產業新聞分類與關鍵字檢索

文字探勘初論 第四組期末專案

蔡逸芃	劉知微	陳承妤	郭太元
資管二	資管二	資管二	資管碩二
B11705034	B10705054	B11705027	R09725015

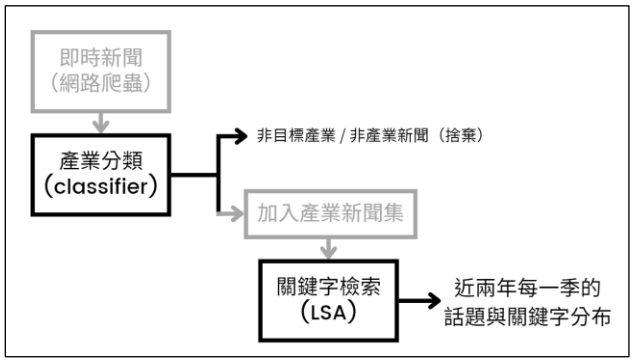
1 動機與目標

在股票市場中，產業新聞會作為消息面反映企業價值。作為投資者，不免需要了解企業所屬產業的動態，而查詢新聞又非常費時，因此想要自動化這個流程，隨時爬取最新的產業新聞並提取關鍵字、觀察近年話題分布的變化。

完整的理想系統架構如下：

1. 決定有興趣的產業類別。
2. 從數個網站爬取最新產業新聞，並記錄其發布時間。
3. 將這些新聞依產業別分類。如果該新聞與使用者有興趣的產業無關，則捨棄新聞，否則加入該產業的新聞集。
4. 對每個產業的新聞集進行 Latent Semantic Analysis，觀察近兩年每季的話題分布變化與關鍵字。

本專案專注於**「將新聞以產業別分類」**與**「關鍵字檢索」**的部分。分類與 LSA 使用的新聞集皆使用 2022~2023 年的歷史新聞，並沒有實時更新。本組有興趣的產業選定為半導體（semi）、工業自動化（auto）、電動車（ev）以及人工智慧（AI）。



圖一：理想系統架構簡圖，黑色為本專案之實作重點。

2 產業新聞分類

2.1 新聞資料

我們使用 beautifulsoup 套件，在數個網站爬取相關產業的新聞。每個產業皆有 1,500 篇以上，來源皆為兩個網站以上。網站的詳細資訊將補充於附錄。

產業	新聞數量	網站數量
AI	2,056	2
auto	1,679	4
ev	2,436	3
semi	2,593	3

表一：各產業新聞數量以及來源網站數量

2.2 分類器選擇

我們將每個產業的 90%新聞作為訓練資料，tokenize 與去除 stop word 後分別使用 Naïve Bayes、KNN (n_neighbors=5)、SVM Linear

(C=1)與 SVM RBF (C=1, gamma='scale')進行分類，並以剩餘的 10%新聞評量成效。

以下我們比較不同分類器之 macro-avg f1 與 micro-avg f1 分數。如需詳細的分類報表與 precision-recall curves 請參考附錄。

分類器	macro-avg f1	micro-avg f1
NB	0.90	0.90
KNN	0.94	0.94
SVM (Linear)	0.97	0.97
SVM (RBF)	0.96	0.96

表二：不同分類器的 f1 分數比較

明顯可以發現 **SVM Linear 表現最佳**，因此，後續分類皆以 SVM Linear 為主。

2.3 分類器的實務應用

或許使用者會覺得疑惑，有些網站之新聞分類完整，為何還需要自己訓練分類器呢？主要原因是，新聞網站的分類不一定是使用者所想要的。例如工業自動化是相對冷門的話題，一般網站可能有「自動化」之類別，但沒有「工業自動化」。訓練分類器的目的，是不論那些網站如何分類新聞，我們的系統都要能應付。

在此衍生出兩個議題：

1. 若網站的新聞分類不好，**爬取到的新聞可能與使用者有興趣的產業完全無關。**例如 Yahoo Finance 的新聞就分得亂無章法，有些甚至只是「報明牌」的新聞。我們應該將其剃除。
2. 雖然分類器的 f1 分數極佳，不過這些測試資料都來自分類器「已經看過」的網站。**對於其他「沒有看過」的網站，其分類成效又如何呢？**

2.4 額外新聞資料—來自其他網站

因此，我們額外爬取了一些產業新聞，其來源網站與章節 2.1 之新聞來源皆不同。

產業	新聞數量	網站數量
AI	25	1
auto	25	1
ev	25	1
semi	25	1
others	96	1

表三：來自其他網站新聞數量及網站數量

注意到「others」類為完全與四個產業無關之新聞。

2.5 Novelty Detection

為了對付「完全無關」的新聞，我們使用了 One-Class SVM，在對新聞進行產業分類前先行將無關的新聞標記為「新奇」，反之則標記為「正常」。One-Class SVM 是非監督式模型，利用已知的「正常資料」(Regular Data) 學習決策邊界，並以這個邊界判斷測試資料是否為「新奇資料」(Novelty)。在此專案中，用來訓練之「正常資料」為章節 2.1 之所有新聞。

我們將 One-Class SVM 的參數 nu 設為 0.1，輸入章節 2.4 之資料，並以 Linear 與 RBF kernels 測試，測試結果如下：

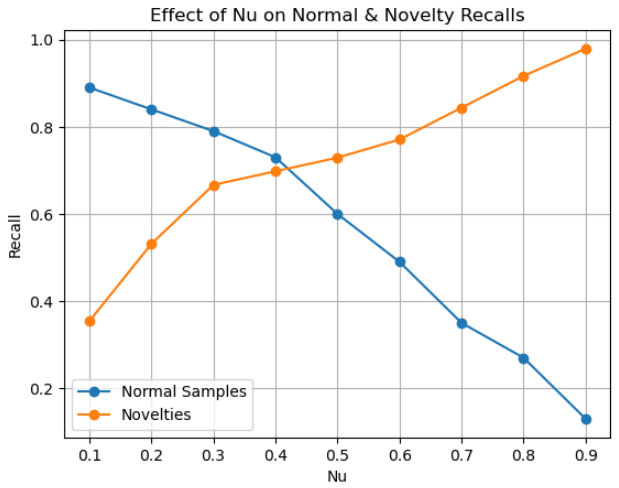
Kernel	Regular Recall	Novelty Recall
Linear	0.94	0.32
RBF	0.94	0.35

表四：One-Class SVM 之不同 Kernel 比較

Linear 的 Regular Recall 與 RBF 相同，亦即四個產業、共 100 篇的正常資料中，有 94%被分類成「正常」。而 Novelty Recall 則是 RBF

略勝過於 Linear，表示 96 篇「others」類的新聞中，有 35%被分類成「新奇」。

以下使用 RBF One-Class SVM 進行參數測試。從原始論文得知，nu 代表原始訓練資料中 outlier 比例的上界（也就是 learning error，因為訓練資料應該都要是 regular data）、support vectors 比例的下界[1]。因此，nu 設定得越大，決策邊界包含的範圍會越小、learning error 會越多，最後造成測試時 Novelty Recall 上升、Regular Recall 下降（更容易偵測出新奇資料，不過正常資料更容易被誤判成新奇資料）。我們以不同的 nu 進行測試，畫出 Novelty 和 Regular Recall 如下圖：



圖二：One-Class SVM 參數 nu 對 Novelty 和 Regular Recall 的影響，橘線為 Novelty Recall，藍線為 Regular Recall。

測試結果和推測相同，Novelty 和 Regular Recall 之間勢必要透過設定 nu 做取捨。我們不希望錯過任何資訊，因此選擇 nu 為 0.1，換取較高的 Regular Recall。

2.6 分類其他網站之新聞

最後，經過 One-Class SVM 分類出「新奇資料」，我們將剩下的資料以 SVM Linear 分類。此 SVM Linear 以章節 2.1 之所有資料進行訓練，並以章節 2.4 中未被分類成「新奇」類的新聞做為測試資料。綜合章節 2.5 與 2.6 的所有分類步驟，分類成效如下：

	precision	recall	f1	support
AI	0.70	0.92	0.79	25
auto	0.80	0.96	0.87	25
ev	0.91	0.80	0.85	25
semi	0.31	0.88	0.46	25
others	0.85	0.35	0.50	96
accuracy			0.63	196
macro avg	0.71	0.78	0.70	196
weighted avg	0.76	0.63	0.62	196

表五：分類報告—其他網站之新聞

macro-f1 為 0.70，micro-f1 為 0.62，算是勉強及格的分數。其中 semi 之 precision 特別差，原因為「未被分類成新奇」的新奇資料大多都被分類成半導體新聞。

撇除「others」類不看的話，SVM Linear 對四個產業、共 100 篇新聞分類的 f1-macro 分數為 0.92，證明分類其他網站之「正常」新聞的表現尚佳，整體分數偏低主要歸因於 Novelty Recall 過低。

3 關鍵字檢索與話題分布

3.1 方法概述

參考了一篇關於智慧工廠話題檢索的論文[2]，其使用 LDA 進行話題檢索，列出每個話題的關鍵字，並在最後以長條圖表現不同時點的話題分布。其中，話題是來自 2014~2017 的所有文章。以其研究結果來看，不同話題仍看得出消長趨勢，舊話題被新話題覆蓋的情形並不嚴重。

我們在章節 2.1 爬取的新聞發布時間皆落在 2022~2023，「話題覆蓋」的問題應該更小。因此，我們決定採用同樣方式，以章節 2.1 中全部的文章進行 topic modeling，並觀察每個 topic 的關鍵字與兩年之中每季的話題分布變化。

稍微不同的地方是，我們嘗試使用 SVD 與 LDA 進行話題檢索，並且最後使用較近代的 BERTopic 進行比較。

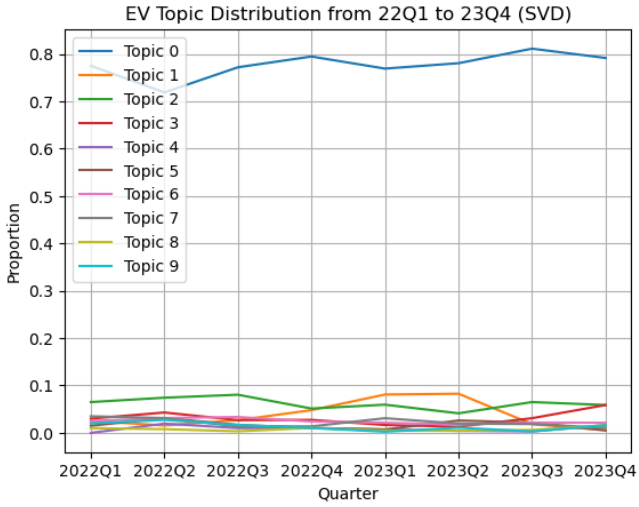
3.2 Singular Value Decomposition (SVD)

使用 SVD 分解出 10 個話題，以半導體為例：

圖三：SVD 分解出的前 6 個半導體話題

首先可以發現，SVD 的確只注重數學結果正確，每個話題之關鍵字語意其實不太相似。其次，”china”在前面 6 個話題就出現了 5 次，可見話題似乎切分得不太乾淨。

接著檢核各個話題在每一季的分布。我們將 SVD 分解出的 topic-document matrix 視為文章的話題分布，找出話題分布中絕對值最大的 topic，作為該文章的話題。接著，分別算出每個話題在一季中的文章數，再除以該季的總文章數，即可得該季的文章話題分布。例如，話題 A 在 23Q1 有 40 篇文章，話題 B 在 23Q1 有 60 篇，則 23Q1 的文章話題分布為「話題 A：0.4，話題 B：0.6」。

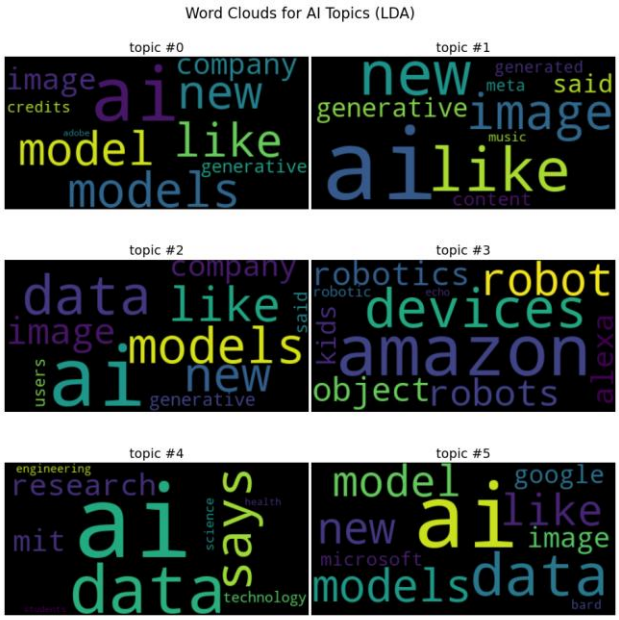


圖四：SVD 分解出的 10 個半導體話題在 22Q1~23Q4 之文章數分布

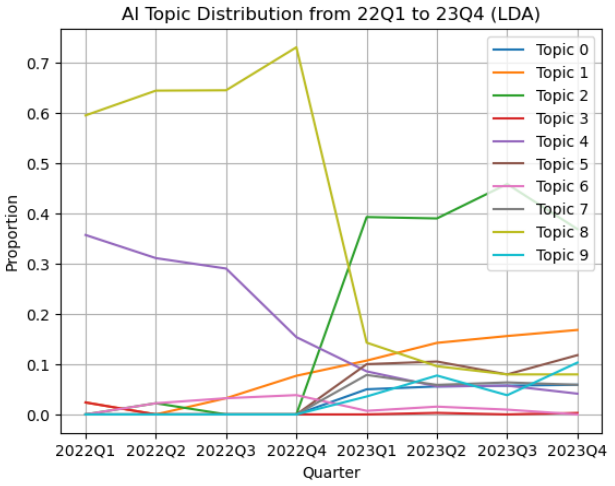
可以明顯觀察到 SVD 將話題分配得非常不平均，第一個話題就幾乎囊括了所有文章。不只半導體，在所有產業均有一樣的狀況。推測原因可能為，SVD 分解中第一個特徵值很大，因此第一個 topic 對恢復 terms-document matrix 的「貢獻」很大，自然會將所有文章歸類到 topic 0。由於 SVD 之檢索效果不符期待，在此不討論如何優化。

3.3 Latent Dirichlet Allocation (LDA)

使用 LDA，先檢索出 10 個話題。畫出不同時間點話題分布的方法與章節 3.2 大致相同，每個文章會被歸類為 topic distribution 中機率最大的主題。在此以 AI 為例：



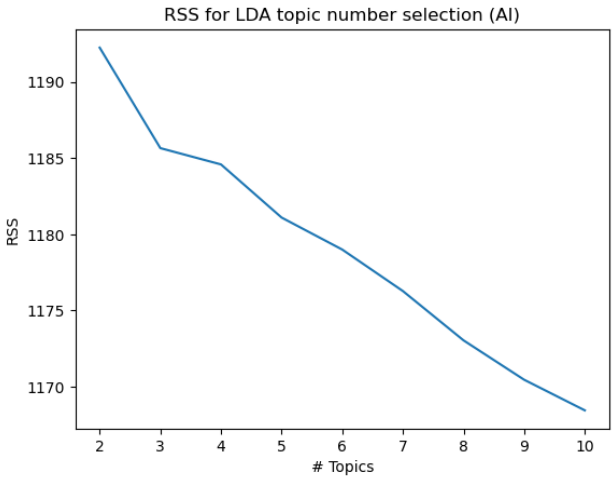
圖五：LDA 分解出的前 6 個人工智慧話題



圖六：LDA 分解出的 10 個人工智慧話題在 22Q1~23Q4 之文章數分布

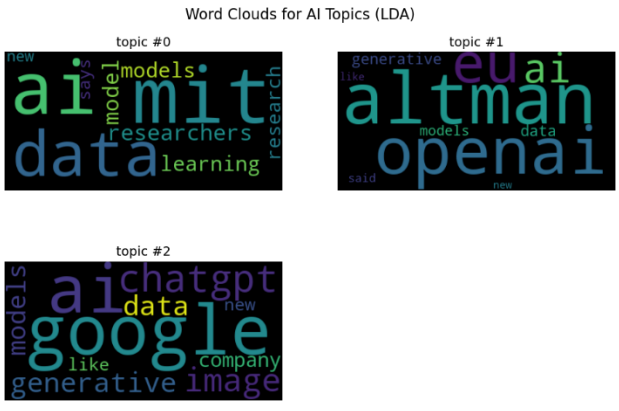
相較於 SVD，LDA 較沒有某個主題獨占鰲頭的問題，不過前 6 個話題仍然重疊性高。參考論文同樣也是使用 LDA，不過話題重疊性低，原因在於話題數量的調整。其使用 cosine similarity 來評量不同話題數量下，屬於不同話題文章的相異度[2]。

在此我們使用不同的方式，利用 Residual Sum of Squares (RSS)與 Elbow Method 來決定話題的數量。文章被歸類完主題後，每個主題成為一個 cluster，依照 K-means 的方式計算 RSS，並取「RSS 不會有大幅度減少」的話題數作為最佳話題數。以下以 AI 之 RSS 折線圖為例：

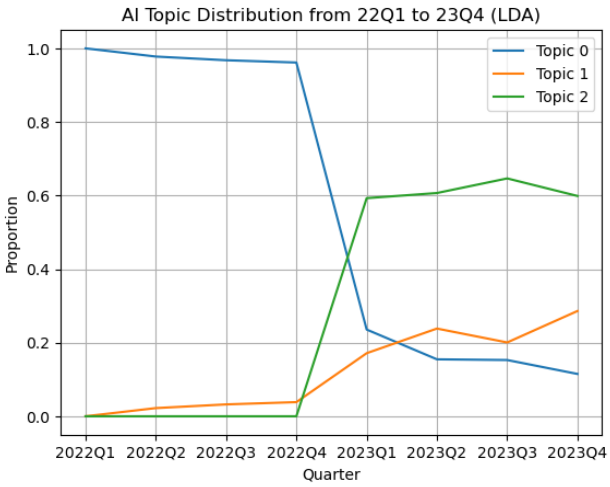


圖七：AI 之 LDA 話題數與 RSS 折線圖

從圖中發現，有鑑於話題數從 3 增加到 4 時，RSS 幾乎不變，話題數應設為 3。以下為話題數為 3 的 LDA 輸出：



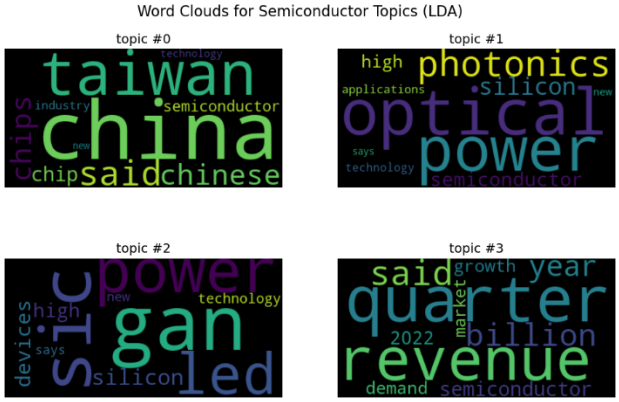
圖八：LDA 分解出的 3 個人工智慧話題



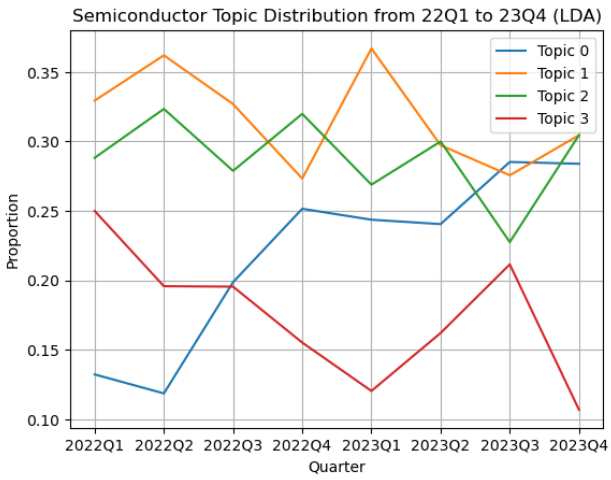
圖九：LDA 分解出的 3 個人工智慧話題在 22Q1~23Q4 的話題分布

經過 Elbow Method 的話題數調整，話題重疊性明顯降低許多。能清楚看到話題 0 與 MIT 和研究相關，話題 1 以 OpenAI、生成式 AI 以及歐洲為主，話題 2 則是 Google 和影像處理相關。話題 1 和 2 近年都有比例上升趨勢，推測和歐洲祭出的 AI 管制政策以及 Google 戮力追趕 OpenAI 有關。

半導體的 LDA 輸出也不錯：



圖十：LDA 分解出的 4 個半導體話題



圖十一：LDA 分解出的 4 個半導體話題在 22Q1~23Q4 的話題分布

話題 0 與兩岸有關，話題 1 為半導體的發光應用，話題 2 為半導體與發光二極體原料，話題 3 與半導體企業財務表現與整體市場需求有關。可以看到話題 1、2 常駐為產業熱門話題，證明發光元件在半導體產業中的地位不減；而話題 0 則在近兩年一路攀升，推測與台積電和半導體禁令有關。

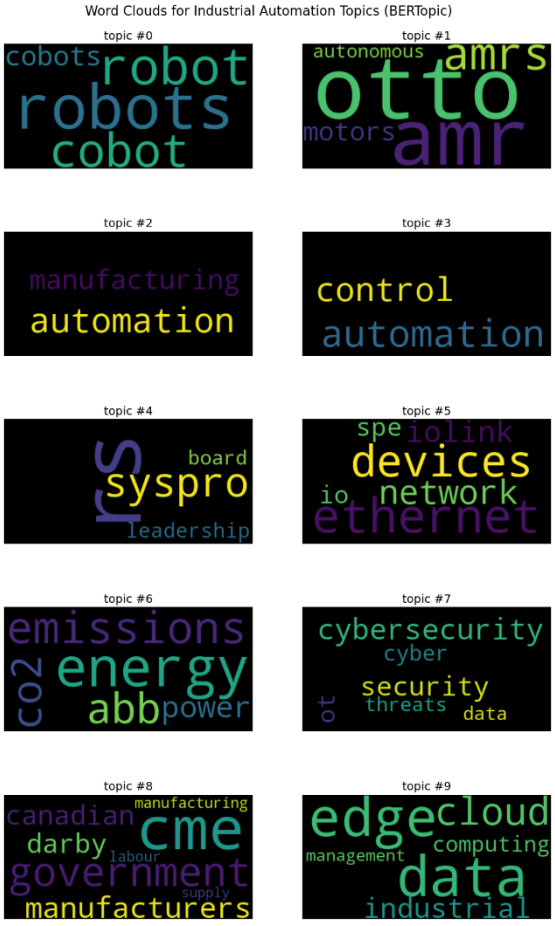
若對其他產業的 LDA 輸出有興趣，可參考附錄。

3.4 BERTopic

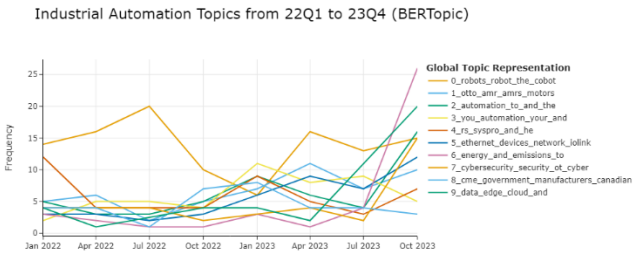
SVD 與 LDA 之共同問題為，並不考慮文字本身的語意，因此提取出的話題關鍵字有時在語意上並不相關。BERTopic 的出現解決了此問題。

從原始論文得知[3]，BERTopic 利用 BERT 先取出 document embeddings，縮減維度後進行分群，再利用 class-based TF-IDF 取出每個群（class）中重要的關鍵字。利用 BERT 的好處為，文章會被依照「語意」進行分群，且不需要特別移除 stop word（若刻意移除 stop word，document embeddings 反而會失真），也不需要特別指定分群數量（若強制指定，分群結果會非常糟糕）。

Python 中的 BERTopic 套件本身就有提供畫出不同時期話題分布的函式，不過縱軸為 Frequency，並沒有像章節 3.2 與 3.3 一樣進行每個時點的 normalization。以下為工業自動化之輸出：



圖十二：BERTopic 分解出的前 10 個工業自動化話題



圖十三：BERTopic 分解出的前 10 個工業自動化話題在 22Q1~23Q4 的話題分布

語意相近的明顯案例為話題 0（機器人）、話題 5（網路與裝置）、話題 6（永續議題）、話題 7（資安）以及話題 9（邊緣與雲端運算）。由話題分布折線圖知，每一年機器人相關話題都是主流，而永續發展議題在近期竄高。

電動車的關鍵字檢索成效也不錯：

圖十四：BERTopic 分解出的前 10 個電動車話題

話題 0 為電池、話題 1 為各種品牌、話題 4 與環保相關、話題 9 甚至分出了電動腳踏車。BERTopic 除了訓練時間較久以外，關鍵字語意的表現遠遠勝過 SVD、LDA 等方法。

其他產業之輸出結果將置於附錄。

4 結論

首先，我們比較了不同分類器，發現 **SVM Linear** 對於產業文章之分類效果最佳，f1 分數能夠達到 0.97，推測這些新聞的分布都是線性可分割，不需要投射到高維向量。

為了在實務上應用分類器，我們利用 **One-Class SVM** 先行篩選掉與有興趣的產業毫不相關的文章。關於參數的選擇，我們決定設定 nu 為 0.1，提高 Regular Recall，缺點為 Novelty Recall 僅有 0.35，拖累了整體分類的評分。此處端看系統使用者的選擇，如果不想要漏訊，就把 nu 設低；如果想要更乾淨的產業新聞集，可以考慮將 nu 設定在 Regular Recall Curve 和 Novelty Recall Curve 的交點。

接下來，我們使用不同的 **Latent Semantic Analysis** 模型與較新的 BERTopic 進行 topic modeling 與關鍵字檢索。**SVD** 的表現因為奇異值分解的數學特性與不可解釋性，較為差強人意。**LDA** 在未調整話題數量時，表現同樣不佳，不過透過 Elbow Method 調整話題數後，話題品質尚佳，也能看出一些重要話題趨勢。

最後使用的 **BERTopic** 受惠於 document embeddings，能夠檢索出語意極為相近的關鍵字以及重要話題趨勢，不過訓練時間較長。對於關鍵字檢索，我們認為 **LDA 與 BERTopic 皆為好用的利器**，能夠產出不同的實用 topic 供系統使用者參考。

綜觀系統表現，這樣的產業新聞分類與關鍵字檢索功能可供一般大眾使用，卻還沒辦法應用於產業投資。**檢索出的關鍵字與話題趨勢大多已廣為人知，較細部的資訊還是要手動查詢。**希冀未來能夠改進此系統，達到發掘產業新聞話題的真正自動化。

5 參考資料

[1] Bernhard Schölkopf, Robert Williamson, Alex Smola, John Shawe-Taylor, and John Platt. 1999. Support Vector Method for Novelty Detection. In Proceedings of the 12th International Conference on Neural Information Processing Systems (Denver, CO) (NIPS'99). MIT Press, Cambridge, MA, USA, 582–588. https://dl.acm.org/doi/10.5555/3009657.3009740

[2] Jung, Y. S., & Chang, T. W. (2018). Text mining based online news analysis about smart factory. ICIC Express Letters, Part B: Applications, 9(6), 559-565. http://www.icicelb.org/ellb/contents/2018/6/elb-09-06-11.pdf

[3] Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794. https://arxiv.org/abs/2203.05794

6 附錄

6.1 章節 2.1 之新聞來源

產業	來源網站
AI	MIT News TechCrunch
auto	Manufacturing Automation automation.com Control Automation Industrial Equipment News
ev	Green Car Reports eVehicle Technology InsideEVs
semi	SIA IEEE Spectrum Asia Financial Semiconductor Today

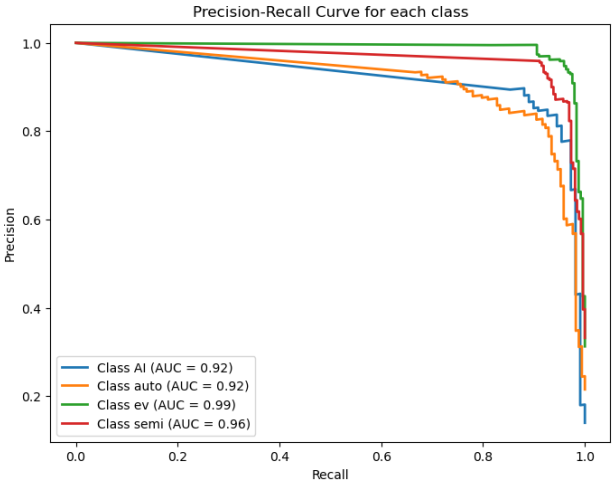
6.2 章節 2.4 之新聞來源

產業	來源網站
AI	Euronews
auto	Rockwell Automation
ev	Global News
semi	NDTV
others	Yahoo Finance

6.3 Classification Reports & Precision-Recall Curves

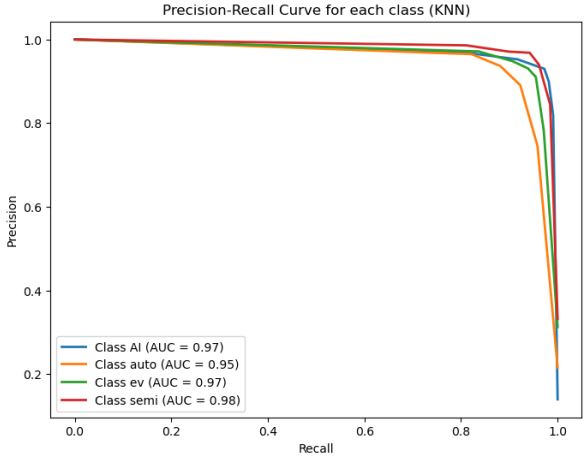
6.3.1 Naïve Bayes

	precision	recall	f1	support
AI	0.87	0.89	0.88	206
auto	0.87	0.82	0.85	168
ev	0.98	0.91	0.94	244
semi	0.87	0.96	0.91	259
accuracy			0.90	877
macro avg	0.90	0.89	0.90	877
weighted avg	0.91	0.90	0.90	877



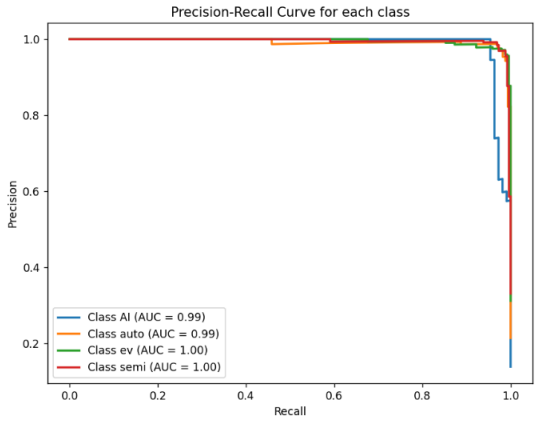
6.3.2 KNN(n_neighbors=5):

	precision	recall	f1	support
AI	0.92	0.98	0.95	206
auto	0.93	0.91	0.92	168
ev	0.93	0.94	0.93	244
semi	0.97	0.94	0.95	259
accuracy			0.94	877
macro avg	0.94	0.94	0.94	877
weighted avg	0.94	0.94	0.94	877



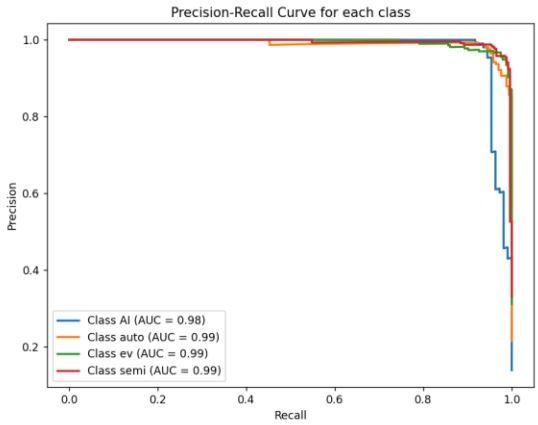
6.3.3 SVM linear:

	precision	recall	f1	support
AI	0.96	0.95	0.96	206
auto	0.98	0.97	0.98	168
ev	0.97	0.99	0.98	244
semi	0.98	0.97	0.97	259
accuracy			0.97	877
macro avg	0.97	0.97	0.97	877
weighted avg	0.97	0.97	0.97	877



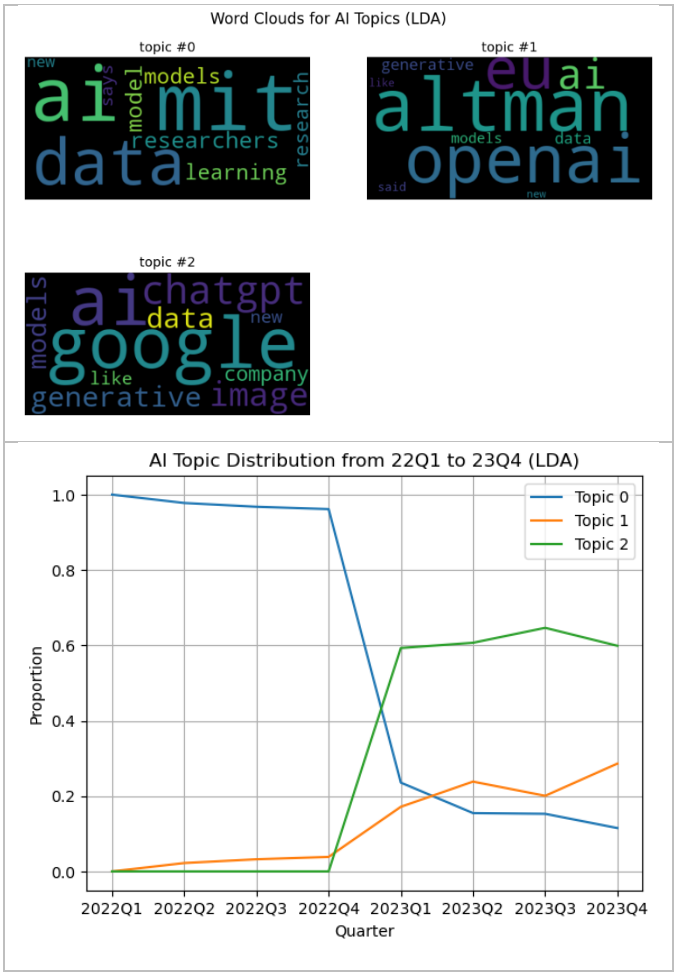
6.3.4 SVM RBF:

	precision	recall	f1	support
AI	0.96	0.94	0.95	206
auto	0.98	0.95	0.96	168
ev	0.96	0.98	0.97	244
semi	0.96	0.97	0.97	259
accuracy			0.96	877
macro avg	0.97	0.96	0.96	877
weighted avg	0.96	0.96	0.96	877

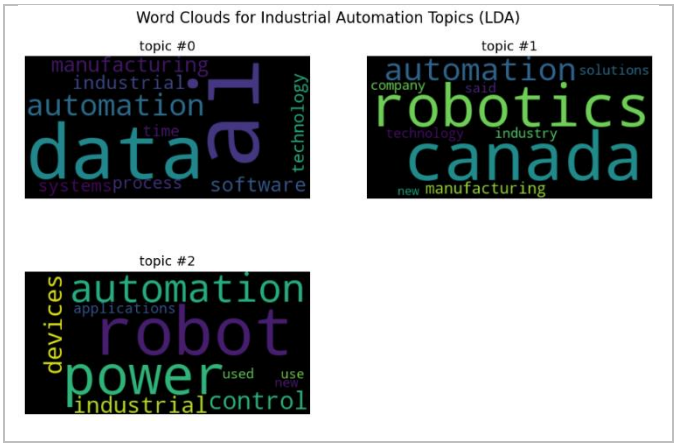


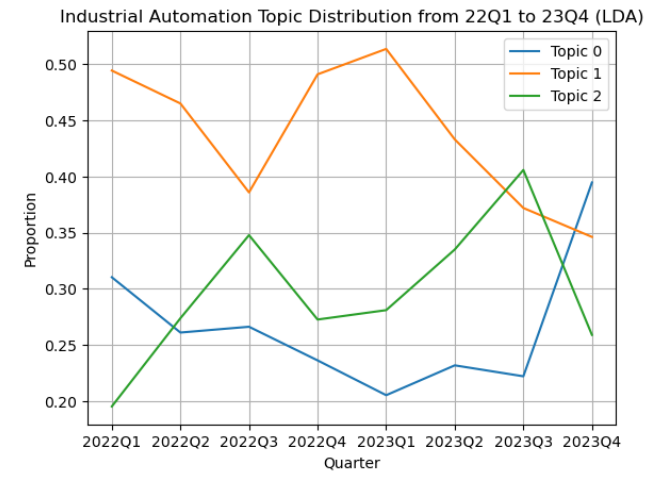
6.4 LDA Topic Modeling

6.4.1 AI

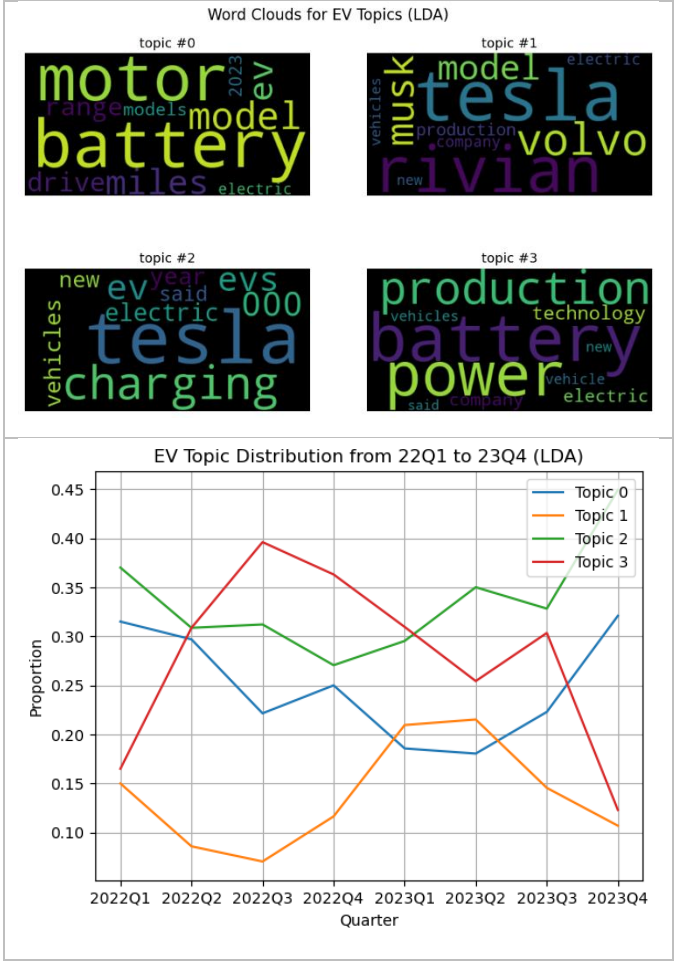


6.4.2 auto

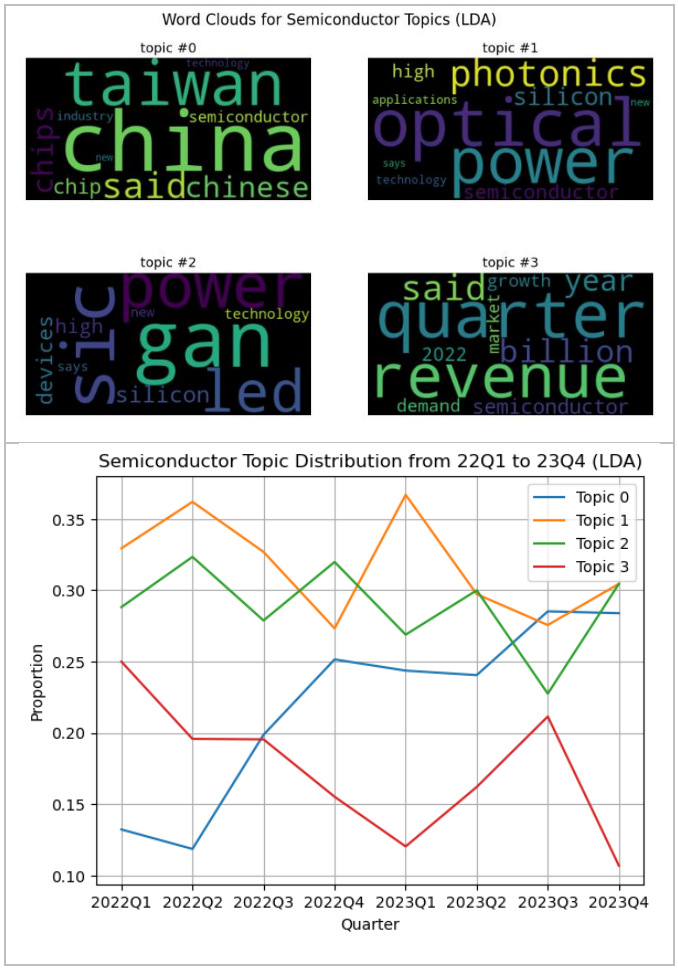




6.4.3 EV

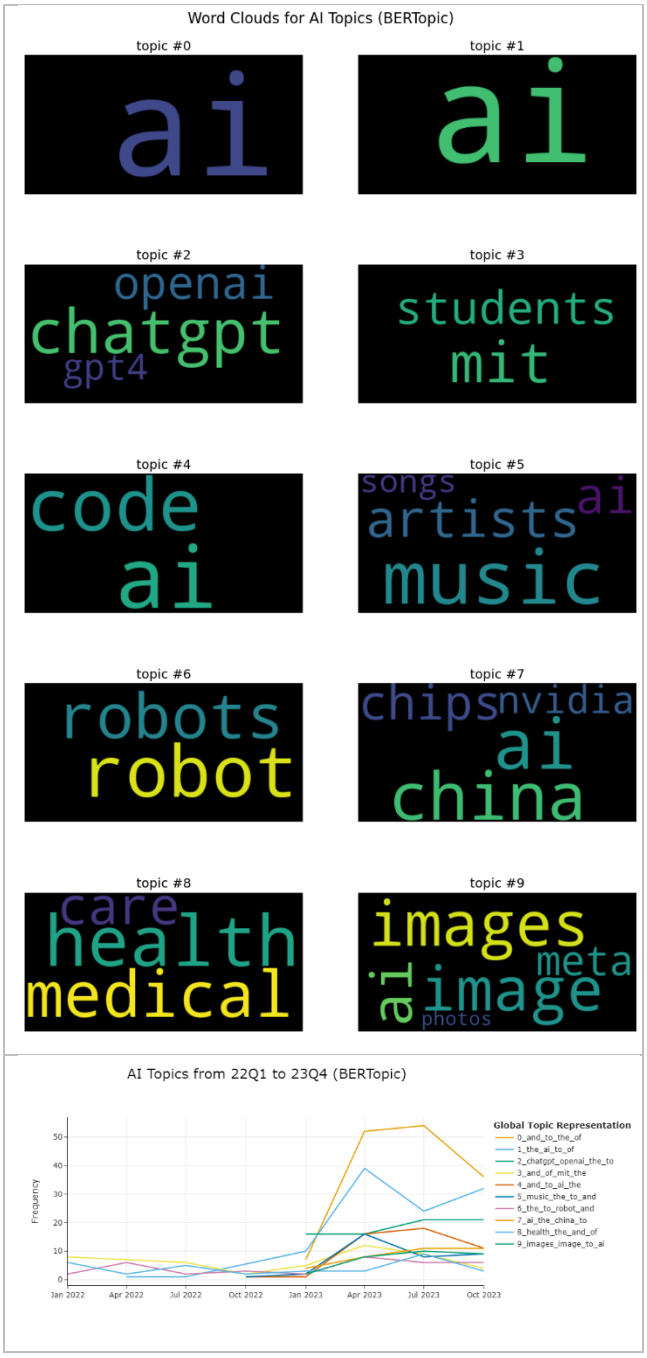


6.4.4 semi



6.5 BERTopic Topic Modeling

6.5.1 AI



6.5.2 auto

6.5.3 EV

6.5.4 semi


Word Clouds for Semiconductor Topics (BERTopic)
topic #0
integrated
photonic
laser
optical
coherent
sivers
photonics
topic #1
amplifier
microwave
amplifiers
5g
rf
mmwave
applications
topic #2
us
sia
semiconductor
research
act
industry
topic #3
solutions
electric
sic
silicon
onsemi
power
elitesic
modules
carbide
topic #4
power
ics
charging
charger
navitas
energy
gan
ganfast
gansense
topic #5
sensing
trumpf
vcsel
vcsels
automotive
laser
applications
lidar
topic #6
said
china
taiwan
tsmc
tsai
us
island
taiwans
topic #7
ns
ledsled
uvc
crayonano
water
disinfection
uv
topic #8
chip
billion
japan
plant
taiwan
tsmc
said
topic #9
samsung
korean
hynix
skus
china
ymtc
chip
memory
south

Semiconductor Topics from 22Q1 to 23Q4 (BERTopic)
Frequency
0
5
10
15
20
25
Jan 2022
Apr 2022
Jul 2022
Oct 2022
Jan 2023
Apr 2023
Jul 2023
Oct 2023
Global Topic Representation
0_optical_photonics_and_integrated
1_rf_5g_microwave_and
2_semiconductor_act_sia_industry
3_sic_power_carbide_electric
4_navitas_power_gan_ganfast
5_vcsel_lidar_trumpf_vcsels
6_taiwan_taiwans_china_said
7_uvc_leds_led_water
8_tsmc_japan_said_billion
9_sk_south_us_china