資料來源：由作者自製

表 6 習近平語料庫八分類之前 25 特徵詞

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
1	藝術	全黨同志	方針	紀念	倡議	兩國人民	貧困地區	工作者
2	文藝 工作者	條例	創業	植物	對話	司法	協同	家庭
3	非公有制	治黨	臺灣同胞	強軍	氣候變化	記者	貧困人口	人生
4	機會	理論體系	兩岸關係	信念	同事	人文	科學技術	教師
5	作品	廣大黨員	基本法	共產主義	議程	行政	農村	價值
6	農民	高級幹部	行政區	軍事	區域	夥伴	差距	職業
7	人士	文選	內地	真理	恐怖主義	人權	精準	分子
8	政治協商	黨的紀律	時刻	國防	框架	對方	資源配置	工人階級
9	民主黨派	有所	書寫	實事求是	能源	雙邊	專家學者	孩子
10	真情	優良傳統	幹事	侵略者	公約	共同利益	產業革命	智庫
11	農村	法規	國家主權	軍國主義	模式	權力	利用	強國
12	問候	違紀	信仰	紅軍	項目	人員	資本	問候
13	多黨合作	實踐證明	海外僑胞	銘記	夥伴	能源	短板	留學人員
14	信任	權力	居民	共產黨人	共商	公民	城鄉	素質

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
15	價值	起草組	自豪	初心	成員	潛力	核心技術	殘疾人
16	領導集體	板塊	大團結	信仰	人文	名義	區域	影響力
17	光榮傳統	黨性	地球	功勳	機構	開幕式	速度	面向
18	作出貢獻	單位	事件	敵人	保護主義	模式	工業	教材
19	建設者	部分	衷心	戰場	人民幣	國際事務	機器人	同學們
20	時代精神	書記	家園	遠大理想	主權	交易	狀態	學校
21	商量	文件	航船	全軍	動能	程式	科學考察	學會
22	作家	成員	偉業	同胞們	高級別	主席國	人人	勞動者
23	精神力量	領導集體	外交政策	征程	總理	外交	角度	職工
24	工人	科學化	作出貢獻	事實	資金	國事訪問	科技成果	大學
25	優秀作品	偉大旗幟	要務	治軍	醫藥	國際形勢	公共服務	機構

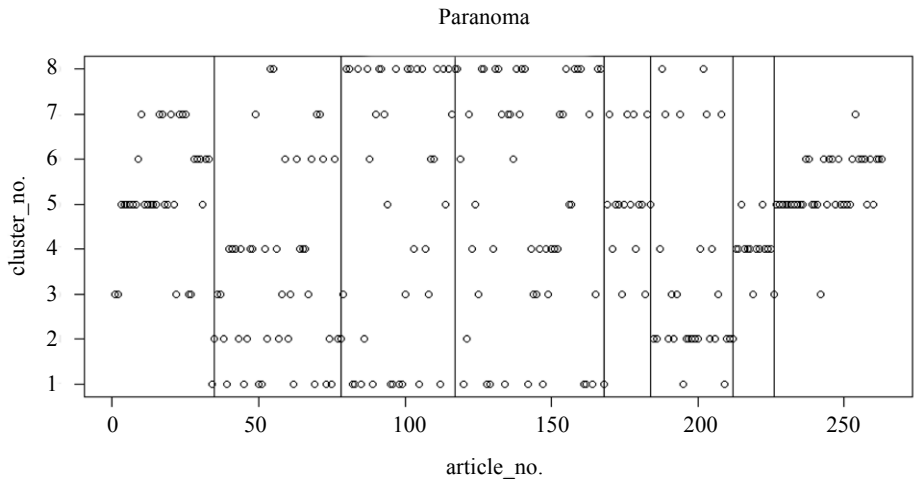
資料來源：由作者自製

由於是資料驅動（data-driven），這些字組合表示「這些字距離其他字更近」，所以如果硬要加上主題，可能不一定合理。而其次，LDA 的特色在於凸顯其分類特徵用詞，因此，可能在總語料庫中出現頻率高，但本表中並未出現；反之，本表可以呈現更為符合「主題」的字詞段落。讀者可以依照其字面意義，大致估計其可能「主題」，因此在文中就不特別命名。

而更進一步，使用在圖 5 中每個點位置代表一篇演講，各領域分類以縱線分隔，縱軸表示主題分布，可以看出機器學習八分群在原始標籤下的意義。與前節二分類相比，是有差異的。

吾人可以發現這樣的分類在：「經濟」、「外交」、「國防」、「生態」四項中分類相當好，在用詞上多集中於少量（甚至僅有一或兩個）主題。而「政治」、「文化」、「社會」、「黨建」較於分散。其原因則是因為本來語詞或演講主題本身就有歧義性，如「政治經濟」是政治還是經濟？「參加金磚會議」是外交還是經濟？「中美高峰會談」是政治還是經濟還是外交？這都很難武斷分類。不過，若有大量文本將原始分類與「機器學習後分類」製表對照如表 7，依然可以看出其分類趨勢。

圖 5 原始標記與機器分群比較圖（八分類）



說明：分類依序為：經濟、政治、文化、社會、生態、黨建、國防、外交

資料來源：由作者自製

表 7 原始標記與機器分群比較表

原始分類	topic_1	topic_2	topic_3	topic_4	topic_5	topic_6	topic_7	topic_8
文化	11	2	3	2	2	4	3	16
外交	0	0	1	0	21	14	1	0
生態	0	0	2	2	8	0	4	0
社會	9	0	5	8	3	1	8	13
政治	8	9	5	11	0	5	3	2
國防	0	0	2	10	2	0	0	0
經濟	1	1	5	0	15	6	7	0
黨建	2	14	3	3	0	0	4	2

說明：Pearson's Chi-squared test: X-squared = 299.82, df = 49, p-value < 2.2e-16, 但空格個數略多，因此可能解釋力有限。

資料來源：由作者自製

相較前節兩分類可能理解為「內外」，多分群比較難以完全被解讀。依照上述表格做了「卡方檢驗（Pearson's Chi-squared test）」得到分類有顯著差異的結



果，證明本分類還是若干有其效果。這樣的分法提供我們兩個觀察點。其一、與二分類相比，「外交」本身又被細分為兩個子類別（Topic\_5, Topic\_6），吾人可以再觀察兩種更深入觀察兩種子類別之主題；其二、以前述兩個子類別（Topic\_5, Topic\_6）作為基準，可以觀察其他分類的分布狀態，其中「黨建」所重複的文章數為「零」，表示「黨建用語」與「外交用語」差異最大，相反的「經濟」與「生態」主題文本數量較接近。

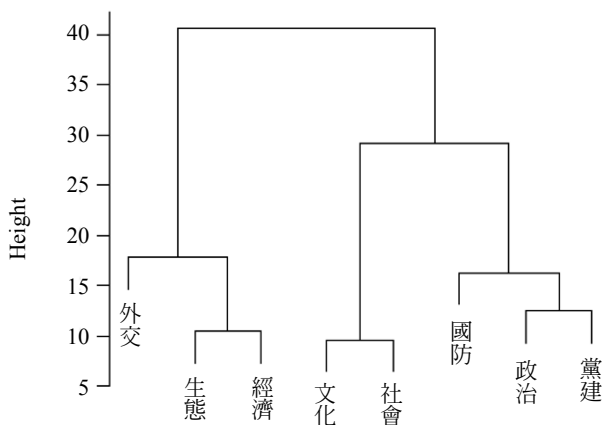
因此，從這些語料中也可以反過來論證，在中共領導的講話中某些「已知領域」的主題呈現若干重合的現象。依照語料庫中所「提及主題關鍵字」之接近程度，以本文結合其相似程度，計算出主題的重合程度（Ward 1963）。作法為使用每個詞在每個主題出現之機率，再由其算出歐幾里得距離（Euclidean Distance）（表 8）。數值越小則越「相似」，因此「外交」與「經濟」最近；和「政治」與「黨建」比較近。因此可以說中共外交的核心議題是促進經濟發展，而「政治」的核心議題是促進黨建，這多數合乎常識。可能是另一種解讀「內外有別」的方式，如果將「外交文本」的用語設定為對外用詞，那麼「政治文本」與「黨建文本」就是屬於對內用詞，介於兩者之間的就是外 / 內用語都有使用的領域。我們也可以將相似程度使用枝狀圖可視化如圖 6。

表 8 各文本歐幾里得距離

	文化	外交	生態	社會	政治	國防	經濟	黨建
文化	0.00	29.22	20.86	9.59	18.55	21.70	23.54	21.05
外交	29.22	0.00	19.47	29.53	28.50	25.67	12.41	29.38
生態	20.86	19.47	0.00	18.36	18.14	10.77	10.44	16.43
社會	9.59	29.53	18.36	0.00	16.19	18.14	21.63	20.52
政治	18.55	28.50	18.14	16.19	0.00	14.04	21.91	12.45
國防	21.70	25.67	10.77	18.14	14.04	0.00	19.10	16.55
經濟	23.54	12.41	10.44	21.63	21.91	19.10	0.00	21.38
黨建	21.05	29.38	16.43	20.52	12.45	16.55	21.38	0.00

資料來源：由作者自製

圖 6 原始標籤相似程度圖



資料來源：由作者自製

## 伍、結論

本篇論文試圖對中共大外宣戰略與「銳實力」的探索上，對於外宣「講中國故事」其採取「內外有別 / 融通中外」的程度。本文試圖使用數位方法，包括文字探勘與自然語言處理做出初探，讀者可以與其他質性論文相互對照，交互佐證。在研究設計上，本文使用「習近平系列重要講話資料庫」中二百餘篇講話作為語料庫，其為中共官方設計架設，具有原始分類八種分別為「經濟、政治、文化、社會、生態、黨建、國防、外交」。本文觀察各文本中「內外有別 / 融通中外」之比率與程度，使用機器分類後，與原始標籤相互參照，試圖回答：一、習時期的大外宣戰略是什麼內涵？體現為具體討論「融通中外與內外有別」之宣傳議題，以及其具體外宣用詞。二、筆者也意識到對於在外宣中的具體操作，如應當如何處理巨量人造資訊流。

上述問題的解答是：

一、在「習近平語料庫」中，我們可以用機器學習兩分群來體現了「內外有別」的用語現象。內外有別的用語上，內用語代表的有「價值、方針、紀念、真理、工作者、全黨同志、強軍、實事求是、信仰、共產主義」等等，外用語代表性的有「倡議、能源、對話、區域、兩國人民、模式、貧困地區、夥伴、氣候變化」等等。較為完整的列表，請參看「表 3 習近平語料庫二分類之前 60 特徵詞」。在

領域的分別上，若以「外交」、「黨建」為代表外用語與內用語的極端形式，則「外交、經濟、生態」偏向外用語、「黨建、政治、國防」偏向內用語，「文化、社會」則介於其中，可能比較屬於「融通中外」。以上結論，可以提供未來研究者觀察中共「銳實力」在各種領域表現時參考。

二、相對於 Nip 和 Sun (2018) 的文獻中使用關聯性短文本，主要研究「行為」，本文使用了「習近平講話」使用了長文本，有著較豐富的內容。因此，就不是從行為來觀測，而是具體要指出內容的差異。目前能做到的模型效果是，依照機器分類方式，能夠大規模識別習近平講話的「模式」，並且在基本的分類（中國政府自己定義下）中取得較準確成果，也初步詮釋成「內外有別」的結果。不過在手段上，直接不掩藏的把對內文稿使用做為網路平臺宣傳的舉措本身（Walker and Ludwig 2017）就呼應了官方的用語「融通中外」。筆者期望提出未來大外宣研究的「內容與行為」途徑，這兩者是可能有差異甚至不同步的，這相當程度暗示研究者，外宣內容與執行可能是由不同（兩個甚至更多）單位來實際操作。

在研究限制與缺陷上，本文也有很多未來需要努力的部分。比如一、本文目前的解釋力僅限於本語料集所蒐集的文獻，而在語料庫向外延伸與擴展主體知識的時候，是否還得以適用，目前無法得知。二、在分類演算法適用上，隨著更新的演算法與討論，也會有更「準確」或其他適用演算法，本文作為初探，僅選擇了目前較為常用的 LDA 演算法。隨著樣本結構的不同，可能總體運算結果不同。因此研究者主觀問題意識（哪些期刊、哪些年份）會在最初的階段起作用。三、若主題分群有理論支撐，比如說一般常識中有「某某部門」或必須分為幾群，也可減少探索分群時間。這需要研究者主體領域知識的協助。而有些零散字詞「以肉眼或常識看來毫無相關」，這涉及資料清洗，如辭典編撰可能會有若干幫助，這部分屬於模型調校，可以藉由經驗累積而改進。上述數點都是本研究需要改進的要點。

關於未來展望，目前成果能做為內容分析的初步效果。也毋寧說這個研究可能與未來的行為模式結合，比如說哪些社交媒體的官方（甚至某些特定）帳號，內容或立場體現了更多「習近平用語」？（結合內容與行為分析）或是「習近平用語」在時間上的轉變（結合時序分析）。在評論意見中，先進指出了演講對象的差異（西方/第三世界國家、雙邊/多邊場合等等），也確實可能造成用語差異。當然，這些都需要更深入的篇章來討論，因此本篇論文期望做為探索性研究，期待將來更多同儕投入。

本文實踐了在「中文政治語料」之主題解析並且順利找出關鍵字；以及實現「中型規模文本（字數約在  $10^3$ ，樣本數量約在  $10^{2-3}$ ）」的 LDA 作法。未來可

望以此為基礎，借鑑其他前人作品探索更大規模文本的解釋。本文最後成果大致符合研究者主體領域知識與一般常識，不過其中仍有未知之處，需要專家學者之協作。筆者承認需要改進的地方甚多，本文也期望提供可重複性（repeatability）研究之語料基礎，未來釋出提供研究者在共同基礎上更多研究討論，邁向改進。而做為探索性學術研究，本文核心旨趣是試圖使用數位研究法來解答一個傳統政策研究問題。筆者必須承認，傳統政策問題「是否應該使用」數位研究方法去解答，本身是一個哲學或是信念問題，可能不存在最終答案或是普世標準。本文的嘗試就如同敲打石頭想要照亮黑暗一樣，火花能夠晶晶閃爍；但落在適當的地方，就能燃起火把，給人光和熱。

\* \* \*

( 收件：107 年 12 月 31 日，接受：108 年 7 月 25 日 )

# Machine Learning: An Application of Text Mining to Xi's Grand External Propaganda Strategy

*Hsuan-Lei Shao*

Associate Professor

Department of East Asian Studies

National Taiwan Normal University

## Abstract

“China’s image” has undergone a dramatic transformation as China keeps expanding its international influence through its rising sharp power. In particular, much of the propaganda work under Xi Jinping has been carried out through a “Grand External Propaganda Strategy”. Growing concerns about the implications of this strategy demand a closer look at its characteristics and practice. Yet, there is a major methodological challenge of analyzing massive amounts of data efficiently and accurately while not relying solely on personal understandings. To solve this challenge, digital methods are required, which will be the topic of this paper.

This research collated Xi’s speeches into a database and developed digital tools to process and analyze “China’s Sharp Power” and “Grand External Propaganda Strategy.” It uses text mining and machine learning, as well as adopting the “Latent Dirichlet Allocation” (LDA) model to extract the main themes from various propaganda texts in order to identify the political messages and ideologies the Chinese top leadership tried to communicate. This paper demonstrates that there are significant differences between speeches on domestic topics: DIPLOMACY, ECONOMICS and ECOLOGY, and on externally directed topics: PARTY-BUILDING, POLITICS and DEFENCE.

This research collected Xi's massive political speeches to a database, developed digital tool to process them, and analyzed "China Sharp Power" and "Grand External Propaganda Strategy" to explore the unknown future.

**Keywords:** Text Mining, Grand External Propaganda Strategy, Xi Jinping, Machine Learning

## 參考文獻

- 人民網，〈習近平系列重要講話資料庫〉，<http://jhsjk.people.cn/>，查閱時間：2018/11/15。People.cn. 2018. “Xijiping xilie zhongyao jianghua ziliaku” [Xi Jinping’s series of important speech databases]. (Accessed on November 15, 2018).
- 何清漣，2019，〈紅色滲透：中國媒體全球擴張的真相〉，臺北：八旗文化。He Qing-liang. 2019. *Hongse shentou: zhongguo meiti quanqiu kuozhang de zhenxiang* [Red Infiltration: The Truth of Global Expansion of Chinese Media]. Taipei: Bachi Culture.
- 李彥冰、荊學民，2010，〈全球化背景下我國對外傳播內外有別原則的困境與出路〉，《湖南大眾傳媒職業技術學院學報》，10 (3)：29-33。Li, Yan-bing, Xue-min Jing. 2010. “Quanqiuhua beijing xia woguo duiwai chuanbo neiwai youbie yuanze de kunjing yu chulu” [Dilemma and Way-out of the Principle of Making Difference between Internal and External in International Communication in Globalization]. *Journal of Hunan Mass Media Vocational Technical College*, 10 (3): 29-33.
- 邵軒磊，2012，〈中國研究議程之系譜：以日本國際政治學會誌為例〉，《問題與研究》，51 (1)：23-54。Shao, Hsuan-lei. 2012. “Zhongguo yanjiu yicheng zhi xipu: yi riben guoji zhengzhixue huizhi weili” [Genealogy of the China Studies Agenda-In the Case of Journal Publications of the Japan Association of International Relations (JAIR)]. *Issues and Studies*, 51 (1): 23-54.
- 邵軒磊、曾元顯，2018，〈文字探勘技術輔助主題分析：以「中國大陸研究」期刊為例〉，《問題與研究》，57 (1)：29-62。Shao, Hsuan-lei, and Yuen-hsien Tseng. 2018. “Wenzi tankan jishu fuzhu zhuti fenxi: yi zhongguo dalu yanjiu qikan weili” [Text Mining-based Topic Analysis: A Case Study of the Journal "Mainland China Studies"]. *Issues and Studies*, 57 (1): 29-62.
- 唐潤華，2015，〈既要內外有別更要內外一體：對我國對外傳播原則的重新審視〉，《新聞與寫作》，2015 (3)：30-34。Tang Run-hua. 2015. “Jiyao neiwai youbie gengyao neiwai yiti: dui woguo duiwai chuanbo yuanze de zhongxin shenshi” [It is Necessary to Have Something Different Inside and Outside and More Inside and Outside: Re-examine the Principle of China’s External Communication]. *News and Writing*, 2015 (3): 30-34.
- 新華網，2013，〈習近平：講好中國故事，傳播好中國聲音〉，<http://www.scio.gov.cn/zhzc/10/Document/1345245/1345245.htm>，查閱時間：2019/9/27。Xinhuanet. 2013. “Xijinping: jiang hao zhongguo gushi, chuanbo hao zhongguo shengyin” [Xi Jinping: Speak China story well, Board China voice well]. (Accessed on Sep. 27, 2019).

- 葉淑慧、蔡明月，2018，〈中國大陸圖書資訊學領域「作者群研究」之學術論文探析：以CNKI收錄之中文學術論文為例〉，《圖書資訊學研究》，12 (2)：1-36。Yeh, Shu-hwei, and Ming-yueh Tsay. 2018. "Zhongguo dalu tushu zixunxue lingyu zuozhequn yanjiu zhi xueshu lunwen tanxi: yi CNKI shoulu zhi zhongwen xueshu lunwen weili" [An Exploration Study of the Authorship Literatures of Library and Information Science in China: A Case Study based on CNKI]. *Journal of Library and Information Science Research*, 12 (2): 1-36.
- Blei, David M., and John D. Lafferty. 2007. "A Correlated Topic Model of Science." *Annals of Applied Statistics*, 1(1): 17-35.
- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet allocation." *Journal of Machine Learning Research*, 3 (4-5): 993-1022.
- Boecking, Benedikt, Margeret Hall, and Jeff Schneider. 2015. "Event Prediction: With Learning Algorithms-A Study of Events Surrounding the Egyptian Revolution of 2011 on the Basis of Micro Blog Data." *Policy and Internet*, 7 (2): 159-184.
- Burscher, Bjorn, Rens Vliegthart, and Claes H. De Vreese, 2015. "Using Supervised Machine Learning to Code Policy Issues: Can Classifiers Generalize across Contexts?" *Annals of The American Academy Of Political And Social Science*, 659 (1): 122-131.
- Chen, Titus C., and Chia-hao Hsu. 2018. "Double-Speaking Human Rights: Analyzing Human Rights Conception in Chinese Politics." *Journal of Contemporary China*, 27 (112): 1-15.
- Grimmer, Justin, and Brandon M. Stewart. 2017. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis*, 21 (3): 1-31.
- Klimeš, Ondřej and Maurizio Marinelli. 2018. "Introduction: Ideology, propaganda, and political discourse in the Xi Jinping era." *Journal of Chinese Political Science*, 23 (3): 313-322.
- Leonova, Olga G. 2019. "Sharp Power- The New Technology of Influence in a Global Word." *Mirovaya Ekonomika*, 63 (2): 21-28.
- Margaret, E.Robert, Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner, Bethany Albertson, and David G. Rand. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science*, 58 (4): 1064-1082.
- Nip, Joyce. Y. M. and Chao Sun. 2018. "China's News Media Tweeting, Competing With US Sources." *Westminster Papers in Communication and Culture*, 13 (1): 98-122.



- Nye, Joseph S. 2018. "How Sharp Power Threatens Soft Power The Right and Wrong Ways to Respond to Authoritarian Influence." <https://www.foreignaffairs.com/articles/china/2018-01-24/how-sharp-power-threatens-soft-power> (Jan. 24, 2018).
- Walker, Christopher, and Jessica Ludwig. 2017. "Sharp Power: Rising Authoritarian Influence, International Forum for Democratic Studies." <https://www.ned.org/sharp-power-rising-authoritarian-influence-forum-report> (Dec. 5, 2017).
- Wang, Ray, and Gerry Groot. 2018. "Who Represents? Xi Jinping's Grand United Front Work, Legitimation, Participation and Consultative Democracy." *Journal of Contemporary China*, 27 (112): 1-15
- Ward, Joe. H., Jr. 1963. "Hierarchical Grouping to Optimize an Objective Function." *Journal of the American Statistical Association*, 58: 236-244.
- Wilkerson, John, and Andreu Casas. 2017. "Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges." *Annual Review of Political Science*, 20: 529-544.
- Yahuda, Micheal, 1972. "Kremlinology and the Chinese Strategic Debate, 1965-66." *The China Quarterly*, 49 (Jan. - Mar.): 32-75.

