

國立成功大學 資訊工程學系  
專題成果報告

**The Political News Opinion Recorder**

意言既出，駟馬難追：  
政事新聞意見紀錄簿

指導教授：高宏宇

組員姓名：王浩、黃子峻

## 摘要

近年來，台灣的新聞資訊量在眾多網路媒體報導的洪流中，以巨量的幅度快速增加，閱讀者對於如此大量的新聞報導，該如何精準且有效率的吸收資訊儼然成為了一大問題。另外，傳統的新聞文章，多半都以單篇純文字的文章結構呈現，而這樣純文字化的結構，並不利於閱讀者快速得到資訊，相當不友善。最後，尤其在台灣的政治類型新聞，多只以當日事件發展進行短篇幅報導，這對於關心政治議題的讀者公民來說，並不利於追蹤、了解政治人物們對於各種議題的態度和意見變化。

因此我們團隊以一目了然、視覺化的政事新聞長期意見紀錄為宗旨，決定了本次專題題目，以中文的自然語言處理（Natural Language Processing）領域內的技術，優化新聞資訊的呈現方式，改善上述問題。我們是利用各家網路新聞作為來源，將現有的新聞資訊進一步匯集、整合，以更利於閱讀者的方式，重新用不同面貌呈現新聞文本內容。

我們所自行設計的演算法框架，能從第一手的新聞純文字文本出發，先從各大知名網路媒體自動化爬取最新的新聞文章文字，接著把單日從各家媒體爬取到的大量新聞中做議題分群，也從每一篇新聞文本中擷取政治人物針對議題所發表的各種正反意見。

隨著蒐集整合新聞日數的增加，系統就可以在每位政治人物專屬的議題意見頁面，以時間軸視覺化的方式，看出每位在新聞文本中出現過的政治人物，他們在關於哪些議題發表過的全部意見言論，並且我們也加上了正負情緒分析，以更利於閱讀者的方式，將新聞以嶄新的面貌重現給淹沒在海量新聞中的公民讀者們。

我們的實驗數據說明了，藉由不同版本的演算法設計結果，與我們自行手動標記的政治新聞意見驗證，我們可以得到 precision 0.689、recall 0.699 的結果，由此可知，整合完的新聞呈現有一定的不失真性與可信度。

經過我們的努力研究，可以讓我們自行設計的演算法去重新剖析解構傳統新聞，藉由整合，賦予嶄新的新聞資料價值，讓新聞的閱讀者，能以不同的角度並且橫跨時間的維度，吸收政事新聞的政治人物發言資訊。

關鍵字：自然語言處理、意見探勘、新聞分群、意見持有者辨識

# 目錄

第一章、 研究動機 . . . . .	1
第二章、 系統設計與原理介紹 . . . . .	2
2.1 系統架構設計 . . . . .	3
2.2 套件演算法簡介 . . . . .	8
第三章、 實驗結果 . . . . .	11
3.1 驗證資料 . . . . .	11
3.2 分析結果 . . . . .	12
3.3 網頁呈現 . . . . .	13
第四章、 結論 . . . . .	16
第五章、 參考文獻 . . . . .	16

## 一、研究動機

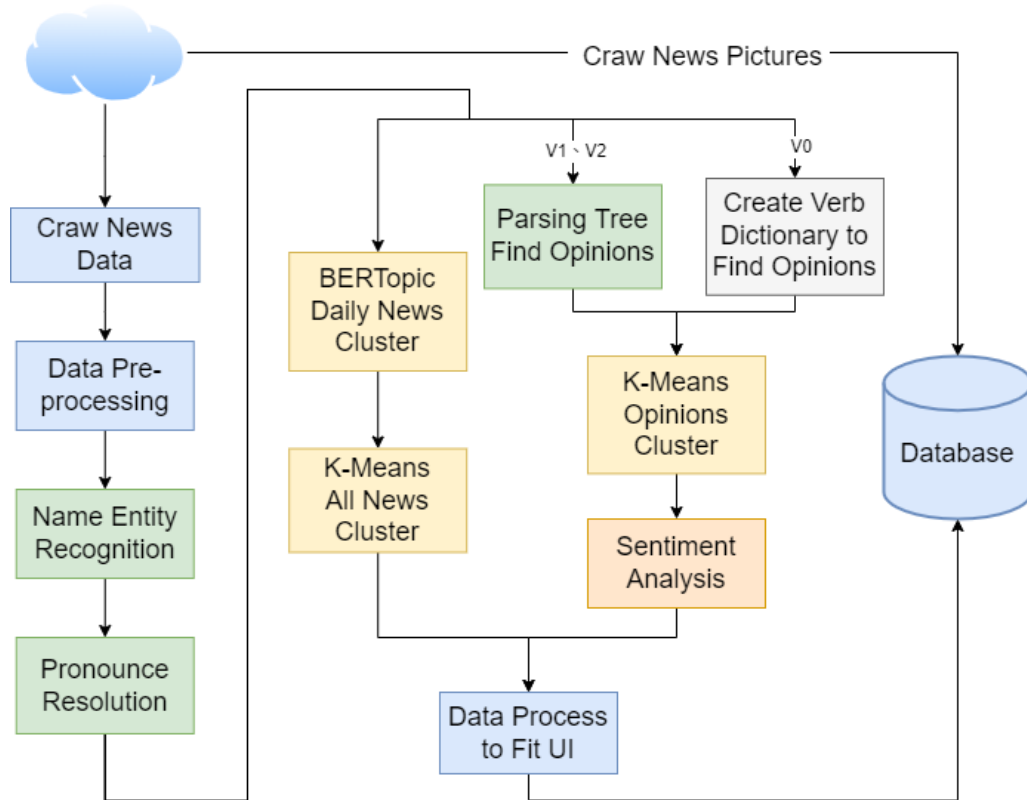
在如今資訊爆炸的時代，光是台灣目前網路主流媒體，就有相當多家，並且當中的政治新聞類別更是動輒上百篇，在這樣的資訊量下，新聞閱讀者們必然難以全部應付。此外，傳統的新聞文章內容也都是以純文字的結構呈現，並且在政治新聞的文章內，我們常常可以看到政治人物會對議題提出看法，但隨著時間流逝，政治人物的言行舉止也可能會隨著民意或是其他因素而產生變化，但如果從傳統的單篇單日新聞結構，公民閱讀者們就很難去追蹤政治人物的意見變化。

但如果以目前的新聞網站角度來思考，為了獲取第一手的最前線政治資料，新聞業者必定會聘請一定規模的報導團隊至現場訪問政治人物，此外，被聘請的文字記者們，每日也都是花費大量的精力與時間來撰寫並且整理當日所獲得的資訊，經過撰寫初稿、複稿，最後校稿等等嚴謹複雜程序，才能產生出我們網路上與傳統電視上所看到的政治新聞，而媒體業者光是一日當中就要耗費大量資源去製作新聞，又更何況新聞追求的是全年無休且即時反應的狀態，對於製作成本相當高的新聞資料，如果新聞還只是停留在過去的呈現模式，閱讀者也保持不變的閱讀方式，沒有辦法好好地吸收藏在每日新聞大量累積之後的資料價值，就會顯得非常可惜。

因此我們認為，如果以人工製作每日新聞為基礎資料集，加以整合目前研究已經有所成果的自然語言處理技術，並將其建構在我們的系統之上，其中，從大量的政治新聞中先進行議題分類，從其中提取出人物意見，再將相同或不同家媒體所報導的政治新聞中的人物議題意見進行整合，最後利用時間順序排列呈現，就能讓讀者透過閱讀依照事件分類的各個人物意見，達到從多家媒體的報導中獲得更全面的人物意見資訊目的。

## 二、系統設計與原理介紹

### 2.1 系統架構設計



▲圖一 系統架構

在我們所設計的架構中，為了拿取到最主流的當日新聞，我們每日都使用 GitHub 上的 Taiwan\_news\_dataset 專案[1]，爬取台灣網路媒體的新聞，其中的網路新聞媒體包含 CNA 中央通訊社、ETtoday 新聞雲、china times 中時新聞網、ftv 民視新聞網、SETN 三立新聞網與 Storm 風傳媒共六家網路主流新聞媒體。而在爬取完的新聞內，原始資料包含新聞標題、新聞文章內容、新聞日期與發佈時間、新聞原網址、綜合類別與新聞文章作者或來源，再經過分析過濾之後，我們需要擷取出來的每篇新聞資訊有新聞標題、新聞文章內容、新聞日期與發佈時間這三種類別。

因為後續任務，在抓取判斷相對應的字詞時，我們會使用到每個文章內部文字的位置資訊，因此我們必須先將每篇文章的空格與換行符號去除，以免使用套件之後造成資訊誤判，擷取出錯誤的字詞內容。

在一般的新聞文章內，我們觀察到，會有兩種情況的意見表述，一為「個人闡述」，另一為「交互對話」。個人闡述表示通常只有一位政治人物在表述個人單一立場的意見，而這樣發表較長的意見時，一個段落中，並不

會在每一句的主詞都直接的寫出完整姓名，而是使用職稱或直接使用姓等等來代稱，如：「國民黨主席朱立倫推動重啟核四公投，新北市長侯友宜的立場格外受到外界關切，侯友宜今表示會尊重公投，不過他提到『沒有辦法解決核廢料，我們沒有資格用核電』。」[2]。

而交互對話形式的意見表述，也會遇見相同的狀況，但不同的是，交互對話的省略行式，在同一段落之中，就會遇見不同人的代詞，情況也更為複雜，如：「國民黨立委鄭麗文 12 日表示，行政院長蘇貞昌應該展現保衛台灣的決心，『我怕到時候第一個投降的是你』，卻遭到蘇反嗆『不會像妳這麼不要臉』，鄭一度走向備詢台要求道歉，鄭麗文表示，請問蘇貞昌，立委質詢政府要拿什麼保衛台灣有什麼不對，『你以為你活在網路上？』」[3]。

為了解決此問題，我們使用由 CKIP Lab 中央研究院中文知識庫小組所提供的專有名詞辨識或實體辨識 (Named Entity Recognition, NER) 專案[4]，此系統可以找出輸入文章內的所有專有名詞，並將其專有名詞附上所屬的類別，其中包含 DATE 日期、EVENT 事件、FAC 設施、GPE 行政區、LANGUAGE 語言等等，根據我們的人工判斷，只需將 PERSON 人物、ORG 組織與 NORP 民族宗教政治團體加入到需要被考慮進行代換的候選名單。此外，為了增加精準度，我們限制候選名單的人物需要介於 2 到 4 個字之間，組織與團體需要大於 3 個字。

而如果同樣的一個字詞出現在不同的候選名單內，我們會進行名單內的數量統計，將候選名單內重複重疊字詞的候選詞過濾，只剩下出現最多次者。如「蘇貞昌、蘇揆、蘇院長、蘇」都出現在同一篇新聞中，通常都會是政治人物的完整本名出現次數最多，故可以統一化政治人物的代稱。而我們判斷主詞的方式也同是利用專有名詞辨識。故我們的設計為先使用專有名詞辨識列出統計之後的候選列表，再從頭逐一進行代換，達到單名代名詞統一代換的目的。

接著，我們的任務就會分成兩個階段，第一部分為新聞爬取下來之後的分群任務，目的就是為了將大量的新聞進行議題的分群，如此一來就可以減少閱讀者的閱讀新聞數量，讀最關鍵的新聞，掌握最核心的議題；另一個部分，就是從傳統新聞文字結構，萃取出政治人物的意見，使閱讀者能用更清楚明瞭的方式讀新聞。以下分別闡述兩個部分的處理。

第一的分群任務部分，我們使用能利用 BERT 與 TF-IDF 演算法產生出可闡述式主題的 BERTopic 套件[5]，其演算法我們也會在下一節進行討論，簡單來說，它把我們爬取完的當日政治新聞的標題文字先用 BERT 模型轉為向量表示，再以 UMAP 演算法將向量降維，用 HDBSCAN 分群，最後使用

c-TF-IDF 演算法來找出適合代表分群的詞彙，而我們也接續使用其最靠近分群中心位置的標題向量，來做為每一群分類的代表新聞。

此外，我們還使用剛剛利用專有名詞辨識系統找出的專有名詞，包含 PERSON 人物、ORG 組織與 NORP 民族宗教政治團體，統計出當日每群新聞內出現次數最高的前五個專有名詞，使用它們來做為新聞關鍵字，能夠方便讀者掌握每一群新聞的內容與方向。

做完當日的分群之後，我們還需要考慮跨日的意見分群，因為除了某些議題會跨日短暫延燒，更有涉獵廣泛的議題需要長期的追蹤關注，如：核能議題、疫情議題等等，所以我們團隊在處理完單日的新聞分群後，以跨日每一群的新聞議題的中心新聞標題作為輸入，放入在 BERTopic 套件內有使用過的子套件 sentence-transformer[6]，並且此子套件我們以 CKIP Lab 中央研究院中文知識庫小組所提供的 ckiplab/bert-base-chinese 模型[7]來作為 embedder，把代表跨日每一群的新聞議題的中心新聞標題再次轉為向量，並做機器學習當中的 k-平均演算法（k-means clustering）分群，至於分群的數量判斷，則是利用在變動分群範圍下的輪廓係數 Silhouette Coefficient Method 找出最適合的分群群數。

如此一來，我們就可以得知，每一個當日分群議題，是屬於哪一群跨日的長期議題，方便閱讀者在閱讀每日新聞時，都能夠得知跨日新聞彼此之間的議題相關性，發展出有別於傳統新聞的框架。


第二部分，意見提取，我們也發展出三種演算法，我們以 version 0, version 1 與 version 2 來稱呼。

version 0 中，我們觀察在新聞文本文章內，一個人物在發表意見所使用的動詞，或者為新聞報導的記者所使用的動詞，高機率出現在以下的列表內，為「說、說明、說道、提到、提出、指出、表示、解釋、回應、宣布、強調、呼籲、認為、發現、批評、評估、痛批、質疑、有感而發、不禁問道」，因此我們先找出文章內這些目標動詞，初步評估意見句會出現的位置，並且也以當下的動詞為基準，依序往前找出前方的第一個主詞作為意見持有者。

主詞的判斷方式，一樣是利用剛剛已經過濾代換過的專有名詞，當我們往前找到第一個對應的主詞時，為了提高精準度，我們還需要再進一步判斷，在主詞前方不能出現「於、在、受」這三種介係詞，否則就易造成主詞誤判，如：「原能會主委謝曉星今於原能會表示，核四廠建築已超過 20 年，包括零件設備都已經老舊風化，重啟可能性應該為零。」，「表示」此動詞，出現在我們所列的目標動詞，故系統會往前找尋第一個人物或組織的

專有名詞，如果不更正，本句找尋出的意見持有者為「原能會」，但如果加上介詞過濾後，系統就可以繼續依序往前找尋到真正的主詞「謝曉星」，如此一來，就能夠藉由此增加精準度。另外，我們再加上主詞與動詞的距離不能超過 30 字的限制，否則易把沒有出現意見持有者的句子搭配上前一句出現的專有名詞，造成錯誤。

#### version 0



原能會表示，作為獨立的安全管制機關，對核四公投「沒有預測立場，也會尊重公投結果」

動詞字典往前  
的第一個專有名詞

▲圖二 version 0 概念

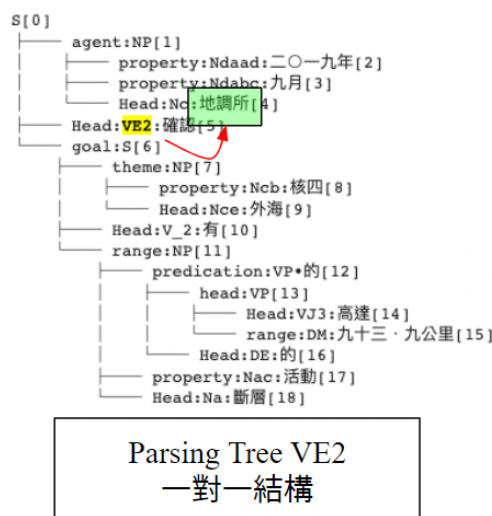
version 1, 2 當中，我們會先使用 CKIP Lab 中央研究院中文知識庫小組所提供的剖析專案[8]，我們也會在下一節介紹，大致而言，句子在此系統中自動剖析利用語法規律和斷詞後的文本做比對，找出可能的短語結構，由於存在歧義的短語結構，因此，利用結構出現的機率及檢測結構中詞與詞之間搭配的合理性成為解決結構歧義的方法。

在 version 1 內，經我們分析之後，我們取剖析後詞類分析為 VE2 的動詞，VE2 為二元述詞，以主事者（agent）為主語，終點（goal）為句賓語，語意多為表語言行為之述詞[9]。例：悲歎、自誇、下令、研究、討論、探索、反省、強調、猜想、說、提到。當我們提取出 VE2 之後，就可以在結構樹內找尋其主詞，根據我們判斷，我們主詞使用語意角色名為 agent 的字詞，agent 表事件中的肇始者，動作動詞的行動者[10]。

除此之外，我們還發現，我們可能會遇到同一個句子內出現不只一個動詞和主詞，如此就會造成我們系統可能會誤判真正的動詞與主詞的配對，因此在 version 1 內，我們限制只擷取句子格式為一個主詞與一個動詞的結構，簡稱為一對一的結構。



## version 1



▲圖三 version 1 概念

在 version 2 的設計中，我們一樣是使用類似 version 1 的設計，但我們此時在動詞的選擇上，多採取了詞類分析為 VC 與 VE11，VC 為動作單賓述詞，語意上需要兩個參與論元，VE11 為動作句賓述詞，後接句賓語的動作及物述詞問類，以主事者（agent）為主語，以終點（goal）為間接賓語，客體（theme）為直接賓語（句賓語），句賓語為疑問句式，且疑問範圍只到包接句[9]。此外，version 2 會處理句子內有多個動詞與主詞的結構，處理原則為將找到的動詞直接往前配對第一個 agent，簡稱為複雜結構。

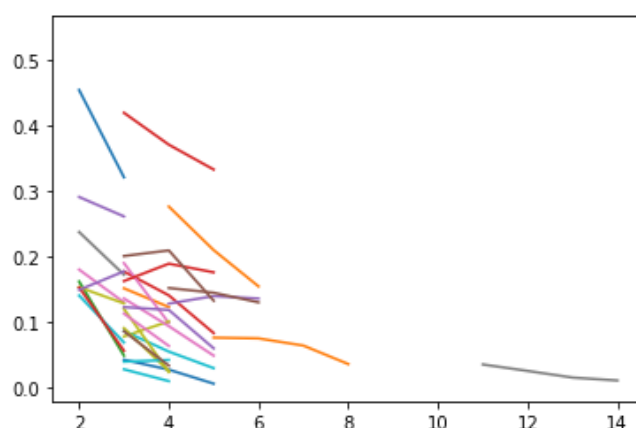
## version 2



▲圖四 version 2 概念

觀察傳統媒體新聞，我們進一步發現，通常有一類型的新聞一個人物說出的意見是本篇唯一重點，故相同的一句話，會不停地出現，為了避免我們系統在一篇新聞內也擷取出過多次幾乎相同的意見內容，所以我們設計意見分群來解決，使用套件 sentence-transformer[6]，並且此套件我們一樣以 CKIP Lab 中央研究院中文知識庫小組所提供的 ckiplab/bert-base-chinese 模型來作為 embedder[7]，把每一句意見都轉為向量表示，並做 k-平均演算法（k-means clustering）分群，至於分群群數的決定，則是利用在變動分群範圍下輪廓係數 Silhouette Coefficient Method 找出最適合的分群群數。而同一群的意見群，我們就取最靠近群中間的意見來當作代表，以免造成過多的相同內容佔據版面。

$$S = \frac{b-a}{\max(a,b)} \quad (a \text{ 表與同群的距離，} b \text{ 表與其他群距離之平均})$$



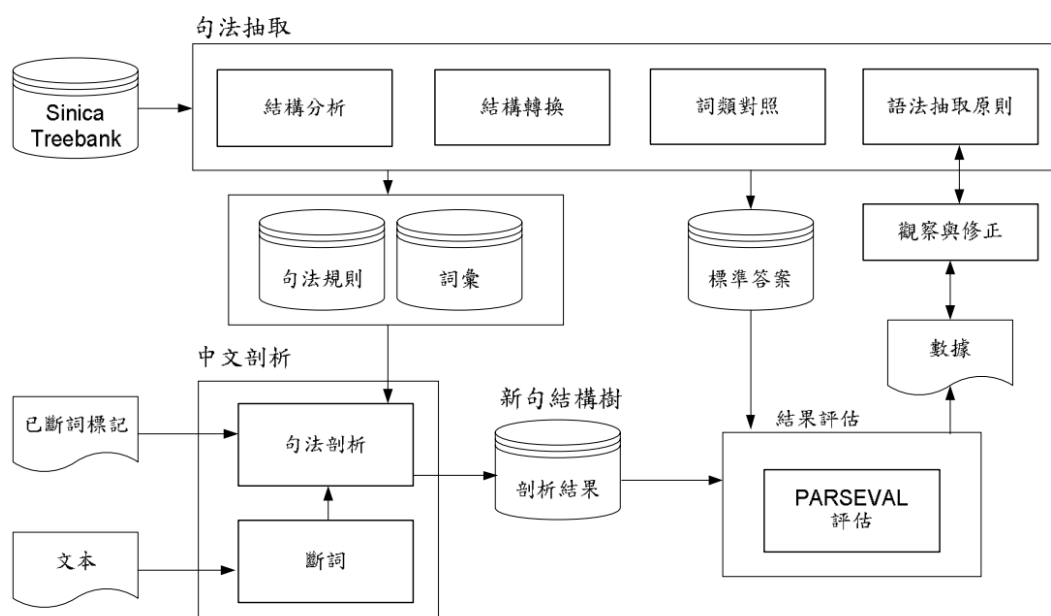
▲圖五 ettoday 手動標記 100 篇新聞的意見分群（橫軸為分群數、縱軸為輪廓係數）

分完群後，為了能夠讓讀者更快速的掌握政治人物的意見，我們使用中央研究院的中文情感語意分析套件 CSentiPackage[11]內的 NTUSD\_traditional 資料，此資料包含了 8276 個負面字詞、2812 個正面字詞，我們根據這些字詞去掃描政治人物所擷取出來的意見，並統計其中的正面與負面意見數量，利用出現較多者來判斷此句意見是正向或是負向的情緒判斷，由此即可以幫助閱讀者更好去理解政治人物意見。

而上述的內容我們都是以新聞和新聞內的意見為單位，但實際上在網頁呈現上，我們還需要進一步將資料以「人物與議題」做進一步的整合切割、重新整理檔案，最後，我們可以將資料庫與演算法內容以 Flask 為網站框架，並將系統架設在 Google Cloud Platform 的 Virtual Machine 上以供閱讀者使用。

## 2.2 套件演算法簡介

### 2.2.1 中文剖析系統

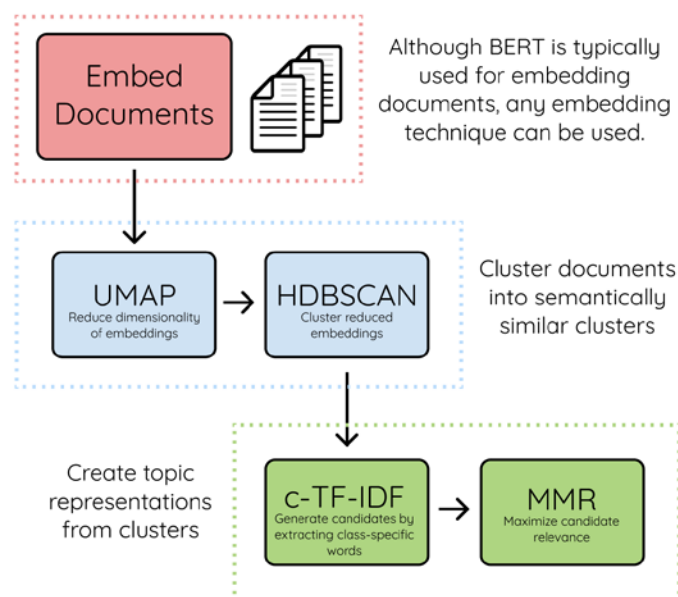


▲圖六 中文剖析系統架構[8]

中央研究院中文知識庫小組中文剖析系統主要為採用機率式無語境規律的模型（Probabilistic Context-free Grammar）為基本剖析架構再加入結構中詞彙搭配關係機率，以解決結構歧義的問題。而在結構決定之後，中文剖析系統也可選擇是否對結構進行語義角色的指派[8]。

分詞與詞類標記採用 CKIP lab 發展的中文斷詞與詞類標記系統。從上圖可知，句子自動剖析是利用語法規律和斷詞後的文本做比對，找出可能的短語結構，由於存在歧義的短語結構，因此，利用結構出現的機率及檢測結構中詞與詞之間搭配的合理性成為解決結構歧義的方法。

### 2.2.2 BERTopic 演算法模型



▲圖七 BERTopic 演算法[5]

BERTopic 套件[5]目的是能夠將文章分群，並且使用可解釋性的字詞來代表某群的性質，而實際的演算法會先用自然語言處理中的 BERT（Bidirectional Encoder Representations from Transformers）模型將輸入的每一個字詞都轉為一個高維的向量來表示，內部使用的方式為利用 sentence transformer toolkit[6]搭配 ckiplab/bert-base-chinese 模型[7]做 embedding。

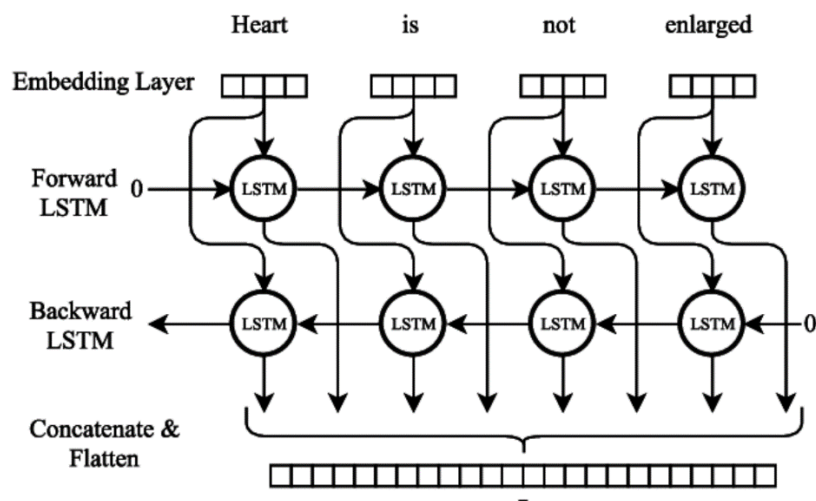
因為 embedding 所得之向量相當高維，表示較為稀疏也在效率上難運算，故以 UMAP（Uniform Manifold Approximation and Projection）演算法將向量降維，而此降維演算法的特色為它能保持原始高維度向量相當大部分的局部結構，使低維向量也能保持一致的局部結構。

接著使用 HDBSCAN 分群（Hierarchical Density-based spatial clustering of applications with noise），也因為在保持局部結構的情況下，HDBSCAN 這種 Density-based 的分群方式，效果會比較好，此外，此分群也不會強制全部的輸入都需要被分到其中一群，離群值並不會被迫分類。

使用 c-TF-IDF 演算法來找出適合代表分群的詞彙，其原理為某關鍵字在一群文章裡面出現的頻率再乘上其關鍵字在其他群文章裡面的稀有度。最後使用 MMR（Maximal Marginal Relevance）來移除過度擬合的單字（跟文章相似度高、但跟其他主題單字完全無關）如：文章作者。

此外，在本系統內，我們並未採取此模型最後的 c-TF-IDF 與 MMR 來擷取出主題關鍵字，而是使用 CKIP Lab 的專有名詞辨識系統[4]，並且進行過濾種類並且統計。

### 2.2.3 專有名詞辨識 (Named Entity Recognition, NER)



▲圖八 BiLSTM[12]

中央研究院中文知識庫小組專有名詞辨識系統[4]主要是為了能夠擷取出輸入文章中的實體文字，其中包含了其中包含 DATE 日期、EVENT 事件、FAC 設施、GPE 行政區、LANGUAGE 語言等等。

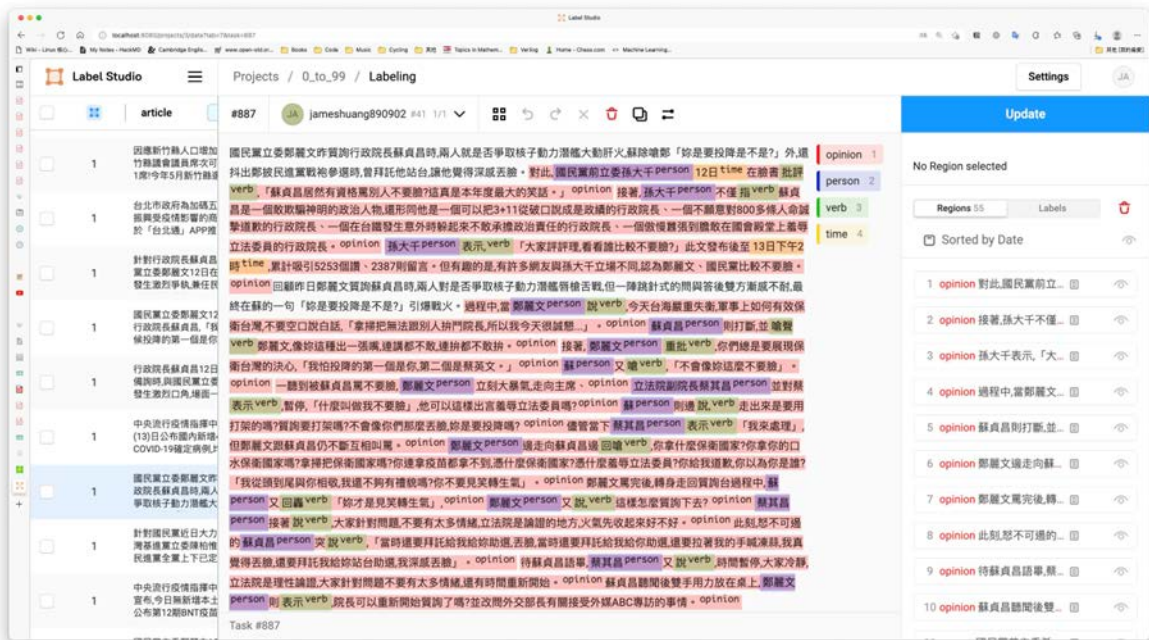
而實作共分四步驟，

1. 以 OntoNotes 中文語料作為訓練集，其中包含了各種類型的語料
2. 利用平衡語料庫等資源來訓練出字詞向量
3. 利用斷詞系統和剖析系統取出語法與語意的特徵
4. 以 BiLSTM (Bidirectional Long short-term memory) 模型為基底，進行訓練，用來預測實體位置與實體類別

## 三、實驗結果

### 3.1 驗證資料

我們的驗證資料來源為已爬取 ettoday100 篇的新聞，並將這些提取出的報導內文作為 csv 格式的欄位資料，並且使用開源的標註工具 Label Studio[13]，以人工方式對新聞中的人物、意見進行標注，而我們標註內容包含意見範圍、句中表示意見持有者表達意見的「動詞」、意見持有者（如：人名、組織與代名詞）



▲圖九 以 Label Studio 進行新聞標註的畫面範例

```
{
  "start": 143,
  "end": 260,
  "text": "蔡英文在中常會上表示，國慶日剛過，今年她針對台灣當前的內外挑戰，提出了「四個堅持」的主張，也希望朝野政黨，能夠依循這2300萬台灣人民的最大公約數，作為我們對內運行民主政治對外團結一致的共同立場，一起來因應國內外各項艱鉅挑戰。",
  "labels": ["opinion"]
},
{
  "start": 151,
  "end": 153,
  "text": "表示",
  "labels": ["verb"]
},
{
  "start": 143,
  "end": 146,
  "text": "蔡英文",
  "labels": ["person"]
}
```

▲圖十 標註後產生的 json 格式範例

## 3.2 分析結果

我們比對的方式為系統找擷取出的意見與我們手動標註的資料，確認彼此之間的意見與人物是否為相互包含的關係，只要有其一包含另一，系統所擷取的意見句就判斷為正確。

	precision	recall	f1
version 0	<b>0.689</b>	0.521	0.575
version 1	0.679	0.518	0.569
version 2	0.609	<b>0.699</b>	<b>0.639</b>

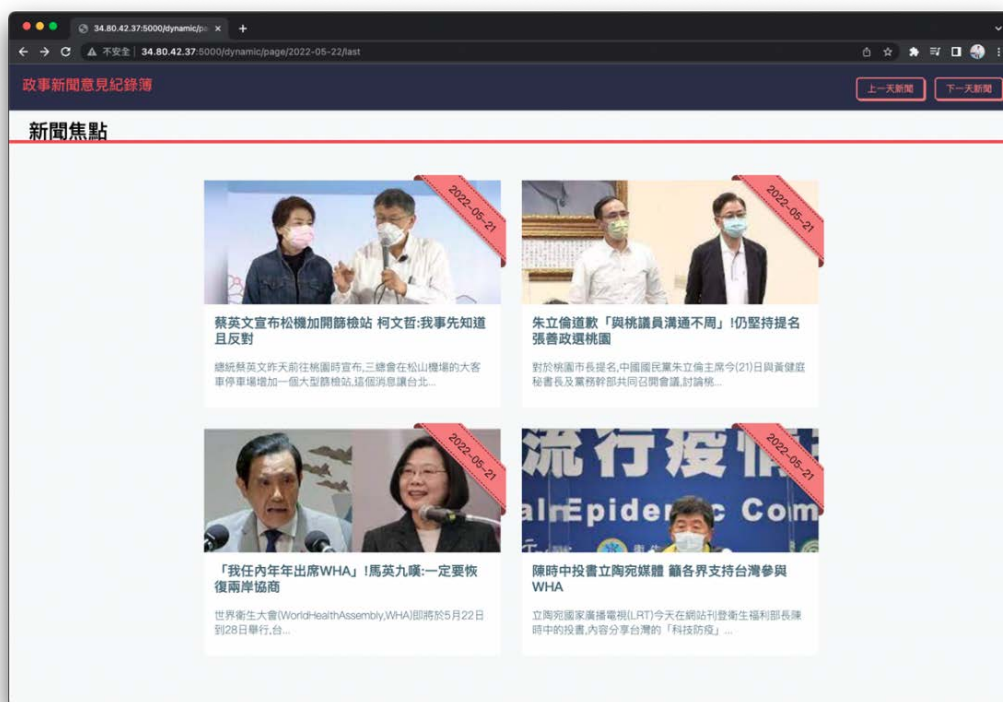
由上結果可知，在使用專有名詞辨識與動詞字典的情況下，precision 的分數會是最高，但其 recall 就會大幅下降，代表有相當大部分的意見並沒有辦法被擷取出來。

另一方面，如果使用中文剖析結構樹為主要的架構，recall 可以上升 0.178，但也因為 version 1,2 收集的意見策略改變，造成了精準度在 version 2 下降 0.08。

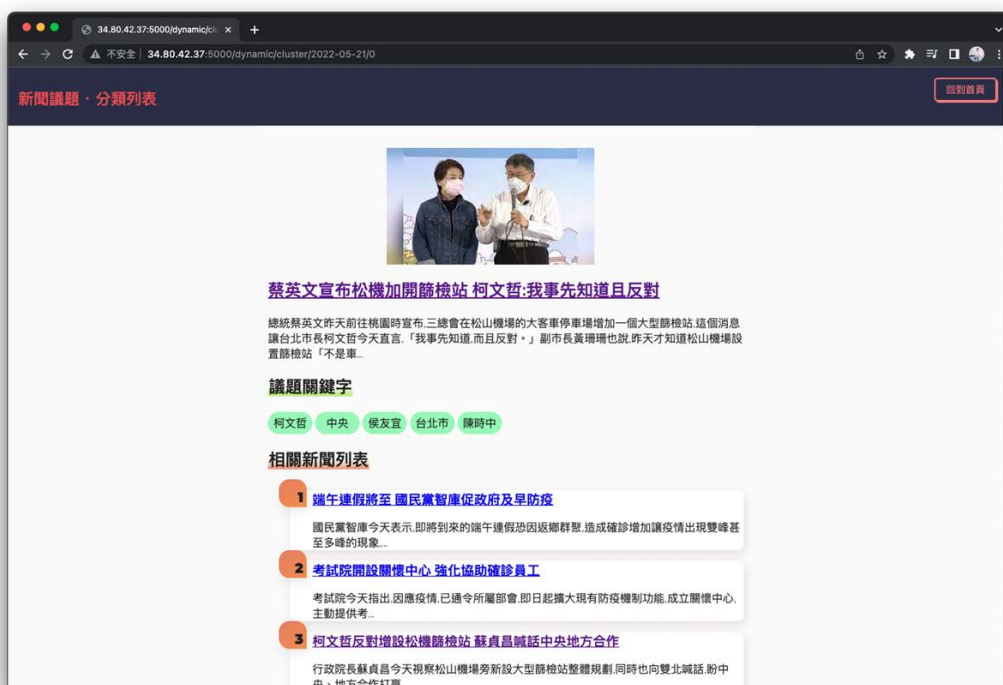
根據系統應用性質，政治人物的意見 precision 精確率的重要性會高於 recall 召回率，再加上 version 1,2 的套件運作時間會相當漫長，故在最終的實際網頁呈現上，我們團隊最後決定使用 version 0。



### 3.3 網頁呈現

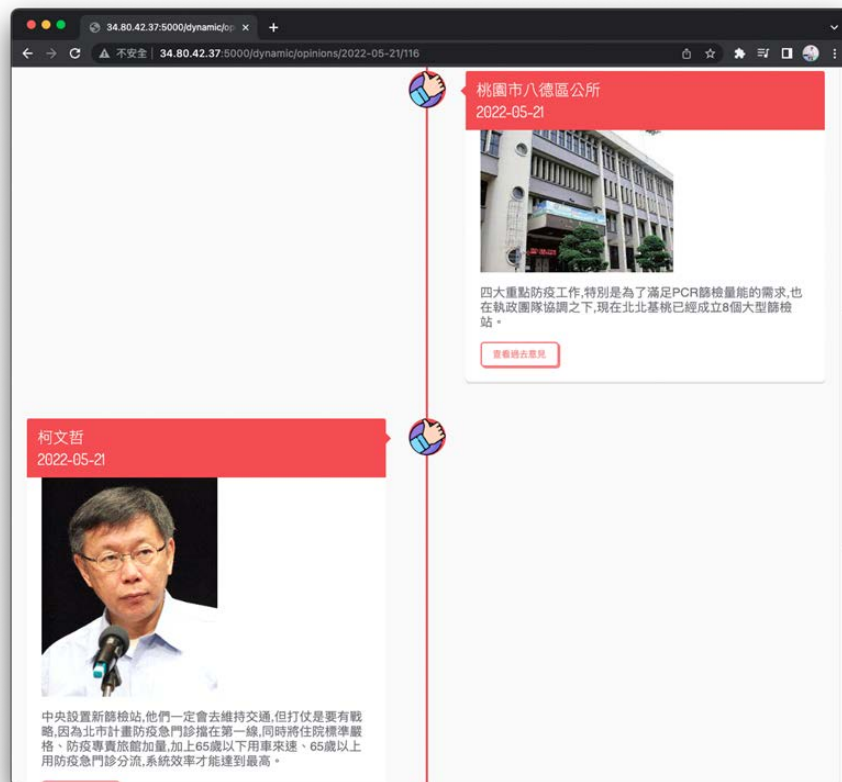
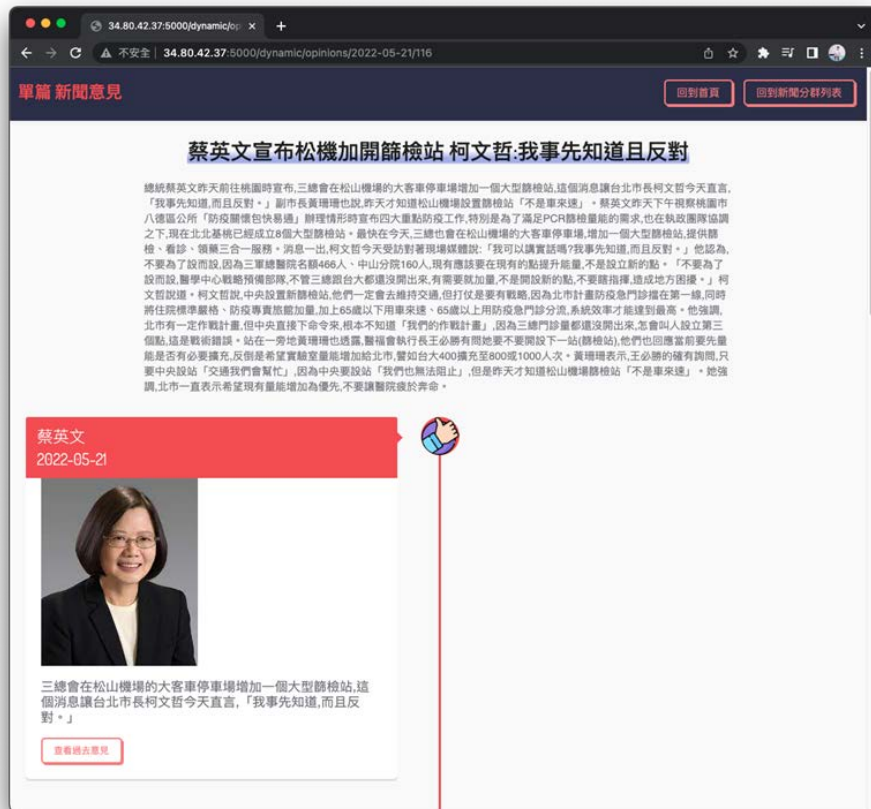


▲圖十一 首頁呈現新聞分群結果

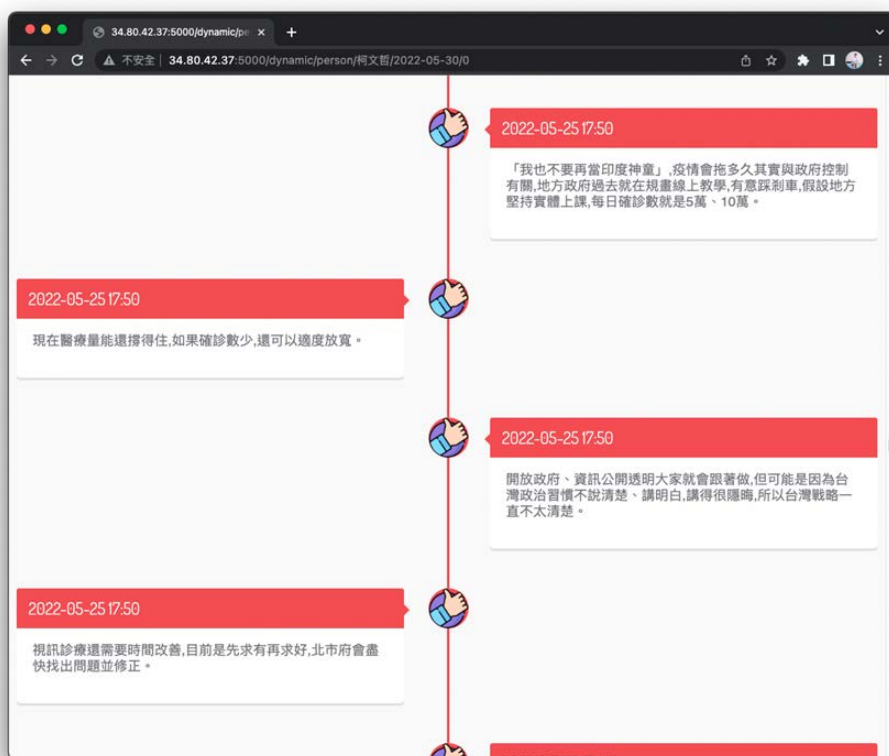


▲圖十二 當日新聞分群列表





▲圖十三 單篇新聞人物視覺化



▲圖十四 政治人物議題長期追蹤

## 四、結論

經過我們系統的設計、實作與架設，可以從網頁呈現得知，有別於傳統政治新聞，只有以單日單一事件為單位，文章以純文字結構表現，我們系統「政事新聞意見紀錄簿」確實可以讓我們自行設計的演算法去重新剖析、解構傳統新聞，以議題分群減少新聞閱讀量，並以人物視覺化方式呈現新聞內容，最後利於追蹤政治人物長期的議題意見，成功藉由整合，賦予嶄新的新聞資料價值，讓新聞的閱讀者，能以不同的角度並且橫跨時間的維度，去吸收政事新聞的政治人物發言資訊。

從實驗數據來看，藉由不同版本的演算法設計結果，與我們自行手動標記的政治新聞意見驗證，我們可以得到 precision 0.689、recall 0.699 的結果，由此可知，整合完的新聞呈現有一定的不失真性與可信度。

但未來我們也希望可以將系統做到提升演算法效能，將運算時間縮短，並且再進一步提高 precision 與 recall 兩大指標，使閱讀者在使用本系統時，能更精準與不遺漏的吸收最新資訊。

## 五、參考文獻

- [1] IKMLab (2021). "Taiwan\_news\_dataset". Retrieved from [https://github.com/IKMLab/Taiwan\\_news\\_dataset](https://github.com/IKMLab/Taiwan_news_dataset)
- [2] 張士哲 (2021 年 10 月 30 日)。談核四公投 侯友宜：沒法解決核廢料就沒資格用核電。TVBS 新聞網。檢自：<https://news.tvbs.com.tw/politics/1620143>
- [3] 杜冠霖 (2021 年 10 月 12 日)。被蘇貞昌嗆不要臉 鄭麗文臉書再飆民進黨被美國吃豆腐講都不敢講。ETtoday 新聞雲。檢自：<https://www.ettoday.net/news/20211012/2099615.htm>
- [4] Peng-Hsuan Li, Tsu-Jui Fu, Wei-Yun Ma. "[Why Attention? Analyze BiLSTM Deficiency and Its Remedies in the Case of NER](#)". AAAI, Feb 2020.
- [5] Maarten Grootendorst (2022). "[BERTopic: Neural topic modeling with a class-based TF-IDF procedure](#)". arXiv:0707.0835
- [6] Nils Reimers, Iryna Gurevych (2020). "[Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation](#)". arXiv:2004.09813
- [7] ckiplab(2022)."ckip-transformers."Retrieved from <https://github.com/ckiplab/ckip-transformers>
- [8] Yu-Ming Hsieh, Duen-Chi Yang, Keh-Jiann Chen. "[Grammar Extraction, Generalization and Specialization](#)". ROCLING, Sep 2004.

- [9] 中央研究院詞庫小組, “中文詞類分析(三版).” CKIP Technical Report No.93-05
- [10] 詞庫小組. “句結構樹中的語意角色”. No. 13-01, Jan 2013.
- [11] Wei-Fan Chen, 陳., & Lun-Wei Ku, 古. (2018). 中文情感語意分析套件 CSentiPackage 簡介. *圖書館學與資訊科學*, 44(1), 24-41. 引自: <https://jlis.glis.ntnu.edu.tw/ojs/index.php/jlis/article/view/732>
- [12] Savelie Cornegruta, Robert Bakewell, Samuel Withey, Giovanni Montana (2016). "[Modelling Radiological Language with Bidirectional Long Short-Term Memory Networks](#)".arXiv:1609.08409
- [13] Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, Nikolai Liubimov (2020-2022). "[Label Studio: Data labeling software](#)". Retrieved from <https://github.com/heartexlabs/label-studio>