113 學年度

資料探勘

Survey Project

Mobile & Social Network

第 6 組

313553028 多工所 王美綺

513557003 資訊專 王琪涵

Content

Chapter 1. A list of all surveyed papers

Chapter 2. Brief Introduction to each paper

Chapter3. A Summarized Table

Chapter4. Comparative Study & Discussions

Chapter 1.

A list of all surveyed papers

 # 1.

Bridging Performance of X (formerly known as Twitter) Users: A Predictor of Subjective Well-Being During the Pandemic

Author: Ninghan Chen, Xihui Chen, Zhiqiang Zhong, Jun Pang

Year: 2024

Source: ACM Transactions on the Web, Volume 18, Issue 1, Article No. 15, Pages 1–23.

DOI: https://dl.acm.org/doi/10.1145/3635033

 # 2.

Partial Data, Potential Exposure: Evaluating Privacy Leakage via GNNExplainer on Social Networks

Author: Liang-Jen Huang, Cheng-Te Li

Year: 2024

Source: IEEE International Conference on Consumer Electronics – Taiwan (ICCE-TW)

DOI: 10.1109/ICCE-Taiwan62264.2024.10674199

Note: Best Paper Honorable Mention

 # 3.

Data-Mining of Social Media Users with Embedding Techniques and Neural Network

Author: Hengshan Cao, Mengjiao Yin, Ying Xi

Year: 2024

Source: IEEE 6th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)

DOI: 10.1109/IMCEC59810.2024.10575693

 # 4.

Dual Graph Networks with Synthetic Oversampling for Imbalanced Rumor Detection on Social Media

Author: Yen-Wen Lu, Chih-Yao Chen, Cheng-Te Li

Year: 2024

Source: WWW '24: Companion Proceedings of the ACM Web Conference 2024, Pages 750–753

DOI: https://doi.org/10.1145/3589335.3651494

 # 5.

Interactive Activities Initiation through Retrieving Hidden Social Information Networks

Author: Yulong Song, Bin Fu, Jianxiong Guo, Xiaofeng Gao

Year: 2023

Source: 2023 IEEE International Conference on Data Mining (ICDM)

DOI: 10.1109/ICDM58522.2023.00063

 # 6.

ANTI-Disinformation: An Adversarial Attack and Defense Network Towards Improved Robustness for Disinformation Detection on Social Media

Author: Kuan-Chun Chen, Chih-Yao Chen, Cheng-Te Li

Year: 2023

Source: 2023 IEEE International Conference on Big Data (BigData)

DOI: 10.1109/BigData59044.2023.10386090

 # 7.

A Machine Learning Based Social Network Data Mining System for Better Search Engine Algorithm

Author: Sheik Erfan Ahmed Himu, Arafat Ibne Ikram, Kazi Mohammed Abdullah, Tasnia Fahrin Choity, Md. Shahazan Parves, Md. Rabiul Hasan

Year: 2023

Source: 2023 IEEE 9th International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)

DOI: 10.1109/WIECON-ECE60392.2023.10456428

 # 8.

Classifying Severe Weather Events by Utilizing Social Sensor Data and Social Network Analysis

Author: Hussain Otudi, Shelly Gupta, Nouf Albarakati, Zoran Obradovic

Year: 2023

Source: ASONAM '23: Proceedings of the International Conference on Advances in Social Networks Analysis and Mining, Pages 64–71

DOI: https://doi.org/10.1145/3625007.3627298

 # 9.

Social Media Sensors for Weather-Caused Outage Prediction Based on Spatio–Temporal Multiplex Network Representation

Author: Rafaa Aljurbua, Jumanah Alshehri, Abdulrahman Alharbi, William Power, Zoran Obradovic

Year: 2023

Source: IEEE Access, Volume 11, Pages 125883–125896

DOI: 10.1109/ACCESS.2023.3327444

 # 10.

Unsupervised Post-Time Fake Social Message Detection with Recommendation-aware Representation Learning

Author: Shao-Ping Hsiao, Yu-Che Tsai, Cheng-Te Li

Year: 2022

Source: WWW '22: Companion Proceedings of the Web Conference 2022, Pages 232–235

DOI: https://doi.org/10.1145/3487553.3524259

 # 11.

Predicting and Analyzing Privacy Settings and Categories for Posts on Social Media

Author: Hsin-Yu Chen, Cheng-Te Li

Year: 2022

Source: 2022 IEEE International Conference on Big Data (Big Data)

DOI: 10.1109/BigData55660.2022.10020677

 # 12.

Social Network Analysis of Popular YouTube Videos via Vertical Quantitative Mining

Author: Adam G.M. Pazdor, Carson K. Leung; Thomas J. Czubryt, Junyi Lu, Denys Popov, Sanskar Raval

Year: 2022

Source: 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

DOI: 10.1109/ASONAM55673.2022.10068640

 # 13.

Finding Potential Propagators and Customers in Location-Based Social Networks: An Embedding-Based Approach

Author: Yi-Chun Chen, Cheng-Te Li

Year: 2020

Source: Applied Sciences, Volume 10, Issue 22

DOI: https://doi.org/10.3390/app10228003

 # 14.

Data Mining and Feature Analysis of College Students' Campus Network Behavior

Author: Liu Kesheng, Ni Yikun, Li Zihan, Duan Bin

Year: 2020

Source: Source: 2020 5th IEEE International Conference on Big Data Analytics (ICBDA)

DOI: 10.1109/ICBDA49040.2020.9101257

 # 15.

Traffic Information Mining From Social Media Based on the MC-LSTM-Conv Model

Author: Yiwen Wang, Zhi He, Jie Hu

Year: 2020

Source: IEEE Transactions on Intelligent Transportation Systems, Volume 23, Issue 2, Pages 1132–1144.

DOI: 10.1109/TITS.2020.3021096

Chapter 2.

Brief Introduction to each paper

#1. Bridging Performance of X (formerly known as Twitter) Users: A Predictor of Subjective Well-Being During the Pandemic

I. Target Problem

COVID-19 的爆發加劇了社交媒體上假訊息的氾濫，而具有影響力的社群媒體用戶被認為能有效降低這類假訊息傳播的影響。

在過去的研究中，特別關注弱勢族群 (例如:移民或是醫療人員)在 COVID-19 下的主觀幸福感 (subjective-well-being) 的影響。但還沒有研究調查這些幫忙降低假訊息的具有影響力的社群用戶的主觀幸福感。在研究這群體會遇到兩大挑戰:

1. 缺乏能準確化社群媒體用戶在促進 COVID-19 相關資訊傳播過程中實際作為連接橋的表現 (Bridging Performance) 的測量工具。
2. 難以獲取同時擁有連接橋表現資料和主觀幸福感評估並且足夠數量的大規模社群媒體用戶資料。

此篇研究基於想了解具有影響力的社群用戶在 COVID-19 下的心理影響，因此需重新構建衡量用戶與社群媒體的橋連接 (bridge performance)，並以此揭示主觀幸福感和橋連接的相關性。

II. Dataset

- **Dataset Link:**
  https://github.com/NinghanC/SWB4Twitter
- **Dataset characteristics**
  資料集為此論文建立，建立方式如下，時間為 2019-2021 的 X (Twitter)作為資料來源，並且分成社群網路 (Social Network) 和貼文的時間軸 (Timeline Tweets)，其中在貼文時間軸中包含用戶 (user)、在 COVID-19 前發的貼文 (tweet before COVID)、COVID-19 期間的貼文 (tweet during COVID)、COVID-19 前每個用戶的貼文 (tweet per user before COVID)、COVID-19 期間每個用戶的貼文 (tweet per user during COVID)。

III. Data Mining workflow and main techniques

1. Data Processing

此步驟為清洗資料並獲得訊息傳播的連結關係和貼文情緒分析。首先，將資料集中貼文的時間軸 (Timeline Tweets) 轉換為級聯樹 (cascade tree) 的形式，以記錄訊息傳播的過程。接著，原始資料集中未針對貼文的情緒進行分類，因此建立了一個端對端的模型來進行分類，將貼文分類為負面、中性或正面。此外，在將原始資料輸入情緒分類模型前，會先刪除所有 URL 和並清理包含使用者名稱的貼文。最後，透過乘上一個量級 (Scale) ，將原本的主觀幸福感 (subjective-well-being) 擴展為三個情緒標籤(即包含消極、積極、中性情感)。

2. Bridging Performance of Individual Users in Information Diffusion and Its Relation with SWB

此步驟為衡量使用者在傳播 COVID-19 資訊方面的連接橋表現 (Bridging Performance)。將上個步驟獲得的級聯樹 (cascade tree)，計算使用者的橋連接值，並藉此定義使用者橋連接幅度，來評估他人觀察到此訊息在傳播中整體的重要性，透過此種測量方式了解哪一個使用者在資訊傳播中發揮更重要的作用，藉以獲得影響力社群用戶。

3. Relation between SWB and Bridging Performance

此步驟將探索由上步驟建立的使用者的橋連接關係是否與積極參與 COVID-19 相關資訊傳播的使用者的主觀幸福感(subjective-well-being) 變化之間存在關係。評估指標包含 UBM 和 in-degree, PageRank , XRank, betweenness centrality, community centrality。

4. Comparing the Bridging Performance of User Subgroups

將使用者分成多個子集去進行分析，藉以了解各個使用者子群集所扮演的角色。

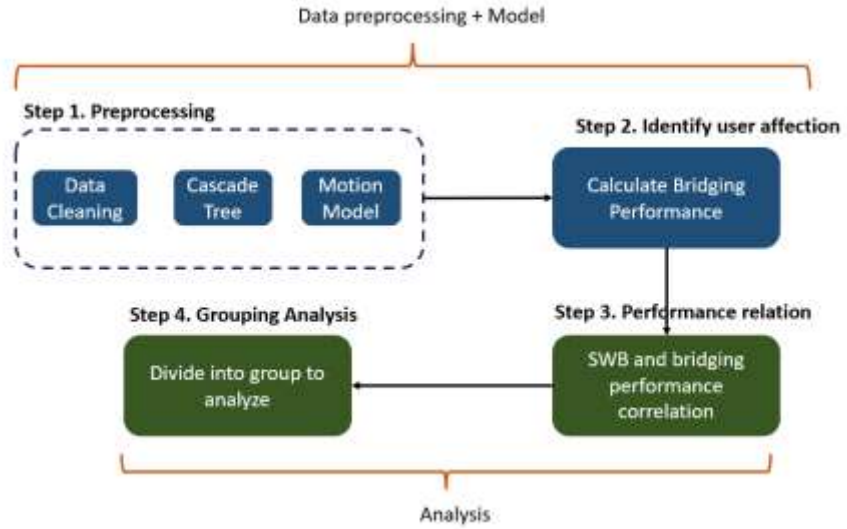

圖 1. The research of overall workflow

IV. Result

1. 此研究提出新的量測方法，藉以量化使用者和使用者的橋連接效能。(如圖 1 workflow 顯示，將 step1 和 step2 做特徵轉換，並拿到相關資料，在 step3 和 step4 做計算和分析)。
2. 在針對個人使用者的測量，實驗結果顯示在疫情期間使用者在訊息傳播方面的橋連接表現與其主觀幸福感之間存在強烈負相關關係。
3. 透過新的測量方法，實驗結果顯示有影響力的使用者和子群體的主觀幸福感下降幅度相對其他使用者更大。
4. 透過對第 4 步驟，針對子群體的橋連接績效測量，重新確認了多語言使用者在社群媒體上傳播訊息的橋接作用，並說明了多語言使用者的主觀幸福感與其橋連接績效之間呈負相關關係。(表 1)

表 1. 多語言使用者的主觀幸福感與其橋連接績效

Operation	User group	#cascade	SBM ρ
	multilingual	20,862	**0.7843**
Maximum	monolingual	5,739	0.2011
	multilingual	20,843	**0.7835**
Median	monolingual	5,758	0.1994
	multilingual	20,849	**0.7838**
Mean	monolingual	5,752	0.2013

#2. Partial Data, Potential Exposure: Evaluating Privacy Leakage via GNNExplainer on Social Networks

I. Target Problem

隨著社群媒體的興起，使用者往往在不經意間披露了過多關於自己的資訊。雖然這些資訊可能並未直接包含私密資訊，表面看似安全，但其中是否有藏有潛在資訊披露的風險。

此研究探討的核心問題是: 當只能合法獲取使用者的部分資料（例如透過特定查詢方法）時，是否能藉此披露使用者之間更廣泛的關聯性，進一步推測並獲取其更為私密的資訊。

II. Dataset

- **Dataset Name**
  Facebook, GitHub
- **Dataset Link**
  https://github.com/benedekrozemberczki/MUSAE?tab=readme-ov-file#datasets

III. Data Mining workflow and main techniques

此篇研究利用 GNNExplainer 生成可解釋性子圖 (Explanatory subgraphs)來描繪透過能夠過查詢取得的部分使用者資料，並通過這些子圖去預測使用者間的關聯性。

1. GNN Model Training

訓練一個基礎的 GNN 模型，使得在 GNNExplainer 生成子圖時，能夠解釋此模型的預測結果。

2. Generate explanatory subgraphs and New Dataset Synthesis

GNNExplainer 主要是透過 mask-out 的方法來識別對模型預測能起到關鍵作用的點和邊，並將他們組合成可解釋性子圖 (Explanatory subgraphs)。另外，此研究有特別強調強調子圖的連接性，若是缺乏邊的連接(不連通的子圖)，則會重新選擇。並且通過結合多個子圖 (這些子圖都是基於相同的 k 個節點生成) ，保留所有重要的的節點特徵和成一個更大的新合成圖，且將這些合成圖也放進資料集成為新的資料集。

3. Reconstruct the original graph using explanatory subgraphs

首先，使用 GNN 模型為每個點學習嵌入表示 (embeddings)。接著，透過點的嵌入量計算內積，如果內積的值越高，代表點的相似度越高，有邊的機率也越高，藉此重建出可能的原圖。

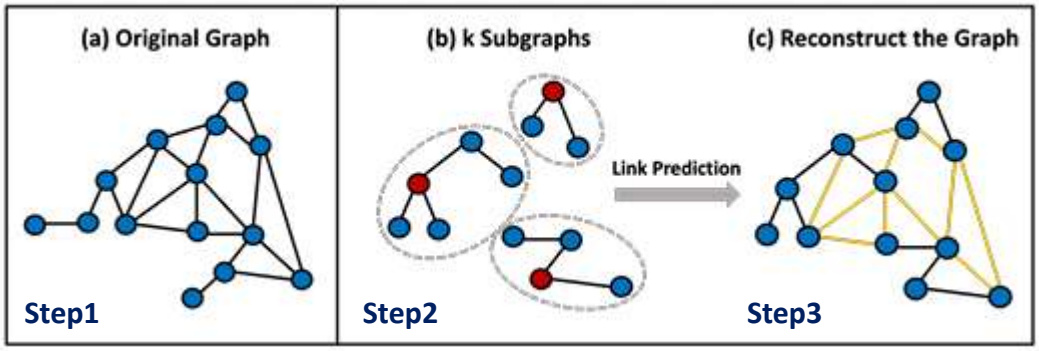

圖 2. The research of overall workflow

IV. Result

1. 使用 GCN、GAT、GraphSAGE 模型中，Facebook 資料集的 ACU 都有達到約 90%。在 GitHub 的資料集中以 GCN 模型表現最佳，有達到 84.97%。從以上實驗結果可以表明可解釋性子圖 (Explanatory subgraphs)，可以接露更多的關係。
2. 從準確率和 recall 中分析，在 Facebook 資料集可以識別出超過 86%的友誼關係，而在 GitHub 的資料集中甚至高達 90%
3. 從上述實驗結果實驗分析(的一點和第二點)，可以顯示如果僅是披露部分資訊，仍有可能導致隱私資訊批露。

表 2. 實驗結果圖

Dataset	Facebook			Github		
	AUC%	Pr%	Re%	AUC%	Pr%	Re%
GCN	90.48	87.15	35.25	84.97	93.12	8.62
GAT	90.94	89.19	64.16	74.94	90.15	11.88
GraphSAGE	89.52	86.19	34.57	83.51	93.49	8.62

#3. Data-Mining of Social Media Users with Embedding Techniques and Neural Network

I. Target Problem

由於結構化數據特徵易於分析且可量化，因此在社交媒體用戶的數據挖掘中發揮著至關重要的作用，但也需要了解到自然語言數據在捕獲使用者生成內容的細微差別的重要性。通過在結構化數據特徵和自然語言數據分析間取得平衡，並充分利用社交媒體數據的潛力，全面瞭解用戶行為、情緒和趨勢。

此研究旨在深入研究使用者在社交媒體平臺上留下的自然語言文本數據，採用特徵嵌入技術和神經網路建模方法來研究變數之間潛在的複雜關係。

II. Dataset

- **Dataset Name**
  Bilibili
- **Dataset Link**
  https://www.kaggle.com/datasets/beats0/bilibili-use
- **Dataset characteristic**
  由 Kaggle 平臺上可用的 800,000 多個條目組成利用 pandas DataFrame 的 .sample() 方法，隨機提取了總共 10,000 條使用者記錄，從而組成一個小規模但具有足夠代表性的數據集，能夠確保神經網路模型進行準確預測。

III. Data Mining workflow and main techniques

1. Data Processing

資料集如表 3 顯示數據集中每個屬性的類型、名稱和相應的解釋。然而，結構化數據中的特徵存在某些共同點，像是它們定量和精確地捕捉樣本與社交媒體平臺互動的各個方面，例如受歡迎程度、創作熱情程度和對平臺使用的忠誠度。因此，這些結構化數據指標不僅可以作為模型構建的特徵，也可以作為模型預測的目標。在此研究中，作者將依次使用 「is_senior」 和 「level」 作為 target label。

2. 嵌入技術

雖然先前研究已廣泛研究基於文本的特徵，但使用者名稱通常被視為唯一識別碼，同時被歸類為索引類型的數據，受到的關注相對較少。本研究通過採用嵌入技術，創新性地將使用者名稱轉換為(10000, 768) 的向量，以預先訓練的大型語言模型 BERT 作為文字嵌入的基礎。鑒於觀察到使用者名稱包含來自中文、英文和日語等語言的混合字元，作者選擇了 “bert-base-multilingual-cased” 模型作為訓練單詞嵌入的基礎。

3. LSTM 網路 (圖 3)

首先，進行數據清理，使用 fillna('unknown')方法處理 “sign” 列中的缺失值。接著，建立了一個 size=10000 的 dictionary，利用 keras.preprocessing.text.Tokenizer 方法對文本進行標記並生成 dictionary index，為所有單詞分配索引。然後使用 X_train_padded = pad_sequences (max_sequence_length = 100) 方法填充序列。鑒於目標為預測二元結果，作者採用 binary_crossentropy 作為損失函數，並使用 adam 優化器。

表 3. 資料集名稱及解釋

Data Type	Feature	Explanation
Index	uid	Identifier, int 1,2...
Image	avatar	Profile Picture
Text(NLP)	name	User name
	sign	User signature, a short text shows their personality
Structured	level	User level, reflects user stickiness
	sex	Users stated gender, male/female/secret
	vip type	Current subscribe plan, int 0/1/2, means non-subscriber/monthly/yearly
	vip_status	Current Membership Status, bool, 1 for yes, 0 for no
	vip_role	Rank of subscriber,int, 0/3/1/7/15
	archive	Number of Submissions as a creator
	fans	Number of followers, reflects popularity
	friend	Number of mutual followers
	like_num	Number of recieved "like"
	is_senior	Senior member or not, bool, 1for yes, 0 for no

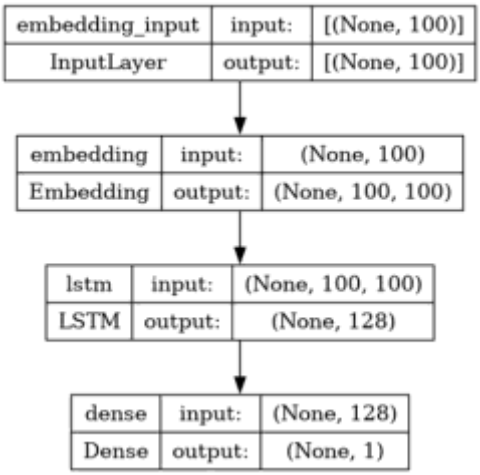

圖 3. LSTM 網路

IV. Results

利用使用者簽章特徵訓練 LSTM 模型的背景下，對兩個流行的學習率值（特別是 1e-3 和 1e-4）進行比較實驗，如圖 4 所示。可以看出兩種配置的模型訓練歷史記錄的視覺化分析中，結果表明，1e-4 的學習率是更合理的選擇，但差異非常微妙，平均為 2.4%。

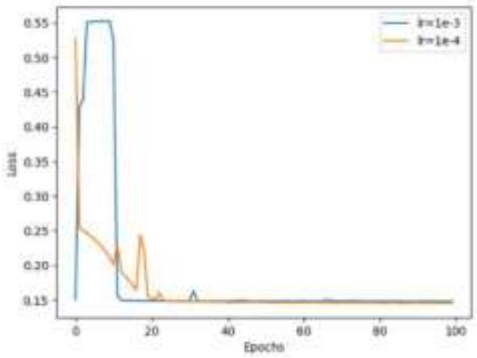

圖 4. 訓練結果圖

作者在表 4 中選擇性地列出了四個關鍵、有代表性的實驗結果。這些結果表明，當模型結構簡單時，增加隱藏神經元數量會導致預測精度大幅下降。此外，與僅依賴結構化資料(即密集特徵)的模型相比，將嵌入納入建模過程可產生顯著優越的結果，表現出超過 3 個百分點的改進。

表 4. 關鍵性實驗結果數據

	Exp1	Exp2	Exp3	Exp4
L.r.	1.00E-03	1.00E-03	1.00E-03	1.00E-03
epoch	100	100	100	500
hidden units	64	64	128	128
feature sets	Embedding +len_sign+ dense 11	Dense11	Dense11	Embedding +len_sign+ dense 11
loss	0.9329	0.963	45.47	0.9629

#4. Dual Graph Networks with Synthetic Oversampling for Imbalanced Rumor Detection on Social Media

I. Target Problem

謠言的偵測能夠有效預防人們被錯誤資訊誤導。但，在謠言偵測裡最大的挑戰在於謠言與真實言論的數量極度不平衡，謠言比起真實言論更不常見。

在過去研究中，透過資料重採樣來解決此問題，但問題在於只適用在特定的特徵才能有效預測。

此研究的目標是想在類別不平衡的情況下判斷言論的真實性，並將貼文的語意資訊和使用者資訊一起有效融入至模型中。

II. Dataset

- **Dataset Name**
  PHEME, RumourEval
- **Dataset Link**
  https://figshare.com/articles/dataset/PHEME_rumour_scheme_dataset_journalism_use_case/2068650?file=4988998
  https://github.com/autoreleasefool/rumoureval

III. Data Mining workflow and main techniques

這篇論文提出一個新的 SynDGN 架構，其過程如下:

1. Learning Tweet Representation

此步驟主要目標在於針對推文串(包含原文和使用者回覆)進行編碼。編碼方式為使用預訓練的 BERT，會將文章對應到相關的句子，以表示此文章狀態。並依據這些推文的相關性，建構出圖。

2. Learning User Representations.

此步驟為學習參與此推文的使用者，想要知道在這篇貼文中最有貢獻的使用者以此最為判斷此貼文是否為謠言的參考。這裡是使用 DTAC 去提取用戶特徵，並且將各用戶用圖表示。

3. Co-attention Tweet Representation and User Representation

為使用者資訊和貼文句子間建模，以了結其中的潛在相關性。使用 Co-attention 去連接 Tweet Graph 的 Tweet embedding 與 User Graph 的 User embedding，。

4. Synthetic Embedding Generation.

提出一個新的合成技術，用於生成少數類別，他的想法為隨機選取謠言的範例中的 embedding 進行插值並合成新資料。其中公式如下:

erumor is the BERT embedding of a randomly sampled rumor instance.

$$h_{syn} = (1-\delta)\cdot e_i + \delta\cdot(e_{rumor})$$

ei is the representation derived 從第 2 步驟

5. Final Prediction

使用 MLP 生成最後預測

$$\hat{y} = \text{softmax}(W_s \cdot h_{syn} + b_s)$$
$$\hat{y} = \text{softmax}(W_r \cdot e_i + b_r)$$

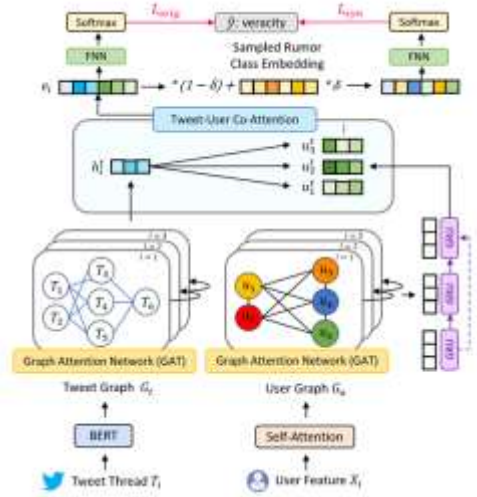

圖 5. The research of overall workflow

IV. Result

1. 從表 5 實驗結果顯示，SynDGN 比起其他模型有更好的分類效果。

2. 圖 6 顯示在類別不平衡下，SynDGN 也比其他模型有較好的效果，說明即使在少數類別樣本度多情況下，還是能 generalize 的不錯。

此篇研究的貢獻為提出一個過採樣的雙網路圖-SynDGN。SynDGN 的核心理念是利用使用者和推文這兩個 GNN，以更有效的做監督式學習。並且設計一個新的有效生成過採樣樣本，基於社交背景對少數類別做過採樣，這比起傳統的 SMOTE 直接作用在特徵空間不同，能夠確保生成的合成樣本不只是重複樣本，而是聯合使用者資訊和推文，產生出有意義且豐富的樣本。

表 5. 模型實驗結果比較

	RumorEval			PHEME		
	P	R	F1	P	R	F1
SVM	72.20	58.33	64.53	75.68	73.42	74.53
CNN	71.05	54.16	61.47	81.06	78.81	79.92
DeClarE	80.76	79.16	79.95	83.78	76.07	79.74
RvNN-BU	64.73	75.00	69.49	70.36	58.88	64.11
RvNN-TD	80.69	79.13	79.90	66.72	57.42	61.72
BaysienDL	71.53	73.48	72.49	77.17	67.85	72.21
AIFN	73.81	70.00	71.85	71.73	72.73	72.23
GEAR	80.00	81.25	80.62	81.11	78.82	79.95
BiGCN	84.23	84.01	84.12	86.87	86.75	86.81
EBGCN	86.48	86.57	86.52	87.11	87.34	87.22
SynDGN	**93.07**	**88.02**	**90.47**	**88.12**	**87.43**	**87.77**

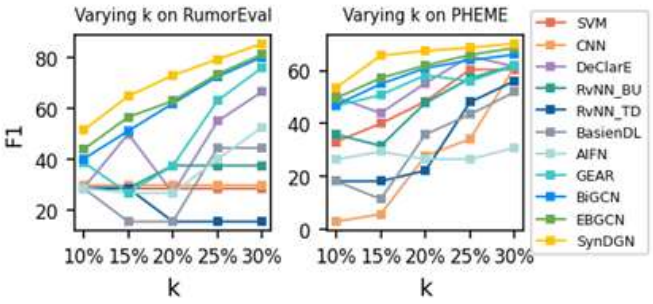

圖 6. 類別不平衡下的實驗結果

#5. Interactive Activities Initiation through Retrieving Hidden Social Information Networks

I. Target Problem

這篇論文探討的是 Interactive Activity Initiation Network (InterAIN)。問題聚焦在未知拓撲的社交網絡中，如何以最小的成本與資源，來啟動用戶間的互動活動，從而實現網絡內的全域覆蓋。不同於傳統的影響力最大化問題，InterAIN 專注於透過邊的互動來驅動用戶參與，而非單純的訊息擴散過程。由於社交平台的拓撲結構通常受到隱私、商業競爭或資料時效性的限制而無法得知

此篇研究提出了一種基於流量偵測器和頂點覆蓋求解器的互動式演算法，透過每輪從網絡中隨機採樣少量邊來逐步構建網絡結構，最終達到有效覆蓋整個隱藏社交圖的目標。

II. Dataset

- **Dataset Name**
  Facebook, Github, Gowalla, Youtube
- **Dataset Link**
  Stanford Large Network Dataset Collection

III. Data Mining Workflow and main techniques:

1. Data Process

Facebook 是一個社群媒體網站，收集 Facebook 用戶的「社交圈」或「朋友清單」。 Github 是一個電腦程式開發者的共享網站，其中節點是共享至少 10 個儲存庫的開發者，邊是他們之間相互的關注者關係。 Gowalla 是一個基於位置的社交網站，其資料集包含友誼網絡。 YouTube 是一個包含社群網路的影片分享網站，其中的資料集包含網站上成員的友誼。

表 6. 資料集資訊

Data set	Node num.	Edge num.	Average degree
Facebook	4,039	88,234	21.8
Github	37,700	289,003	7.7
Gowalla	196,951	950,327	4.8
Youtube	1,134,890	2,987,624	2.6

2. 空圖結構

由於對網絡拓撲處於未知狀態，因此構建過程從一個空的圖結構開始。

3. 邊緣探索

流量監測器透過模擬網絡流量，隨機取樣網絡中用戶之間的互動，將有頻繁互動的用戶對（u, v）推斷為潛在的邊 {u, v}，最後在每一輪中返回少量未被覆蓋的新邊，並將這些邊添加到當前的部分圖結構中。

4. 節點選擇

VC 求解器根據流量監測器返回的部分邊集，使用貪心策略選擇連接更多邊的節點，計算臨時的最小頂點覆蓋集。

5. 更新圖結構

VC 求解器覆蓋的節點和邊被添加到圖結構中。

6. 多輪迭代

流量監測器會略已選的邊並繼續隨機抽樣新的邊，發現更多未被覆蓋的關係。VC 求解器基於新的邊集更新覆蓋節點，進一步擴展圖結構。

7. 停止生成

當流量監測器無法找到更多未被覆蓋的新邊，或當 VC 求解器已覆蓋所有可能的邊時，流程停止。

8. 最終結果

構建出一個接近完整的社交網絡拓撲

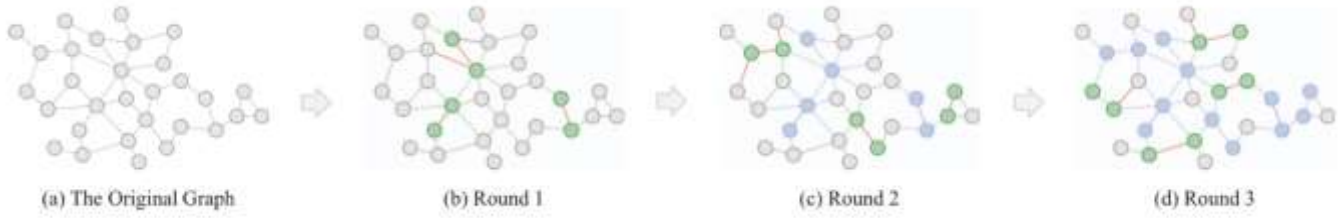

圖 7. The research of overall workflow

9. Main Techniques:

- 隨機取樣與探索:使用隨機流量監測器進行邊的探索，以應對未知網絡拓撲。
- 貪心演算法: VC 求解器運用貪心策略，選擇最小節點集來覆蓋更多邊，實現接近最佳的網絡構建效率。
- 互動式優化:流量監測器與 VC 求解器之間的多輪互動過程中，僅需掌握少部分的邊資訊，即可逐步覆蓋整個圖結構。

IV. Result

資料集中的評估結果如表 7 所示。特別是，沒有 vc-comp 演算法會循環而不是輸出最終的解，因此作者運行該演算法 50 輪，並在表 7 中顯示最終的覆蓋率。如表 7 所示，除了 vc-comp 演算法外，所有演算法都完成了產生覆蓋整個圖的頂點集，其中選擇的節點數量也類似。作者提出了 vc-node 演算法和 vc-edge 演算法都會產生 2-approximation 結果，在這裡也可以發現這些演算法產生的結果並不比原始頂點覆蓋的結果差，與理論分析相對應。從下表也可以發現，兩種演算法只需要一小部分邊（由偵測器隨機提供），同時傳回整個圖的頂點覆蓋。也就是說，在類似現實世界社群媒體的情況下，此演算法能夠隱藏大部分拓樸資訊。

表 7. 實驗評估結果

Data set	Algorithm	Node chosen	Edge requested
Facebook	vc	3,714	(All)
	vc-node	3,596	3,105
	vc-edge	3,594	3,347
	vc-comp (99.9%)	3,456	2,405
Github	vc	19,667	(All)
	vc-node	17,882	30,817
	vc-edge	17,540	23,582
	vc-comp (93.6%)	9,135	10,492
Gowalla	vc	104,761	(All)
	vc-node	97,821	178,750
	vc-edge	96,731	125,146
	vc-comp (73.5%)	24,441	34,308
Youtube	vc	339,992	(All)
	vc-node	323,358	854,893
	vc-edge	306,904	416,509
	vc-comp (68.9%)	39,021	48,402

#6. ANTI-Disinformation: An Adversarial Attack and Defense Network Towards Improved Robustness for Disinformation Detection on Social Media

I. Target Problem

雖然文本分類領域近年來取得了顯著進步，但現有方法往往忽視了虛假信息不斷發展的性質，其中犯罪者利用對有毒內容的擾動來逃避檢測或審查。為解決此問題，作者提出了一個新的框架，即 Adversarial Network Towards Improved robustness for Disinformation detection (ANTI-Disinformation)，它利用強化學習技術作為對抗性攻擊，以使用對語氣最少的影響進行最有效的修改。作者將文本編輯過程構建為代理的環境，並提出推送獎勵(push reward)和相似性獎勵(similarity reward)，以分別最大化擾動和語義守恆。此外，作者提出了一種防禦模型來恢復被攻擊的文本樣本，並希望通過結合對抗性訓練來展示創建更健壯的檢測系統的潛力。這種對抗性攻擊和防禦範式不僅可以促進對與虛假資訊傳播相關的不斷變化的行為的適應性和主動回應，還可以增強模型的穩健性和泛化能力。

II. Dataset

- **Dataset Name**
  Instagram, Vine, Twitter
- **Dataset characteristic**
  Instagram 和 Vine 數據集是針對網路欺淩檢測的部分，而 Twitter 數據集則是針對謠言檢測的部分。其中數據統計量如表 8 所示:

表 8. 資料集資訊

Datasets	Instagram	Vine	Twitter
# Posts	2,211	882	742
# Positive	676	283	372
# Negative	1,535	599	370
# Comments	159,277	70,385	216,805

III. Data Mining workflow and main techniques

作者提出的架構由兩個主要的部份組成，分別是對抗性攻擊模型以及對抗性防禦模型。

1. 對抗性攻擊模型 (圖 8)

在此模型中，作者利用強化學習來製作對抗樣本。強化學習已經取得了重大進步，尤其是在遊戲場景中。因此，作者採用深度 Q 學習(DQN)作為方法的基礎，並將攻擊過程制定為基於文本的遊戲。具體來說，每個註釋 ci 表示給定文章中基於文本的遊戲。作者的目標是在這些文本遊戲中戰略性地採取行動，以有效降低模型的準確性。通過反覆運算參與這些文本遊戲，可以使模型學習如何製作高質量的對抗範例來降低模型的性能。

2. 對抗型防禦模型 (圖 9)

對抗性的例子不僅用作攻擊模型的手段，還用作增強其穩健性的工具。鑒於拼寫錯誤的單詞和語法錯誤在社交媒體平臺上很常見，此類雜訊會顯著影響模型的準確性。通過防禦對抗性攻擊，作者可以提高模型的準確性和穩健性。因此，作者提出了一種新的防禦框架來實現這個概念來解決問題。

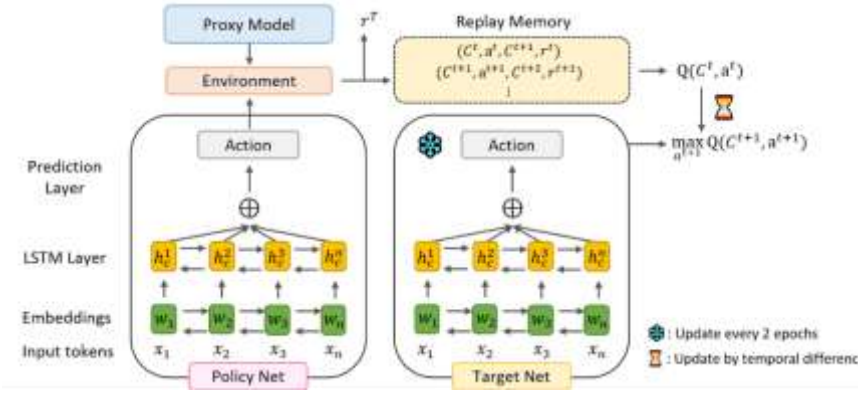

圖 8. 對抗性攻擊模型

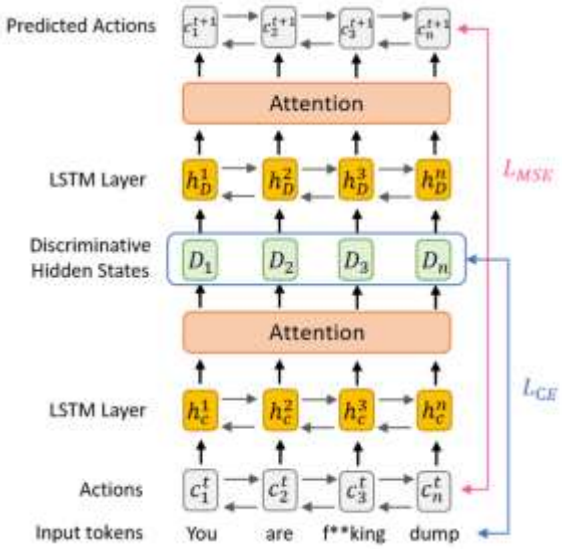

圖 9. 對抗型防禦模型

IV. Result

為了確保在不同設置下進行全面評估，作者將實驗分為兩部分：白盒攻擊(white box attacks)和黑盒攻擊(black box attacks)。在白盒攻擊場景中，可以完全訪問目標模型的架構和參數，能夠執行定製的攻擊。而黑盒攻擊場景中，對目標模型的瞭解則是很有限的。為了評估提出的方法的表現，作者使用了四種流行的語言模型，即 LSTM、BERT、RoBERTa 和 XLNet 作為代理模型。

白盒攻擊: 在表 9 中展示了白盒攻擊的結果，展示了將作者提出的攻擊和防禦方法與四種文本分類模型相結合的影響。這樣的結果表明，用作者提出的攻擊模型生成的文本可以有效地破壞眾所周知的文本分類模型的有用性。Instagram 數據中所有模型的準確性下降百分比均超過 50%，Amazon Vine 數據中超過 20%，Twitter 數據中超過 40%。

表 9. 白盒攻擊實驗結果

Dataset	Instagram				Vine				Twitter			
	Clean	Att.	∇	Def.	Clean	Att.	∇	Def.	Clean	Att.	∇	Def.
LSTM	82.7	41.0	-50.4%	83.7	77.8	48.4	-37.8%	79.8	87.0	38.0	-56.3%	87.5
BERT	83.7	14.0	-83.3%	84.3	77.8	60.8	-21.9%	72.6	88.1	28.1	-68.1%	89.3
RoBERTa	85.9	41.0	-52.3%	81.6	79.8	62.0	-22.3%	80.4	90.6	38.1	-42.1%	88.1
XLNet	77.8	22.7	-70.8%	78.3	79.3	61.8	-22.1%	78.3	89.3	41.8	53.2%	88.1

黑盒攻擊: 在表 10 中，作者展示了黑盒攻擊的結果。“Clean” 行表示模型在不執行任何攻擊的情況下的準確性分數。“Permutation” 行表示當句子中的單詞隨機排列時的性能。此外，作者通過隨機交換形容詞和名詞來生成對抗性示例（行 “Adj. Swap” 和 “Noun Swap”）。另外包括了 DeepWordBug，這是一個專門用於攻擊針對文本分類的深度學習模型。最後一行“RL Attack” 對應於作者提出的攻擊方法。可以看出作者提出的方法產生較低的準確率分數，證明了它在生成高品質的對抗性樣本作為被攻擊文本方面的有效性。

表 10. 黑盒攻擊實驗結果

Dataset	Model	Clean	Permutation	Adj. Swap	Noun Swap	DeepWordBug	RL Attack (Ours)
Instagram	BERT	83.7	78.9	82.1	83.2	75.1	**67.0**
	RoBERTa	85.9	50.1	51.1	50.8	50.2	**49.1**
	XLNet	77.8	49.2	51.8	49.6	48.5	**48.1**
Vine	BERT	77.8	67.5	75.7	74.7	48.2	**39.6**
	RoBERTa	79.8	64.9	66.8	65.7	65.2	**64.9**
	XLNet	79.3	64.9	65.8	66.9	**64.8**	**64.8**
Twitter	BERT	88.1	85.0	86.2	85.6	67.2	**46.0**
	RoBERTa	90.6	82.5	86.8	87.5	60.3	**55.0**
	XLNet	89.3	87.5	86.8	87.5	70.3	**54.0**

#7. A Machine Learning Based Social Network Data Mining System for Better Search Engine Algorithm

I. Target Problem

挖掘社交大數據並不是一件容易的事，需要適當的演算法才能獲取相關資訊，只有正確的資料探勘流程才能幫助開發高效率的大數據分析。在資料過濾過程中導致的三個主要問題是(1)虛假資訊過濾和(2)不適當的資料過濾以及(3)基於假定的資料回傳。主要的搜尋引擎不信任社交網路訊息，但社交網路平台上的信息比任何其他搜尋引擎上的信息都多。為了利用他們的數據，需要多種演算法來從社交網路中找到準確的資訊。

此研究設計了一個系統，該系統將獲取社交網路數據並根據資訊有效性進行分析。

II. Dataset

作者自 Instagram 與 Twitter 收集了 10,000 個用戶數據，以應對不斷變化的社會經濟條件。每個樣本都針對此事提供了自己的觀點。主要問題是 COVID-19 大流行期間對微型企業系統和大型企業系統的影響造成的損失數量。輸出統計如表 11 所示:

表 11. Covid-19 人們觀點統計圖

User Opinion	Percentage of people agreed opinions
People who think small business causes much damage on pandamic	73% people of total Analysis
People who thinks both small and large business causes damage	8% people of total analysis
People who think only large business causes much damage on pandemic	19% people of total analysis

III. Data Mining workflow and main techniques

1. 假消息辨識 (圖 10)

Fake Detector 解決了兩個主要組件：代表特徵學習和可信度標籤推斷，它們共同構成了深度擴散網路模型 Fake Detector。假訊息會帶來更多負面意見。所以，透過分析，評論往往可以發現假新聞。另一方面，作者的數據共享可能性也回傳了假新聞機率。

2. 資料過濾 (圖 11)

年齡較小的人經驗較少，因此他們分享的資訊接受度較低。年長的人會因為經歷而更容易接受。作者使用歐幾里得測量來產生數據處理值。

3. 基於使用者選擇的資料有效性 (圖 12)

有時使用者定義的邏輯不起作用，在這種情況下，作者在這裡使用基於使用者意見的假設邏輯傳回的線性迴歸分析。作者收集了 1000 個使用者的資料來解決社會經濟條件變化的問題。每個樣本都對問題給出了相應的看法。

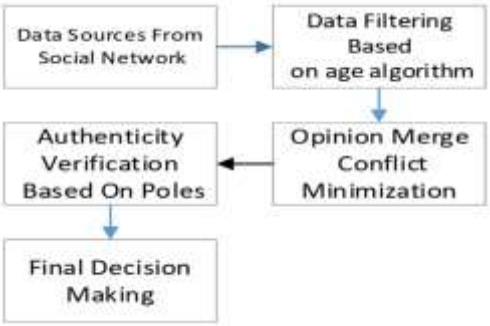

圖 10. 假消息辨識流程圖

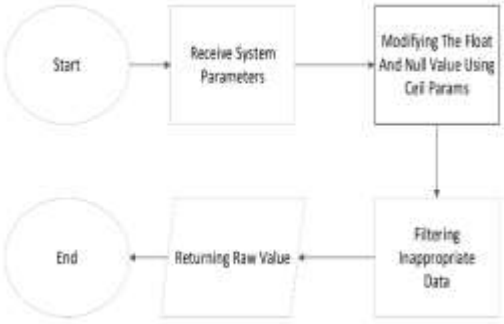

圖 11. 資料過濾流程圖

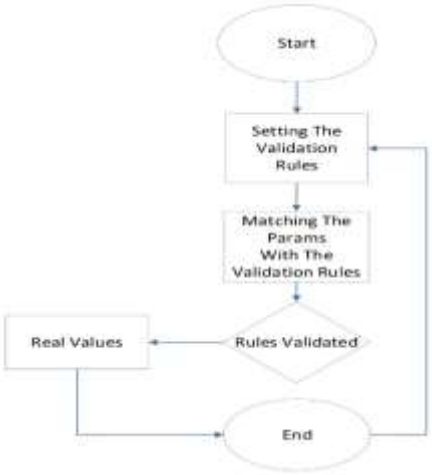

圖 12. 基於使用者選擇的資料有效性

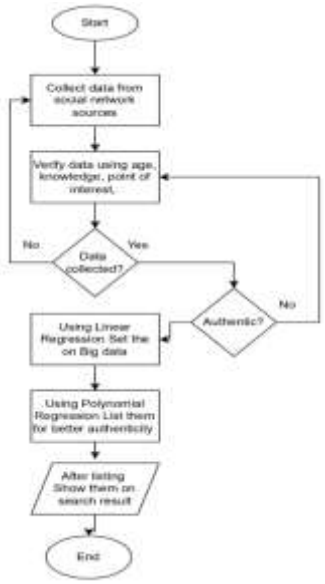

圖 13. The overall workflow of research

IV. Results

為了對所提出的方法進行初步評估，作者分析了一些社交媒體事實並將其與隨機搜尋引擎結果進行比較。根據該分析發現了一些準確性。整體表現如表 12 所示:

表 12. 社交媒體事實與隨機搜尋引擎結果比較圖

Topic	Social Media Result	Google Result	Bing Result	Accuracy On our Search
Covid-19 medicine efficiency	73% of people think they are effective	50% effective based on research results	52% effective based on search result	73.9%
Global warming facts	Most major analyses on social media focused on greenhouse effects and industrialization	Among major issues, google focused on greenhouse effects and volcanic activity	Among major issues, Bing focused on greenhouse effects	78%
Financial investment in a global stock exchange	56% of people think that investing in a global stock is worthy	google highly recommends investment	Bing returns to authentic results	50%

在這裡可以看到結果和準確性的一些變化。因此，根據這些結果，可以看出社群媒體對於搜尋議題的重要性。

#8. Classifying Severe Weather Events by Utilizing Social Sensor Data and Social Network Analysis

I. Target Problem

不確定的異常天氣災害，會對人類社會帶來嚴重影響，因此，若能有效預測異常天氣發生，將能降低災情的發生和損失。隨著機器學習的蓬勃發展，各領域都有重大的突破，但在預測異常天氣仍具有挑戰。主要原因在於缺乏高品質的天氣資料，其中包含氣象站感測器故障引起的資料缺失，以及不同天氣類別間資料的不平衡。

在過去的研究中，有些方法是利用模擬和特徵重建的方式來進行資料過採樣 (oversampling)，然而，此種方法往往因為存在內在假設和謬誤，所以產生的數據無法完全模擬真實世界，而導致演算法性能被高估。

此篇研究的目標是通過將氣象數據與地理標記的社交媒體結合，利用多模態 (Mutlimodel) 資料來提升異常天氣的預測。

II. Dataset

這篇論文的資料集並不只是單純使用某個開源資料集，而是作者們自己從各方收集並聚合而成，包含天氣資料集 (從 ASOS 收集 9 個阿拉斯加州的氣象站資料)、地域性的社群資料集 (專注在阿拉斯加州的 Anchorage、Hoonah 這兩個城市)。另外，這兩資料集並未在論文中提供連結。

III. Data Mining workflow and main techniques

1. Data Collection, and Preprocessing

因為是針對阿拉斯加的天氣進行預測，所以從 ASOS 收集 9 個阿拉斯加州的氣象站資料。另外，提取地域性相關的推文，推文特點包含需是天氣相關推文，且必須是在此地區發布的推文，另外也移除所有重複的推文。

2. Feature Extraction

- Weather Data
  對各氣象站計算在特定期間的各特徵的平均 (Mean of Variables)、變異數 (Variance of Variables)，且針對平均和變異數的範圍 (Variable Range Mean and Variance)做統計。在填補缺失值是針對此統計分析作填補，並對其做標準化。
- Social Sensors Data
  推文會先經過清洗、聚合、標記，並在使用 BERT (Bidirectional Encoder Representations from Transformers)做特徵提取。

3. Data Integration

將天氣數據和由上步驟從 Twitter 提取的特徵做整合，資料合的過程是透過時間戳將兩資料集對齊，以確保時間一致性。

4. Training Dataset

針對天氣進行標記，各類型天氣災害對應至一個數字。並將所有資料集的 70%作為訓練，其餘做為測試。

5. Baseline Traditional Methods vs Multi Layer Perceptron

Baseline 是挑選 Random Forest 和 Logistic Regression 模型。並與多層感知器 (Multi-Layer Perceptrons, MLPs) 進行比較。MLP 能考慮稀有事件模式中的非線性特徵，因此藉此評估 MLP 在捕捉稀有事件的複雜性與非線性關係方面的優勢。

圖 14. The overall workflow of research

IV. Result

在此研究中創建了四個不同的資料集，藉以比較不同資料集的準確率，來證明參考社群網路的推文對於天氣預測有良好的效果。結果表 13 所示:

表 13. 實驗結果圖

Result					
Dataset	*weather*	*social*	*proximity*	*F1 range in 5 folds*	*SD*
A	✓	✓	✓	**[0.78, 0.83]**	**0.01**
B	X	✓	✓	[0.72, 0.78]	0.06
C	✓	✓	X	[0.71, 0.77]	0.02
D	✓	X	✓	[0.26, 0.33]	0.03

Dataset A: 結合天氣和考慮地域性的推文

Dataset B: 僅使用地域性貼文，不考慮天氣

Dataset C: 結合天氣和不考慮地域性貼文

Dataset D: 僅使用鄰近氣象站隨機抽樣的天氣數據

從結果顯示，利用 Dataset A，也就是包含社交媒體的資訊，可以有更高的 Accuracy、F1-score。如果只依賴天氣資料或社群資料，效果都不如結合兩資料集效果來的好。

此研究證實了利用社群媒體的貼文對於天氣預測是有幫助的，並提供資料特徵提取的想法和流程，為天氣預測提供新的觀點。

#9. Social Media Sensors for Weather-Caused Outage Prediction Based on Spatio–Temporal Multiplex Network Representation

I. Target Problem

惡劣天氣變得更加普及且嚴重，對基礎設施和生活品質帶來重大影響。社交網路的使用者可以觀察以及分享有關天氣相關造成的破壞事件。在過去有不少研究是將目光放在利用社群媒體預測自然災難的發生，但很少有研究是將社群媒體與停電數據做結合。

此研究是想透過學習時空多重網路 (spatio-temporal multiplex network)來提高與天氣相關的停電預測準確性。

II. Dataset

這篇論文的資料集並不只是單純使用某個開源資料集，而是作者們自己從各方收集並聚合而成，包含停電相關資料、電線路 (BPA 傳輸線)、天氣相關資料 、閃電、植被、社群媒體資料 (twitter、Reddit)。另外，這兩資料集並未在論文中提供連結。

III. Data Mining workflow and main techniques

1. Data Collection

由 (II.Dataset)中我們可以知道此研究相關收集的資料，在論文中有針對各項說明詳細的資料收集方法，在此報告中我僅將會針對社交媒體資料的收集做說明。在 twitter 和 Reddit 這兩社群平台，都是利用關鍵字收集與停電相關的貼文，並收集發此推輪的使用者個人訊息，依照區域性分類，並且最後只留下選定的地區，其相關推文。

2. Multiple Social-Power Network Creation

- 在此步驟是建立一個多層網路，總共有 6 層，各層名稱與解釋如下表格。

Layer	Explain
BPA transmission line layer	若兩地區共享同條 BPA 輸電線，則為此兩頂點建立邊，且權重表示共享的輸電線數量。
Power outage layer	若兩地區在同一天發生停電，則為此兩頂點建立邊。
Weather layer	若兩地區在同一天既發生停電又報告惡劣天氣，則此兩頂點建立邊。
Lightning layer	若兩地區在同一天發生閃電，則此兩頂點建立邊，且權重表示當天的閃電次數總數。
Vegetation layer	根據頂點之間的植被特徵相似性（以歐幾里得距離衡量）建立邊。若距離小於閾值(<0.5)則此兩頂點建立邊，否則不連接。
Social media layer	若兩地區在停電期間皆有社交活動（如 Twitter 或 Reddit 的相關貼文），則此兩頂點建立邊，且**權重示該期間的社交活動數量**。

- 多層網路的連接

每層網路的頂點集相同，跨層將各層相同的頂點座連接，而內層頂點的連接方式依照上表的特性決定兩頂點間是否有邊。使用此多層連接的目的在於將各層不同特性考慮進步分析此差異。比如:天氣、閃電、社交平台的討論等，這些特徵如何導致停電，增強其關聯性。

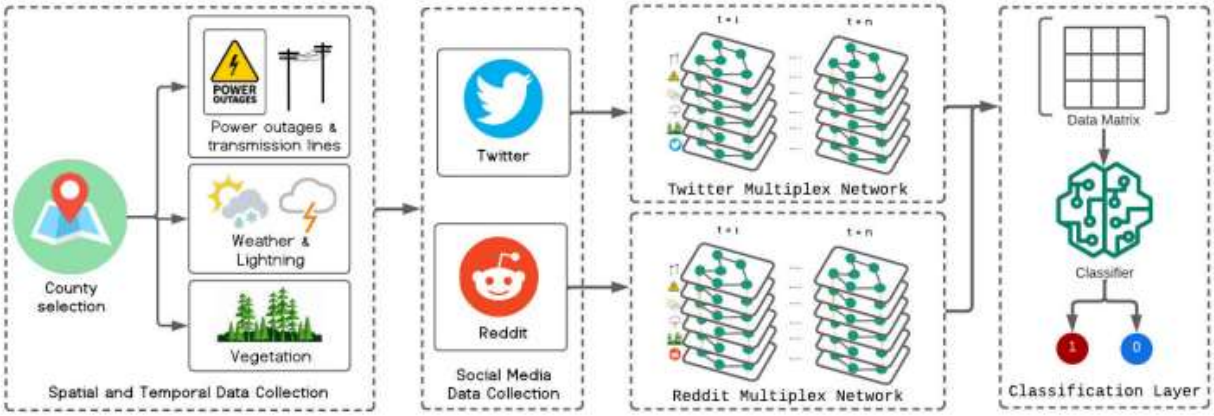

圖 15. 多層網路的連接的整個流程圖

3. Model Setup and Training

此研究為評估使用第 2 步驟的多層網路對於停電的四冊是否能提高準確率，故將某個地區在某時間內若有停電則設為 1，否則為 0。此外，資料集是相當不平衡，採用 SMOTE 解決資料不平衡。模型採監督式學習，模型則分別測試了 logistic regression、 neural network support vector machine、 random forest、xgboost、decision tree。將資料的 75%為 training data，其餘為 testing data。並使用 5-fold validation 驗證模型的穩定度。

IV. Result

在此研究中創建了三個不同的資料集，藉以比較不同資料集的準確率，來證明參考社群網路的推文和互動對於停電預測有良好的效果

Dataset A: weather features

Dataset B: weather and lightning features

Dataset C: weather, lightning, and vegetation features. Social sensors are Twitter (T) and Reddit(R)

結果如表 14:

表 14. 實驗結果圖

	Baselines						The proposed approach					
Features	*SetA*		*SetB*		*SetC*		*SetC + R*		*SetC + T*		*SetC + R + T*	
Classification layer	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
LR	0.73	0.81	0.74	0.82	0.75	0.82	0.79	0.85	0.81	0.86	**0.82**	**0.87**
DT	0.83	0.87	0.84	0.88	0.85	0.88	0.87	0.90	0.88	0.90	**0.90**	**0.91**
RF	0.84	0.88	0.85	0.88	0.86	0.89	0.88	0.90	0.89	0.91	**0.91**	**0.92**
NN	0.84	0.88	0.85	0.88	0.86	0.89	0.89	0.91	0.90	0.91	**0.91**	**0.92**
SVM	0.85	0.88	0.86	0.89	0.87	0.89	0.90	0.91	0.91	0.92	**0.92**	**0.93**
XGB	0.86	0.88	0.87	0.89	0.88	0.90	0.91	0.92	0.92	0.93	**0.93**	**0.94**

從上圖可以得知不管在選擇哪個模型，在有考慮社群網路作為感測的資料集，預測的準確度和 F1-score 都是最高，這表明社群網路的互動 (例如:貼文或是討論度)對於預測停電都是有幫助的。

此篇研究的貢獻包含:

1. 證實導致停電的惡劣天氣與從社交平台(twitter、Reddit)提取使用者對於資訊的傳播 (如討論或是發貼文)是有正相關性。
2. 提出新的架構方式-多層網路的連接 (spatio-temporal multiplex network)可以比僅依賴天氣特徵此種簡單網路有更好的學習效果，並設計出能增強模型的網路結構。

#10. Unsupervised Post-Time Fake Social Message Detection with Recommendation-aware Representation Learning

I. Target Problem

社群媒體 (Social Media) 的蓬勃發展，讓人們可以更簡單且隨時隨地分享訊息並與他人互動，但這也造成假訊息 (fake messages) 的氾濫，因此如何偵測假訊息尤為重要。

過去的研究多採用監督式學習 (supervised learning) 的方法，透過分析使用者評論 (user comments)、社交網路 (social networks)、使用者屬性 (user attributes)、使用者與訊息互動 (user-message interaction) 來進行偵測。然而，在現有的架構中面臨兩大挑戰:

1. **依賴使用者回饋(user feedback to the source message):** 在模型訓練時假設使用者回饋是可取得的(例如: 訊息評論、轉發者等)。但如果在原訊息 (source message) 剛發布時，使用者的回饋還沒產生，因此在那時間點，便無法立即預測原訊息的真實性。
2. **標記訊息的難處:** 在監督式學習 (supervised learning)，需先針對訊息去做標記。然而，訊息標記的品質不一定，且也相當耗時和耗費人力。

此篇研究旨針對新發布訊息，使用非監督學習(unsupervised learning)實現**在沒有真實性標記和沒有使用者回饋的情況下**，及時辨識假訊息。

II. Dataset

- **Dataset Name**
  Twitter-15, Twitter-16
- **Dataset Link**
  https://github.com/OwenLeng/rumor-detection-include-twitter15-twitter16data-
- **Dataset characteristics**
  這兩資料集包含一個訊息集及相對應的轉發(retweet)用戶列表。在此研究中，他們只選擇那些標籤為 true 和 fake 的訊息作為 ground truth。另外，他們透過使用者 ID 從 Twitter API 去獲取使用者屬性的相關資訊。

III. Data Mining workflow and main techniques

此研究提出一個新架構 **Rec**ommendation-aware **M**essage **R**epresentation (RecMR)，核心想法為在訊息發布時利用 RecMR 來找出對此訊息感興趣的潛在用戶，並模擬可能的互動方式，而不需藉由真實的互動方式，來達到偵測假訊息。

使用的特徵包含使用者自我描述的字數 (number of words in a user's self-description)、使用者名稱數量 (number of words of user's screen name)、追蹤此使用者的數量 (number of followers)、此使用者追蹤的數量 (number of following)、此使用者發布限時動態的數量 (number of creating story)、此使用者是否開啟定位功能(whether user allows the geo-spatial positioning)。

訓練時只使用真實訊息作為輸入，而在模型評估時包含真實和假資料。

此系統可以分成 4 大部分:

第 1 部分，將原訊息 (Source Code) 透過 GNU 做嵌入表示 (embedding)。

第 2 部分，將第 1 部分的結果與用戶相關資訊一起輸入到 User-Recommend Module，並產生 K 名用戶推薦的列表。

第 3 部分，利用第 2 部分產生的用戶特徵，產生傳播序列 (Propagation Representation)和用戶間互動圖 (Graph Representation)。

第 4 部分，將原訊息的嵌入表示 (Source Code Embedding)、推薦系統用戶的傳播 (Recommendation Propagation)、推薦系統用戶間互動圖(Recommendation Graph)做串接輸入到 AutoEncoder。AutoEncoder 為 unsupervised learning，會將重建出來與真實輸入的訊息計算誤差，若是誤差較大表是假訊息的機率較高。

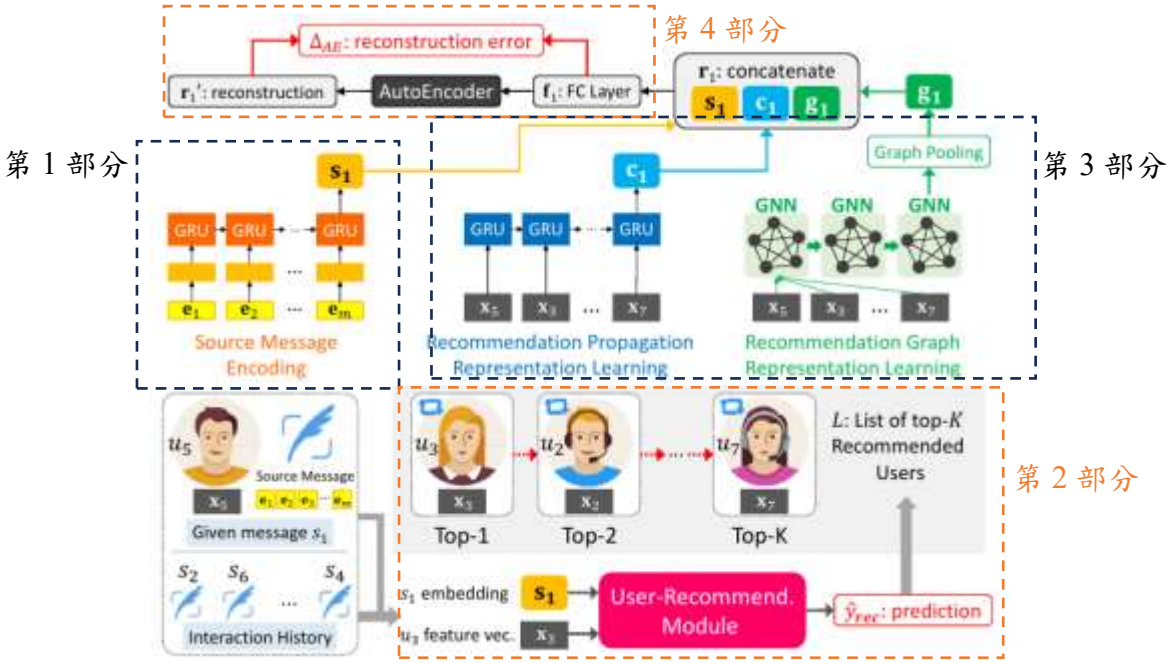

圖 16. The overview of RecMR framework

IV. Result

1. 此實驗的 RecMR 的表現比 UFNDA 來的好，證明 RecMR 成功模擬用戶訊息互動，並且能有效辨別假訊息。
2. 使用推薦系統並模擬用戶互動比真實用戶互動表現更好，推測可能是因為使用此系統能根據訊息對相關用戶的偏好，能收集到更更多的真實性證據。
3. 移除推薦系統用戶間互動圖(Recommendation Graph)表現會略變差，說明用戶的互動討論的價值。

此研究最大的貢獻在於探勘用戶資訊並建構出推薦用戶系統，以模擬用戶互動，因此不需要有真實用戶的情況下，提升非監督式學習。

表 15. Experimental results on Twitter15 data

		Precision		Recall		F1	
		@30%	@60%	@30%	@60%	@30%	@60%
	UFNDA (no users)	0.7249±0.0400	0.6895±0.0451	0.3575±0.0197	0.4534±0.0296	0.4788±0.0264	0.5471±0.0358
K=10	RecMR/GT	0.7206±0.0410	0.6992±0.0471	0.3554±0.0202	0.4598±0.0310	0.4760±0.0271	0.5548±0.0374
	RecMR/RD	0.7220±0.0414	0.7008±0.0435	0.3561±0.0204	0.4608±0.0286	0.4769±0.0273	0.5560±0.0345
	RecMR-G	0.7393±0.0456	0.7127±0.0565	0.3646±0.0225	0.4686±0.0371	0.4883±0.0301	0.5655±0.0448
	RecMR	**0.7483±0.0376**	**0.7202±0.0535**	**0.3690±0.0186**	**0.4836±0.0352**	**0.4943±0.0249**	**0.5786±0.0424**
K=30	RecMR/GT	0.7348±0.0263	0.7039±0.0393	0.3624±0.0129	0.4628±0.0259	0.4854±0.0173	0.5585±0.0312
	RecMR/RD	0.7272±0.0245	0.7006±0.0418	0.3506±0.0121	0.4607±0.0275	0.4704±0.0162	0.5559±0.0331
	RecMR-G	0.7351±0.0329	0.7142±0.0429	0.3625±0.0162	0.4696±0.0282	0.4856±0.0217	0.5667±0.0340
	RecMR	**0.7454±0.0385**	**0.7211±0.0404**	**0.3761±0.0170**	**0.4742±0.0332**	**0.5000±0.0228**	**0.5721±0.0300**
K=50	RecMR/GT	0.7257±0.0363	0.6957±0.0427	0.3579±0.0179	0.4574±0.0280	0.4794±0.0240	0.5520±0.0338
	RecMR/RD	0.7213±0.0377	0.6945±0.0435	0.3557±0.0186	0.4567±0.0286	0.4765±0.0249	0.5510±0.0345
	RecMR-G	0.7346±0.0420	0.7103±0.0499	0.3623±0.0207	0.4670±0.0328	0.4853±0.0277	0.5635±0.0396
	RecMR	**0.7406±0.0290**	**0.7183±0.0437**	**0.3717±0.0143**	**0.4723±0.0311**	**0.4945±0.0192**	**0.5699±0.0375**

#11. Predicting and Analyzing Privacy Settings and Categories for Posts on Social Media

I. Target Problem

由於用戶經常沒有意識到隱私管理，容易將自己的貼文暴露在隱私洩露的情況下，因此，如何有效預測並推薦社群媒體貼文的隱私設定與隱私類別，從而提升隱私管理，避免隱私洩露，便成為這篇論文研究的目標。

II. Dataset

作者使用 A personal privacy preserving framework: I let you know who can see what 的資料集，其中總共有 11,370 則貼文。透過將種子關鍵字清單提供給 Twitter 搜尋服務，針對預定義分類法中的每個隱私類別收集了貼文。每個貼文都與 32 個隱私類別相關聯，如「關係狀態」、「宗教」、「健康狀況」等，這些類別透過 Amazon Mechanical Turk (AMT) 進行標記。此外，每個貼文還與四種隱私設定的機率分佈相關聯：家庭、親密、休閒和戶外。在參考文獻中，根據人類日常行為將社群媒體貼文分為八個面向:「地點」、「醫療」、「關係」、「中立狀態」、「情緒」、「活動」、「個人屬性」和「人生里程碑」 。作者就參考文獻將 32 個隱私類別劃分為這 8 個面向。也就是說，作者對隱私類別有兩個粒度級別，並將討論這兩個級別的預測效能。下圖展示了 32 個隱私類別和 8 個面向的分佈和關係。

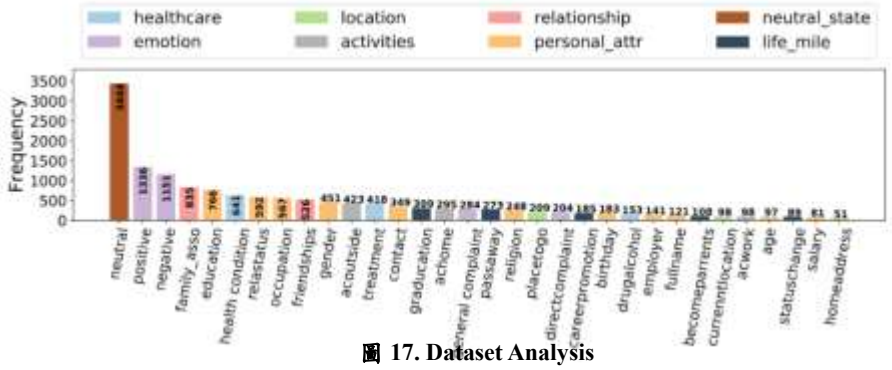

圖 17. Dataset Analysis

III. Data Mining Workflow and Main Techniques

主要步驟:貼文特徵提取 -> 透過圖嵌入進行貼文特徵學習 -> 提出多任務學習模型

1. 特徵提取:

- Linguistic Inquiry and Word Count: 透過心理語言學字典分析套件 LIWC，提取 70 維特徵向量，也就是 70 個使用者個性類別。
- 隱私詞典: 基於語言的隱私關鍵字區分 隱私相關與非隱私相關詞彙。
- 情緒偵測: 使用史丹佛 NLP 分類器，提取 5 種情緒極性特徵，分別是非常消極、消極、中性、積極和非常積極。
- 句子嵌入: 使用 Sentence2vec，提取貼文的語意嵌入向量。
- 元數據: 從貼文元資料提取標籤、俚語、表情符號、圖像等 互動特徵。

2. 特徵學習

使用 FeatWalk 方法在特徵矩陣 X 上進行隨機遊走，完成從貼文 i 到貼文 j 的步行。透過多次重複此過程，為每個貼文 i 生成一個貼文序列 $Q(i)$，作為在特徵空間中的隨機遊走軌跡。接著將隨機遊走路徑視為自然語言處理中的句子，並應用 skip-gram 架構來學習貼文之間的隱藏關係與相似性。在

這個架構中，每個貼文(在特徵矩陣 X 中)被視為一個單詞，而隨機遊走生成的貼文序列則被視為句子。最終獲得每個貼文 i 的學習特徵表示 $FL(i)$。最後，將提取的特徵向量 $FE(i)$ 與學習的特徵向量 $FL(i)$ 連接，作為最終特徵，並用於預測貼文的隱私設定。

3. 多任務學習

將串連的兩個特徵向量作為模型的輸入，由於隱私類別和隱私設定具有潛在的相關性，所以建立一個 128 維的共享隱藏層，並通過此層提取通用特徵。接著，在隱私類別與隱私設定兩個分支分別使用 sigmoid 與 softmax 激活函數進行預測，最後，利用 $L = \lambda \cdot Lc + (1-\lambda) \cdot Ls$ 損失函數來評估模型準確性。其中，Lc 為二元交叉熵損失函數，Ls 為分類交叉熵損失函數，$\lambda \in [0,1]$ 是一個權重參數，用來平衡兩個任務的損失。

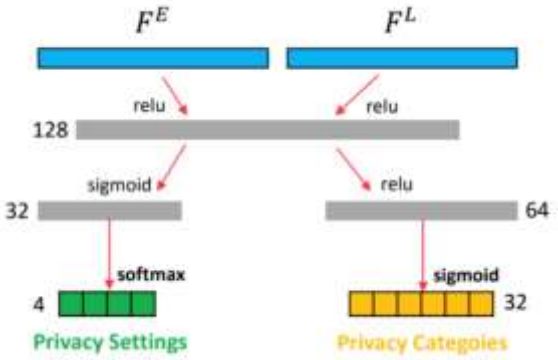

圖 18. 多任務學習

IV. Results

預測結果如圖 19 和圖 20 所示，分別為 8 個面向和 32 個隱私類別的結果。在八個隱私方面中，可以明顯觀察到「關係」、「個人屬性」、「中立狀態」、「情感」和「醫療保健」的預測效果較好。可能的原因也許是更多的貼文屬於這些方面，並且可以提取更多可區分的文本證據(例如相關單字)作為預測的特徵。「人生里程碑」、「活動」和「位置」方面可能具有屬於其他方面的與隱私相關的線索，因此不太可預測。而在 32 個隱私類別的細粒度結果的部分，可以發現與「關係」相關的類別(即友誼和家庭關聯)具有更高的預測表現，原因同樣可能是因為他們有更多的貼文和更多的證據性文字。此外，也可以發現到幾乎一半類別的 F1-score 接近 0，從圖 17 數據集的分佈圖可以看出，這應該是由於該類別的訓練貼文數量過少所導致。

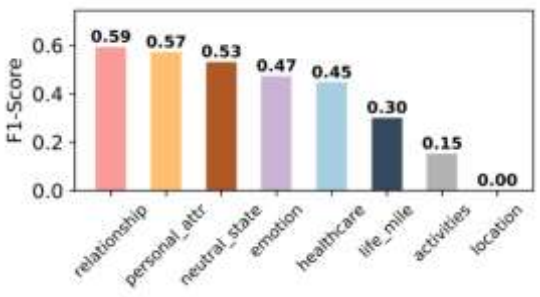

圖 19. 八面向實驗結果圖

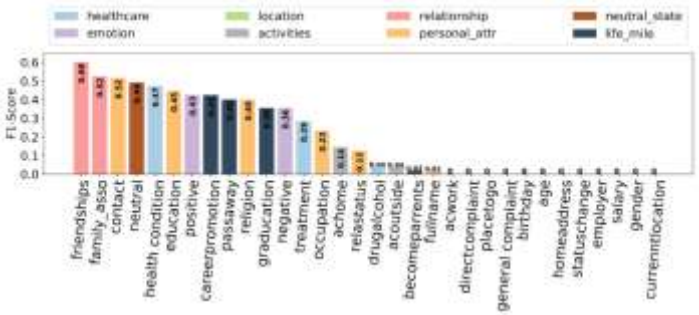

圖 20. 三十二個隱私類別實驗結果圖

12. Social Network Analysis of Popular YouTube Videos via Vertical Quantitative Mining

I. Target Problem

在此論文中，作者提出了一個演算法以挖掘有趣的 quantitative frequent patterns。作者的 qEclat（或 Q-Eclat）演算法擴展了常見的 Eclat 演算法，以便能夠垂直挖掘 quantitative patterns。與現有的 MQA-M 演算法(專為水平 quantitative frequent patterns 挖掘而構建)相比，作者的評估結果表明 qEclat 挖掘 quantitative frequent patterns 的速度更快。

II. Dataset

作者使用四個不同的定量交易數據集評估演算法的性能：

- 兩個合成數據集: 假設有 n 筆交易和 I 個不同的項目。每個項目都有在特定交易中發生的機率 prob，其中 0≤prob≤1。如果該項目出現在交易中，則該項目的出現次數將遵循 Poisson(λ)分佈加 1。作者設置 n = 1000、I = 50 和 λ = 1。這兩個定量交易資料庫的 prob 值分別為 0.2 和 0.8。
- 來自 UCI ML 儲存庫的兩個真實數據集: 作者修改了 chess 和 mushroom 數據集，使其成為定量交易資料庫。每當一個項目在交易中出現時，它不是只出現一次，而是它的出現次數遵循 Poisson(λ)分佈加 1.
- UCI ML dataset link: http://archive.ics.uci.edu/ml

III. Data Mining workflow and main techniques

1. Q-Eclat 演算法

首先，如果定量交易資料庫是水平格式，作者將它轉換為垂直格式。下一步是計算 C1 裡面的所有候選 1-itemexpsets，其中定義 Ck 為候選 k-itemexpsets 的集合。之後，作者計算與每個候選 1-itemexpset 關聯的 tidset，tidsets 可以根據定量交易資料庫的垂直表示計算出來。然後作者通過計算其相應 tidset 中的元素來計算每個候選 1-itemexpset 的 support。最後，作者基於兩個新的 pruning rulse 從 L1 移除一些 itemexpset，其中定義 Lk 為 frequent k-itemexpsets 的集合。

接著，將 k 設定為 2 並開始執行 main 迴圈。main 迴圈的第一步是使用 Lk−1 生成 Ck。如果有兩個頻繁的(k−1)-itemexpsets 在 Lk−1 之中且他們的第一個(k−2)-itemexp 是相同的以及它們的最後一個 itemexp 參考不同的項目，那就添加一個候選 k-itemexpset 到 Ck，它由第一個（k−2）-itemexps 和兩個 itemexpset 的最後一個 itemexp 組成。之後，會從 Ck 中刪除每個有包含一(k−1)-itemexpset 不屬於 Lk−1 的候選 k-itemexpset 。下一步是創建對應於 Ck 中每個候選 k-itemexpset 的 tidsets 。這可以使用 tidsets 的遞迴定義來完成:

$$tidset(X) = tidset(W \cup \{y\}) \cap tidset(W \cup \{z\})$$

在計算了 tidsets 之後，作者計算了 Ck 中每個候選 k-itemexpset 的 support。任何候選 k-itemexpset 其 support ≥ minsup 都會被添加到 Lk。使用兩個 pruning rules，作者從 Lk 移除不感興趣的 itemexpsets。到此便結束 main 迴圈開始遞迴，直到 Lk 變為空。

2. pruning rulse:

假設 X 包含一個格式為(z≤r)itemexp，如果 Y 具有跟 X 相同的 support，除了(z≤r)被替換為(z≤r+s)，那麼 Y 可以從 Lk 移除。

假設 X 包含一個格式為(z≥r)itemexp，如果 Y 具有跟 X 相同的 support，除了(z≥r)被替換為(z≥r-s)，那麼 Y 可以從 Lk 移除。

其中，X 和 Y 是 Lk 中的 itemexpsets，z 是一個項目，而 r 和 s 是正整數。

IV. Results

圖 21 顯示了每個真實定量交易數據集中各種 minsup 值的兩種演算法的運行時間(為了節省空間，省略了合成數據集的結果，但結果相似)。在所有情況下，作者的 Q-Eclat 傳回相同的 itemsexpsets 集合的運行時間都比現有的 MQA-M 演算法短。

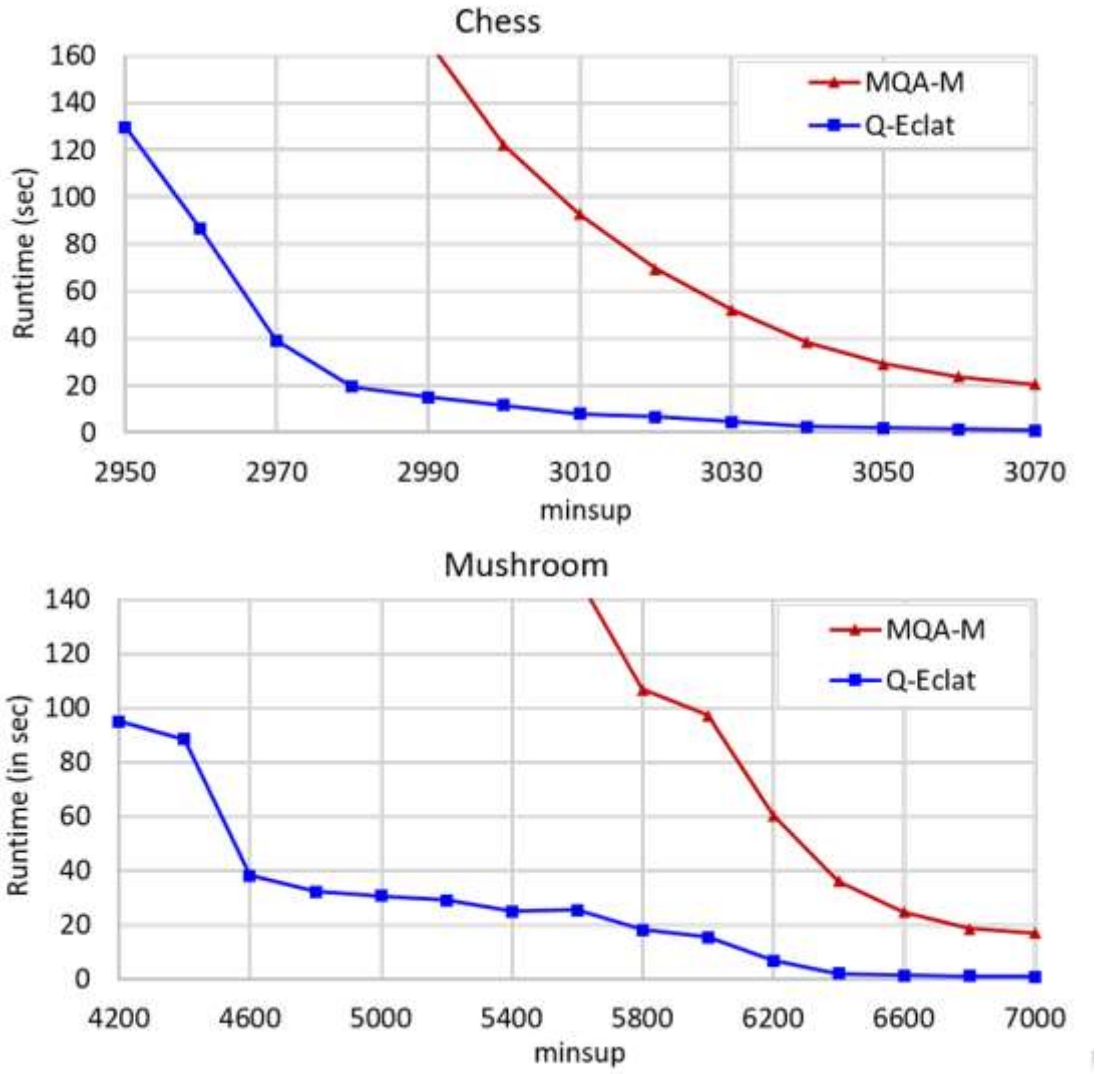

圖 21.實驗結果圖

#13. Finding Potential Propagators and Customers in Location-Based Social Networks: An Embedding-Based Approach

I. Target Problem

在社群媒體上打卡並分享位置 (Points-of-Interests)以成為生活日常，可以利用此對使用者關係做社交連接，形成位置的社交網路 (Location-based Social Networks)。

在 Location-based Social Networks 中最重要的議題為，找到有效的資訊傳播者，以能最大化顧客數量。在過去研究是利用找到影響力大的使用者，但這些都是依賴過去資料，所以不確定對未來是否一樣有效。

此研究主要是想找到 Targeted Propagator Discovery (TPD) 和 Targeted Customer Discovery (TCD)，以找到潛在的目標傳播者和客戶。

II. Dataset

- **Dataset Name**
  Instagram
- **Dataset Link**
  https://developers.facebook.com/docs/instagram-platform
  資料集可以透過 Instagram API 獲取

III. Data Mining workflow and main techniques

因為要找到 Targeted Propagator Discovery (TPD) 和 Targeted Customer Discovery (TCD)。

1. Random Walk Mechanism

利用 skip-gram 模型來學習使用者和 POI(Points-of-Interests)的特徵表示。計算使用者間的關聯性，如果使用者間有更多共同存取的 POI，代表其關係度越高，或是兩個使用者有較多的共同朋友，則代表其關係較為相向，並生成相關圖與賦予使用者間，邊的權重。透過共同訪問關係高的使用者，進行的隨機遊走傾向於收集與目標 POI 具有相同類別的鄰居，此方法更有可能吸引具有不同場地類別的鄰居，以便表示學習可以產生相似的嵌入，

2. Embedding Learning

基於 skip-gram 模型架構的 LBSN2vec 和 PLBSN2vec 的嵌入學習。採用三層神經網路，由輸入層、隱藏層、輸出層組成。

3. Making Prediction via Similarity

TPD (Targeted Propagator Discovery)任務:

因為相似偏好的使用者間會具有相似的嵌入向量，表示其關係較為相近，所以使用者存取 POI 的機率就越高。所以此篇論文的想法就是，透過計算所有用戶的影響力，並對其排名，找出能那些能夠產生最多人訪問給定 POI 的用戶。

TCD Targeted Customer Discovery 任務:

TCD 則是考慮 POI 偏好設計訪問評分指標。

IV. Result

1. 在圖 22 的 TPD 的結果中，可以發現 PLBSN2vec 能夠穩定地具有更高的 Precision 值。
2. TCD 是以店家立場去思考，所以他們更關心的是-是否能找到所有潛在客戶，所以採用的評估工具為 recall，在圖 23 結果顯示 PLBSN2vec 有較好的 recall。

此篇研究的貢獻在於找到 Targeted Propagator Discovery (TPD) 和 Targeted Customer Discovery (TCD)，並提出一個網路嵌入模型 LBSN2VEC，基於 LBSN 中的社交網路和使用者 POI 簽到資料來共同學習 POI 和使用者的特徵表示，從實驗結果顯示，LBSN2vec 在 TPD 和 TCD 任務中都可以優於過去的方法。

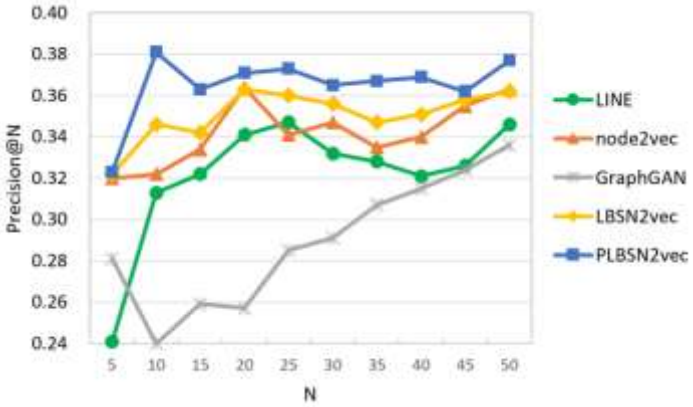

圖 22. TPD 實驗結果圖

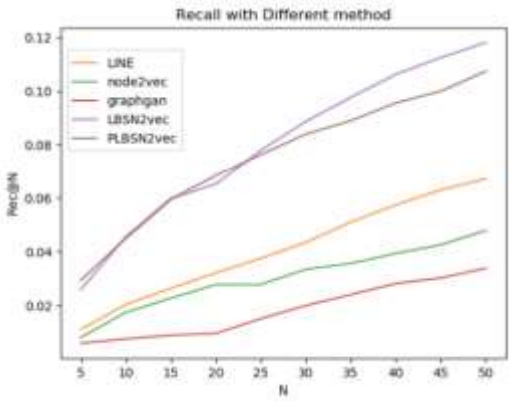

圖 23. TCD 實驗結果圖

#14. Data Mining and Feature Analysis of College Students' Campus Network Behavior

I. Target Problem

作者認為學生事務管理的一大問題來自學生輔導員精力有限與學生行為多樣性之間的矛盾，導致許多潛在問題學生失去早期介入的機會。

此研究透過大數據來了解學生的校園網路行為特徵，同時對學生的在校行為進行量化分析，從而在大型資料中挖掘對學生輔導員有用的信息，進一步解決學生事務管理問題。

II. Dataset

資料的收集來源主要包括學生在校園使用網路的資料和學生在校的行為資料。其中，校園網路使用資料主要包括以下 8 個: ID、上線日期、離線日期、線上時間、離線時間、入境流量、出站流量、總流量。其他學校資料包含成績、校園卡片消費數據、圖書借閱數據、體能測試數據等。作者進行資料清理後，最後將作為分析的資料為大學 39 個月內 3245 位學生網路行為的 28 個分析指標。

表 16. 資料集資訊

Index number	Index Name
1	Time
2	Flow
3	AveTime
4	AveFlow
5	DayTime
6	NightTime
7	AveDayTime
8	AveNightTime
9	WeekdayTime
10	WeekendTime
11	AveWeekdayTime
12	AveWeekendTime
13	DayFlow
14	NightFlow
15	AveDayFlow
16	AveNightFlow
17	WeekdayFlow
18	WeekendFlow
19	AveWeekdayFlow
20	AveWeekendFlow
21	FlowTime
22	DayFlow/DayTime
23	NightFlow/NightTime
24	WeekdayFlow/WeekdayTime
25	WeekendFlow/WeekendTime
26	Times
27	Flow/Times
28	Time/Times

III. Data Mining workflow and main techniques

1. 資料處理

此研究首先對 3245 名學生的 28 項研究指標資料進行平滑處理，接著透過 B-spline basis function 對平滑後的曲線數據進行近似，將平滑後的數據轉換為 functional data，並對其進行進一步的主成分分析。基於變異數比例大於 90%的原則來選取主成分，並提取選定的主成分係數，將選定的主成分係數向量作為下一步資料探勘的基礎。

2. 資料探勘

此研究將函數資料主成分分析所得的係數向量進行 Cluster 分析，作者使用 K-means clustering analysis 對具有不同校園網路使用模式的學生群體進行分組。透過分析不同類別的特徵，可以更了解學生在校的網路使用模式，有助於發現那些需要高度關注的網路重度使用的學生族群。由於 K-means clustering method 須先確定最佳 K 值，也就是聚類的數量，所以作者透過 Hubert 指數和交叉驗證來確定為 4 個類別，如下圖所示。其中，實作過程主要基於 Rstudio 中 Nbclust 套件和 Factoextra 套件的演算法。

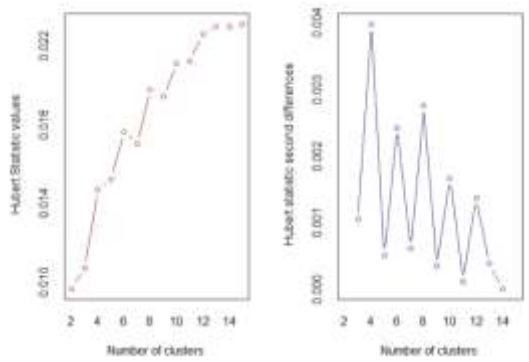

圖 24. 資料探勘-Cluster 分析

首先，K-means clustering algorithm 以上述建構的 28 個研究指標作為輸入，最終輸出每位學生的類別標籤。透過分析，將 3245 名學生分為 4 類: 第一類至第四類依序為 230 人、958 人、1707 人、350 人，下圖也可觀察到不同類別的男女比例。

表 17. Cluster 分類結果

Category	Basic Information				
	Number	*Male*	*Percentage of Male*	*Female*	*Percentage of Female*
1	230	58	25.2%	172	74.8%
2	958	795	83.0%	163	17.0%
3	1707	1362	79.8%	345	20.2%
4	350	251	71.7%	99	28.3%

此外學生群體校園網路使用特徵包含線上時長、線上流量、流量/時間比等，隨後將進一步分析各群體流量/時間比的差異。由圖 25 可知，第一類至第四類的 flow/time ratio 依序增加，對各群體學生 39 個月的 flow/time ratio (如圖 26)進行分析可以發現，當學生進入大學二年級後(10-20 個學期月)，flow/time ratio 顯著下降，且每個類別都有相似的規則。隨後，進入大學三年級(20 個學期月後)，第三類和第四類的 flow/time ratio 明顯增加，其中第四類明顯高於其他三類。

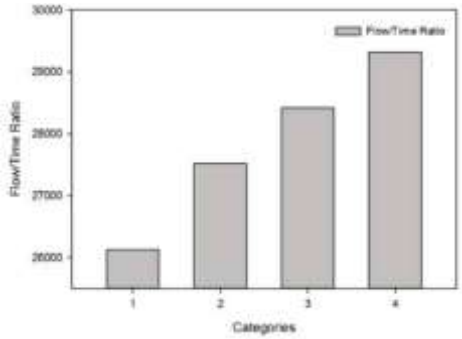

圖 25. flow/time ratio

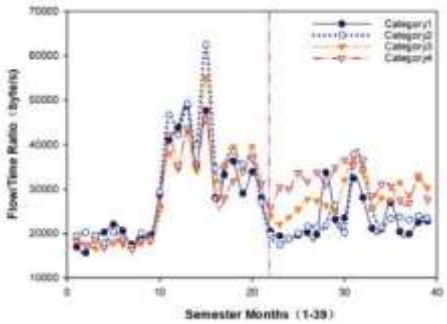

圖 26. flow/time ratio of 39 month

除了上述學生校園網路使用的特性與分析外，其他學生在校行為的數據也值得進一步綜合考量，例如在成績表現方面，統計了各類別學生的算術平均分數和加權平均分數，結果顯示，一、二、三、四類學生的平均分數依次下降。

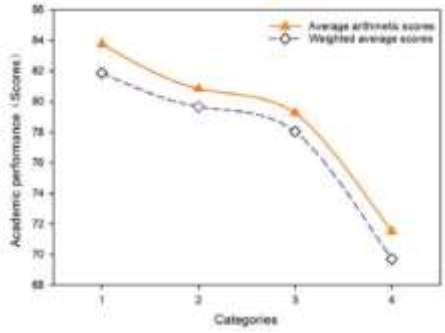

圖 27. Average performance

VI. Results

經過分析所得到的四類學生群體中，總流量結果顯示，第三類略高，第一類最低，第二類和第四類接近。四類學生類別流量差異較小。從流量/時間比來看，第一類、第二類、第三類、第四類依序上升，第四類明顯高於前三類，也就是說第四類學生屬於經常進行高流量網路行為群體。經結合其他學生在校行為數據，第四類的課程成績、長跑考試成績、食堂早餐數量均低於其他三類，表示這類學生的學習能力、身體健康、生活習慣已大受影響，因此校園網路的使用程度可能是重要的影響因素。最後，學生輔導員也能夠就上述的分析結果找到需要高度關注的學生，並及時給予幫助。

15. Traffic Information Mining From Social Media Based on the MC-LSTM-Conv Model

I. Target Problem

隨著經濟的快速發展和城市的蓬勃發展，氣機車的數量也隨之升高，交通問題已成為世界各大城市的困擾。交通問題不僅影響城市間的物流運輸，也影響生活品質。

此研究提出一個基於深度學習的方法來處理交通資訊的社群媒體資料，建立一系列規則來獲取有關交通擁堵的詳細資訊。

II. Dataset

- **Dataset Name**
  Sina Weibo (新浪微博)
- **Dataset Link**
  論文中沒有提供連結，但有提到可以使用 Sina Weibo 的 API 獲取資料集。

III. Data Mining workflow and main techniques

1. MC-LSTM-Conv to Extract Microblogs about Traffic Jams From Mass Data

此步驟主要達成兩件事。首先，將文字轉換成可以被深度學習模型理解的數學向量和矩陣。接著，再利用 MC-LSTM-Conv 進行特徵提取和分類，並利用這些特徵更好地從眾多數據中識別有關交通擁堵的微博。以下將針對 MC-LSTM-Conv 做詳細介紹:

模型包含兩個通道 channel 1 和 channel 2，兩個通道架構相似，包含捲積層、LSTM 層以及池化層，用於將鋸子轉換為特徵向量，將這兩個通道的結果輸入到全連接層，通過 soft-max 層完成分類。
通道的目的不一樣，channel 1 是找到相鄰單字間的小範圍依賴關係，而 channel 2 是讓每個單字嵌入受到前文影響，獲取上下文資訊。

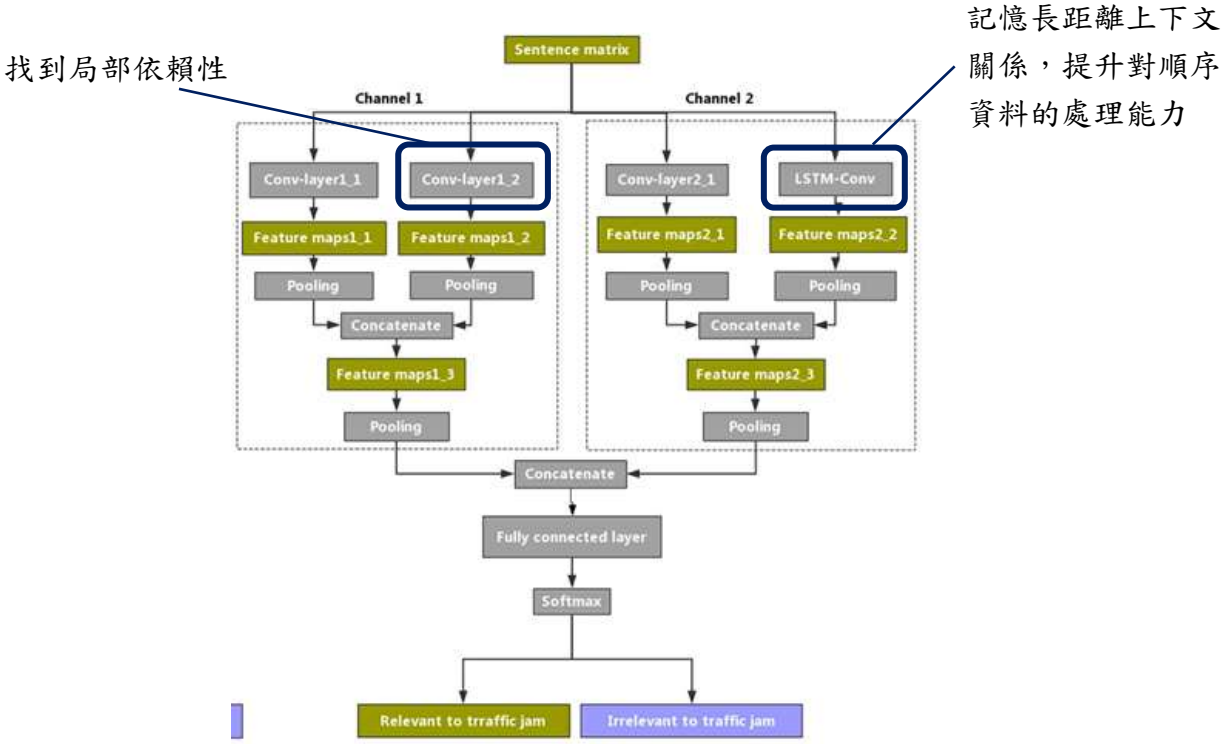

圖 28. MC-LSTM-Conv

2. Keyword Fuzzy Matching Method to Mine Detailed Traffic Information

透過 MC-LSTM-Conv 方法，從大量的新浪微博資料中獲得了一組關於交通擁堵的打卡微博。為了提供更詳細的各地交通擁堵信息，將反映交通擁堵事件的微博分為四類：交通事故或大型活動、道路建設相關、交通號誌故障或設定不合理有關、政府機構的工作相關，使用 Google 搜尋語法，計算關鍵字與關鍵字間的距離，為每個類別建立相關規則，以此作為分類依據。

圖 29. The overall workflow of research

IV. Result

1. 探討了 MC-LSTM-Conv 模型中 Conv 層的最佳參數設置，並透過與其他模型的比較驗證了其有效性和優越性。
2. 關鍵字模糊匹配方法在提取交通擁堵詳細資訊方面的表現，可以發現 F1 四個類別都高於 0.80，說明微博分類規則是有效的。

表 17. 各方法分類表現實驗結果圖

Id	Method	Acc	Pre	Rec	F1
1	MC-LSTM-Conv	**0.923**	**0.923**	0.920	**0.921**
2	MC-Conv	0.917	0.919	0.912	0.915
3	CNN	0.919	0.914	**0.922**	0.918
4	LSTM-CNN [23]	0.905	0.902	0.903	0.902
5	LSTM-CNN [24]	0.910	0.907	0.908	0.907
6	LSTM	0.893	0.885	0.896	0.890
7	RNN	0.869	0.870	0.870	0.870
8	SVM	0.757	0.755	0.752	0.753

本文提出了一個名為 MC-LSTM-Conv 的深度學習模型，有效地從微博中提取語義特徵。與其他深度學習模型相比表現更好。

本文也採用關鍵字模糊配對的方法，將與交通壅塞相關的微博分為四類，每一類都反映了具體的路況。

表 18. 不同類別下的分類影響

Id	Pre	Rec	F1
1	0.886	0.802	0.840
2	0.816	0.886	0.849
3	0.879	0.804	0.837
4	0.895	0.993	0.942

Chapter 3.

A Summarized Table

 # 1.

Bridging Performance of X (formerly known as Twitter) Users: A Predictor of Subjective Well-Being During the Pandemic

Year: 2024

Target Problem: 了解具有影響力的社群用戶在 COVID-19 下的心理影響

Adopted Approach:構建衡量用戶與社群媒體的橋連接，並以此揭示主觀幸福感和橋連接的相關性

Datasets: Twitter

Result & Contribution:疫情期間使用者在訊息傳播方面的橋連接表現與其主觀幸福感之間存在負相關

 # 2.

Partial Data, Potential Exposure: Evaluating Privacy Leakage via GNNExplainer on Social Networks

Year: 2024

Target Problem:使用者透漏部分內容時，是否能進一步推測並獲取其更為私密的資訊

Adopted Approach: GNNExplainer 生成可解釋性子圖，並通過這些子圖去預測使用者間的關聯性。

Datasets: Facebook, GitHub

Result & Contribution:結果顯示若僅披露部分資訊，仍有可能導致隱私資訊批露

 # 3.

Data-Mining of Social Media Users with Embedding Techniques and Neural Network

Year: 2024

Target Problem:深入研究使用者在社交媒體平臺上留下的自然語言文本數據

Adopted Approach:採用特徵嵌入技術和神經網路建模方法來研究變數之間潛在的複雜關係

Datasets: Kaggle

Result & Contribution:自然語言文本數據如個人化簽名透過嵌入技術納入神經網路建模過程，能對社群媒體分析和用戶分析領域造成重大影響。

 # 4.

Dual Graph Networks with Synthetic Oversampling for Imbalanced Rumor Detection on Social Media

Year: 2024

Target Problem:在類別不平衡的情況下判斷言論的真實性

Adopted Approach:提出一個新的合成技術，用於生成少數類別

Datasets: PHEME, RumourEval

Result & Contribution:設計一個新的有效生成過採樣樣本，比傳統 SMOTE 有更好效果

 # 5.

Interactive Activities Initiation through Retrieving Hidden Social Information Networks

Year: 2023

Target Problem:如何在未知拓撲的社交網絡中以最小的成本與資源來啟動用戶間的互動活動從而實現網絡內的全域覆蓋。

Adopted Approach:用一種基於流量偵測器和頂點覆蓋求解器的互動式演算法，透過每輪從網絡中隨機採樣少量邊來逐步構建網絡結構，最終達到有效覆蓋整個隱藏社交圖的目標。

Datasets: Facebook、Github、Gowalla、Youtube

Result & Contribution:經由作者開發的隨機隱藏輸入計算模型及其對應的近似演算法，且透過大量實驗，能夠實現少量的邊緣取樣來達到有效的頂點覆蓋。

 # 6.

ANTI-Disinformation: An Adversarial Attack and Defense Network Towards Improved Robustness for Disinformation Detection on Social Media

Year: 2023

Target Problem:解決虛假資訊不斷發展的問題，提出“Adversarial Network Towards Improved robustness for Disinformation detection”概念

Adopted Approach:提出的架構主要由兩個部份組成，分別是對抗性攻擊模型以及對抗性防禦模型

Datasets: Instagram、Amazon Vine、Twitter

Result & Contribution:利用強化學習作為對抗性攻擊並引入防禦模型增強偵測系統，經過實驗結果證明能處理複雜的對抗性攻擊，以保護線上使用者免受虛假資訊

 # 7.

A Machine Learning Based Social Network Data Mining System for Better Search Engine Algorithm

Year: 2023

Target Problem:設計一套系統並透過系統獲取有效的社交網路信息

Adopted Approach:透過特徵學習和可信度標籤構成 Fake Detector，並用來辨識假新聞或消息

Datasets: Instagram、Twitter

Result & Contribution:以公眾輿論為基礎開發的系統，能有效偵測與處理社交平台上的假訊息。

 # 8.

Classifying Severe Weather Events by Utilizing Social Sensor Data and Social Network Analysis

Year: 2023

Target Problem:因缺乏高品質天氣資料，通過將氣象數據與社交媒體結合，提升異常天氣預測

Adopted Approach:探勘有用資料，並適當的整合氣象數據與社交媒體資料

Datasets:論文自己收集，可參考上面看更完整解釋

Result & Contribution:利用社群媒體貼文預測天氣是有幫助的，並提供資料特徵提取的想法和流程

 # 9.

Social Media Sensors for Weather-Caused Outage Prediction Based on Spatio–Temporal Multiplex Network Representation

Year: 2023

Target Problem:學習時空多重網路來提高與天氣相關的停電預測準確性

Adopted Approach:建立 Multiple Social-Power Network Creation 增強其關聯性

Datasets:論文自己收集，可參考上面看更完整解釋

Result & Contribution:提出新的架構方式-多層網路的連接，結果顯示有更好的學習效果

 # 10.

Unsupervised Post-Time Fake Social Message Detection with Recommendation-aware Representation Learning

Year: 2022

Target Problem:新發布訊息時，因沒真實性標記和沒使用者回饋，而無法及時辨識假訊息

Adopted Approach:提出 RecMR 架構找出對此訊息感興趣的潛在用戶，並模擬可能的互動方式

Datasets: Twitter-15, Twitter-16

Result & Contribution:透過模擬用戶互動，在不需有真實用戶的情況下，辨識假訊息

 # 11.

Predicting and Analyzing Privacy Settings and Categories for Posts on Social Media

Year: 2022

Target Problem:如何有效預測並推薦社群媒體貼文的隱私設定與隱私類別

Adopted Approach:利用 LIWC 分析提取貼文特徵，再利用圖嵌入方法進行特徵學習，最後進行隱私類別和設定的多任務學習。

Datasets:使用 A personal privacy preserving framework: I let you know who can see what 此篇 dataset

Result & Contribution:透過圖嵌入方法進行貼文特徵學習實現有效預測/推薦社群媒體貼文的隱私設定

 # 12.

Social Network Analysis of Popular YouTube Videos via Vertical Quantitative Mining

Year: 2022

Target Problem:設計一種挖掘有趣的定量頻繁模式的演算法

Adopted Approach:透過 Q-Eclat 演算法將定量交易資料庫的水平格式轉為垂直並計算 Ck 中所有候選 k-itemexpsets 的集合，接著使用 pruning rulse 移除資料並找出 frequent k-itemexpsets。

Datasets:自行合成的數據集與 UCI Machine Learning Repository

Result & Contribution:提出的定量垂直 Q-Eclat 演算法挖掘定量頻繁模式所需的運行時間比現有的 MQA-M 演算法更短，進一步增強定量頻繁模式挖掘的方法。

 # 13.

Finding Potential Propagators and Customers in Location-Based Social Networks: An Embedding-Based Approach

Year: 2020

Target Problem:找到 TPD 和 TCD，以找到潛在的目標傳播者和客戶

Adopted Approach:提出一個網路嵌入模型 LBSN2VEC，找出有影響力用戶。

Datasets: Instagram

Result & Contribution: LBSN2vec 在 TPD 和 TCD 任務中都可以優於過去的方法

 # 14.

Data Mining and Feature Analysis of College Students' Campus Network Behavior

Year: 2020

Target Problem:透過資料探勘來了解學生的校園網路行為特徵，同時對學生的在校行為進行量化分析，從而挖掘對學生輔導員有用的信息。

Adopted Approach:對資料進行平滑處理再利用 K-means clustering method 將學生分為四群，以進行學生的各項行為分析。

Datasets: 學生校園網路使用資料和其他校園行為資料。

Result & Contribution:透過資料探勘方法幫助學生輔導員找出校園網路重度使用者族群或是因而導致行為不佳的學生，以及時給予必要協助。

 # 15.

Traffic Information Mining From Social Media Based on the MC-LSTM-Conv Model

Year: 2020

Target Problem:利用社群媒體資料，建立一系列規則來獲取有關交通擁堵的詳細資訊

Adopted Approach:提出 MC-LSTM-Conv 模型，並利用關鍵字分類

Datasets: Sina Weibo (新浪微博)

Result & Contribution:提出新的模型，能有效地從微博中提取語義特徵，並在後續分類有更好表現

Chapter 4.

Comparative Study & Discussions

在分析了 15 篇關於 Social Network 的論文後，我們依據其研究主題將其分為以下四大類別：

類別	論文編號
1. 假訊息相關研究	#4, #6, #7, #10
2. 隱私問題研究	#2, #5, #11
3. 社交平台關係探勘與分析	#1, #3, #12, #13
4. 特殊應用研究 (天氣/大眾運輸工具/教育)	#8, #9, #14, #15

接下來，我們將依照上述四個類別的論文逐一分析其優缺點，並延伸探討其他可能的未來工作。最後，我們將針對我們選定的主題-社交網路 (Social Network) 進行深入分析，最終提出總結與討論。

1. 假訊息相關研究

論文

 # 4. Dual Graph Networks with Synthetic Oversampling for Imbalanced Rumor Detection on Social Media

 # 6. ANTI-Disinformation: An Adversarial Attack and Defense Network Towards Improved Robustness for Disinformation Detection on Social Media

 # 7. A Machine Learning Based Social Network Data Mining System for Better Search Engine Algorithm

 # 10. Unsupervised Post-Time Fake Social Message Detection with Recommendation-aware Representation Learning

優缺點

[#4]這篇論文旨在利用提出的新合成技術生成更具有意義的資料點，而不只是依賴 SMOTE 技術，因為 SMOTE 可能生出重複且缺乏意義的點。我覺得這篇論文的優點，在於為假訊息相關研究提供了一個解決訊息真實與虛假資料不平衡的有效方法，讓未來在延伸相關問題時，能夠產生更多有價值的資訊點。此外，我覺得這篇論文有將文章的語意也作為模型的訓練資料，這不僅使得模型能夠更全面地理解使用者的行為和意圖，還提升了對假訊息的辨識能力。

然而，對於論文中考量語意的部分，我抱持一個疑問: 就是在多種語言環境下，模型訓練生成的點是否有一樣的表現?因為這篇論文有提到語言訓練的部分是採用 BERT 語言模型，但原版的 BERT 是基於英語的大型語言庫，在非英語的語言比較弱。雖然 BERT 有在多語言版本下的模型，但因為語言規模較小，所以我有點好奇在這樣的情況下。比如: 如果是中文的貼文下，採用此篇的方法並只加入社群使用者資料和採用此篇方法並加入社群使用者資料和語意哪種訓練結果會比較好?

[#6]這篇論文旨在提出一個不僅能對抗和防禦的模型，以促進對於不斷變化虛假資訊傳播的行為有相關適應性。我覺得這篇論文的優點在於提出這種對抗和防禦型模型，來偵測對於假訊息的偵測，跟另外 3 篇有很大的不同，其他篇比較像是加強一般流程的某個細節，這篇是直接針對框架的模型去進行修改，讓模型透過自我學習不斷增強以適應假訊息不斷的變化，假訊息不斷的變化也相當符合當代的情況。

這篇我覺得沒有太大缺點，但我有點好奇這樣的方法，如果隨著訓練資料量增加，是否會有梯度爆炸的機會呢?因為我似乎沒有在這篇論文的架構圖中看到使用 ReLU、LeakyReLU、Batch Normalization，也沒看到論文提到 Gradient Clipping，不太確定是否是有加入沒寫出來，還是其實不會遇到這個問題，因為過去我比較常聽到是在圖片中利用生成對抗網路，所以不確定是不是因為都是用文字的關係，所以資料量沒有那麼大，不太會遇到這個問題。

[# 7]這篇論文主要提到如何有效挖掘出有效資料。這篇我覺得更過去以往在挖掘社群網路資料的有效性時，通常沒有考慮資料的真偽性，所以可能就也把假消息一起放進去做學習，當然成效也包含這些假消息。而這篇論文最大的優點在於有先對假消息做辨識，使用 Fake Detector 學習特徵並有可信度的標籤，以偵測假新聞，並將這些資料做特殊處理。這讓使得在挖掘社群網路資料是相當有功效的，有點像是在做資料的前處理，把那些異值刪掉，已得到更純粹的資料集。

這篇我覺得沒有太大缺點，但我有點好奇這樣的方法，在這個前處理的成本有多大。這篇是單純只在找社交媒體上的關聯性，但找到的這些關聯性如果實際應用在實例中，比如:想建立推薦系統，推薦系統背後就是要找到社交媒體背後的祕密，並進而推薦用戶相關資訊，但這個的前處理等於做了其他篇論文一篇在做的事情，感覺成本有點大的前處理，但如果這個辨別假訊息的模型預訓練好像也可以，但就又跟上篇論文提到的假訊息不斷變化，這個預訓練模型好像準確度又變得不高了，所以我想問的是這樣的成本和結果有成正比嗎?

[# 10]這篇論文主要是透過模擬使用者互動來預測假使一篇貼文出來時，沒有任何回應有辦法判斷嗎?我覺得這篇論文的優點在於提供一個新觀點，其他篇論文基本上都是用已存在的使用者回應，但這篇論文是利用使用者的社交關係去推論當這篇貼文出來時，其他使用者可能會怎麼會應，用模擬方式去取代必須要用已存在的使用者回應。

這篇論文提出的觀點我相當喜歡,不覺得有缺點。用過往社交圈重建模擬去模擬取代必須要有的資訊，此種應用的情境是有即時性的且意義在於不需要及時會饋。這讓我思考如果是應用在自然災害上的能源分配上，因為自然災害發生時，可能無法立刻獲得實際回應結果，如果利用社交平台的發布災況，去建立社交平台的分析，並模擬災況等級，是否就能及時針對災情的資源進行分配。

Possible improvements and extensions

總和來說，在假訊息相關研究，這幾篇論文都給我帶來許多啟發，啟發內容也在上面各段提及。接下來我會針對假訊息相關研究的這個主題做討論,這四篇論文都有提及假訊息對生活或是社交媒體上的影響，並設計相關方法去預防，但並未考慮”無惡意的假訊息傳播者”，因為並不是所有人的媒體識讀能力相當高，所以就有”無惡意的假訊息傳播者”的出現，他們的行為是可能有分享過假訊息，但在接收到反駁訊息時，也會分享正確訊息，已對自己分享錯誤資訊做一個更正，這類人不應該被歸類在惡意傳播者，所以若是能針對群體有更仔細的分類，而不單單只是分成二元分類(惡意或非惡意)，這樣對於真正的惡意使用者也能更有效的預防且阻止。

2. 隱私問題研究

論文

 # 2. Partial Data, Potential Exposure: Evaluating Privacy Leakage via GNNExplainer on Social Networks

 # 5. Interactive Activities Initiation through Retrieving Hidden Social Information Networks

 # 11. Predicting and Analyzing Privacy Settings and Categories for Posts on Social Media

優缺點

關於上述三篇所提到的技術在處理社交網路隱私問題時,均展現了對複雜網路數據結構和用戶行為的深刻洞察。這些技術的優勢在於能針對不同場景提供針對性的解決方案，例如：GNNExplainer 生成的解釋性子圖有助於理解隱藏用戶的關聯性;基於流量偵測和頂點覆蓋求解的互動式演算法在重建隱藏網路結構方面具有高效性；結合 LIWC 分析套件和圖嵌入的多任務學習能有效捕捉貼文文本特徵與隱私設置之間的關聯性。然而，這些方法依然各自面臨了一些限制，像是解釋性子圖可能不小心揭露網路結構，因而存在隱私洩露的風險，而在重建隱藏網路結構中多輪採樣的方式可能引發用戶對數據訪問頻率提高的疑慮，另外，語言特徵分析對多語言的處理可能存在進步的空間。因此，未來若是能針對這些限制進行改善，也許能進一步提高技術的多元應用能力。

Possible improvements and extensions

1. 提升模型的隱私保護機制：像是引入隱私保護演算法，以確保數據處理過程中不洩露用戶個人信息。例如，GNNExplainer 生成的子圖可以通過隱私優化模型來控制其揭露的資訊範圍，使能夠同時兼顧透明性與隱私保護。

2. 多模態數據的隱私處理：現有技術主要針對單一模態的數據進行處理，例如文本或結構化數據，無法充分利用社交網路中的多模態資訊，像是音訊和影片。也許未來可設計多模態整合模型，同時加入隱私保護策略以便處理更多情況。

3. 高效的數據採樣與分析：在基於流量偵測的演算法中，利用強化學習設計自動採樣策略，使系統能動態調整採樣的範圍和頻率，減少重複且多次訪問數據帶來的風險與資源浪費，同時提升識別重要隱藏網路結構的效率。

4. 法律和道德倫理：將現行的隱私保護法規如 General Data Protection Regulation(GDPR) 嵌入模型設計，確保技術應用符合法律與道德要求。例如，設計能自動識別和避免處理未經授權的敏感數據的系統。

3. 社交平台關係探勘與分析

論文

#1. Bridging Performance of X (formerly known as Twitter) Users: A Predictor of Subjective Well-Being During the Pandemic

#3. Data-Mining of Social Media Users with Embedding Techniques and Neural Network

#12. Social Network Analysis of Popular YouTube Videos via Vertical Quantitative Mining

#13. Finding Potential Propagators and Customers in Location-Based Social Networks: An Embedding-Based Approach

優缺點

以上四篇研究分別從圖結構分析、自然語言處理、頻繁模式挖掘和嵌入學習等不同角度出發，探索了社交平台用戶關係的挖掘與分析。每種技術都在特定的情境下表現出強大的應用潛力與技術優勢，不過，卻也存在著各自需要面臨的挑戰。

首先，橋連接分析方法透過級聯樹以計算用戶的橋連接值，進而有效衡量了用戶在傳播訊息中的重要性。然而，這種方法對於結構化數據的依賴性較高，並且主觀幸福感的變化與橋連接表現之間的因果性難以直接證明。

其次，由於自然語言處理的預訓練技術 BERT 和時間循環神經網路 LSTM 的文本建模方法適用於處理多語言文本並捕捉語義特徵，因此可以看出上述兩種方法在處理自然語言數據方面的價值。然而，該方法對於大規模語言模型的計算資源需求較高，運行效率也相對面臨挑戰，且 LSTM 在處理長文本、長序列時可能因為難以捕捉序列中遠距離的依賴關係，導致性能下降明顯。

第三，Q-Eclat 演算法在高效挖掘定量頻繁模式方面表現優異，對於大規模數據的分析提供了優化路徑。然而，此技術偏重於挖掘頻繁模式，缺乏對數據語義層面的深入理解，難以應對社交行為中的複雜互動模式。

最後，LBSN2vec 模型利用嵌入學習，成功構建使用者與 Points of Interests(POI) 的特徵表示，實現了對目標傳播者(TPD)與目標顧客(TCD)的高效識別，特別適合基於地理位置的社交網路分析。但該方法主要依賴使用者與 POI 的數據，對於非地理相關的社交互動探索能力有限，並且需要大量簽到數據來保證模型的精確度。

Possible improvements and extensions

1. 結合因果推斷與深層解釋：在橋連接分析中，主觀幸福感與訊息傳播表現的關係比較像是統計相關，而非因果關係。若是引入因果推斷方法，也許可以進一步揭示訊息傳播行為對幸福感的直接影響，同時避免模型過度依賴相關性假設。
2. 引入壓縮技術：對於 BERT 和 LSTM 的文本建模技術加入模型壓縮方法，如 Quantization 和 Pruning，以降低模型規模，同時保持準確性，減少大規模語言模型對資源的依賴。
3. 結合語意分析技術：在頻繁模式挖掘前，加入自然語言處理的嵌入技術(例如： Word2Vec)來學習數據中的語義特徵，將語義嵌入整合進頻繁模式挖掘過程，使算法更具語意敏感性。
4. 混合多元數據嵌入學習：除了社交網路中地理位置數據以外，可以考慮結合其他社交網路訊息，像是用戶的興趣標籤、互動紀錄等資訊，拓展嵌入模型的輸入型態範圍，使其不僅依賴於地理位置數據，進一步增強模型對用戶行為變化的處理能力。

4. 特殊應用研究 (天氣/大眾運輸工具/教育)

論文

 # 8. Classifying Severe Weather Events by Utilizing Social Sensor Data and Social Network Analysis

 # 9. Social Media Sensors for Weather-Caused Outage Prediction Based on Spatio Temporal Multiplex Network Representation

 # 14. Data Mining and Feature Analysis of College Students' Campus Network Behavior

 # 15. Traffic Information Mining From Social Media Based on the MC-LSTM-Conv Model

優缺點

[# 8]、[# 9]這兩篇都是對異常天氣的預測與應用，而且都是出自同一個實驗室，論文架構與想法都蠻類似，所以在這裡我將會一起討論。這兩篇都是將社交媒體的特色-使用者能隨時隨地將在地情況分享到社交平台，這樣對於異常天氣的資料收集無異是增加機會。我覺得這兩篇的優點都是提出一個新觀點，看似將沒有相關的議題放在一起作為預測。

但，同時這兩篇也讓我不禁思考，在這假訊息崛起的時代，如何確保收集到的資料都是真實的，在這兩篇論文中都沒有考慮訊息的真假，只有考慮訊息是否符合地域性，所以我認為這是此篇最大可以改進的地方，因為異常天氣的資料量本身就已經夠少，如果再透過連接社群資料的假資料，反而會使資料集更顯得雜亂無章，所以我覺得再訓練前，可以針對從社群媒體收集到的資料做清洗。

[# 14]此篇論文是從學生的社交平台的行為與在校表現結合,去幫助學生輔導員挖掘並掌握學生動態。我認為此篇的優點在於分析相當仔細，不但是以不同 cluster 方式對學生分組，又針對學生的不同上網時間去做分析，將學生對於社交平台的特徵都考慮相當清楚。

但，我對於此篇的實際應用抱有懷疑，因為本身是幫助學生輔導員更好掌握學生動態，但這可能就會犧牲學生的隱私，所以覺得這項研究更適合在研究領域，而不適合在應用層面。

[# 15]此篇論文是將社交平台去預測現在的交通情況。我覺得這篇論文最大的優點是在建立 MC-LSTM-Conv，MC-LSTM-Conv 的目的是進行文字特徵提取和分類，在這篇論文不單就只考慮文字的局部型，而是同時利用 LSTM 記憶長距離上下文關係，提升對順序 資料的處理能力。這讓我學習到如何對於貼文進行分析。

此篇論文我覺得沒有太大的缺點，只是有點小疑問，因為這篇論文在提取出關鍵字後，是使用 Google 搜尋語法，計算關鍵字與關鍵字間的距離，為每個類別建立相關規則，以此作為分類依據。那為什麼不用模型訓練，更適合不同情況，而不只是被特定語句限制，如果在這裡用類似像 Random Forest 這種方法會有更好表現且適合更多元情況嗎?

Possible improvements and extensions

總和來說，在特殊應用研究 (天氣/大眾運輸工具/教育)中，這些論文提供不同觀點，帶出社群平台與我們生活各處的連結，如何好好應用是值得我們思考的。

Mobile & Social Network improvements and extensions

在以 Social Network 為主題的應用，其核心概念在於建立社群平台上的使用者互動關係的連接，並將其應用於各種領域上，例如:辨識假訊息、隱私披露風險分析、社交網路深層探勘、天氣預測、交通運輸管理以及教育創新上，為許多領域提供創新的解決方法。

經過閱讀多篇論文後，我們不禁思考: 若將 Social Network 作為主題，並透過 Google Scholar 作為平台，為所有研究學者建立其社交網路，以增加學生找相關論文的便利性，是否可行?以下將是我們經過閱讀後的論文並運用其相關技術想到的提案:

I. Problem Statement

對於研究學者，尤其是剛入門的初學者而言，尋找優秀且高品質的論文是一項相當巨大的挑戰。其原因在於初學研究學者對於會議和期刊的等級，以及頂尖教授或研究學者的研究領域和貢獻缺乏了解，所以可能造成花了大量時間找到的論文也有可能是錯誤資訊，這往往需要花費大量時間和經驗累積，甚至依賴他人指導，才能找到合適的研究資源。這使初學者與研究領域間形成明顯的鴻溝。

II. Our Observe

事實上，在 Google Scholar 已提供相關有用資訊，包含:顯示研究者的研究領域和其過去論文的相關合作者。我們將以專注於 Social Network 領域的李政德教授為例(現任教於成功大學)，說明我們觀察到可以利用的有用資訊。在圖 29 中顯示李教授的 Google Scholar 頁面，其中包含李教授的研究領域和過去的合作學者。接著，我們隨著點開一個過去共同合作者 Yu-Che Tsai 如圖 30 顯示，可以發現這位的共同合作者的研究領域也是跟李教授重疊，另外這位共同合作者的共同合作者，基本上也是與李教授重疊，我們觀察其他領域的研究者其 Google Scholar 也有相同情況。這主要是因為在學術界中的研究學者通常會有屬於跟自己研究領域相仿的圈子，而這層關係可以從各研究學者的研究生涯開始慢慢累積。

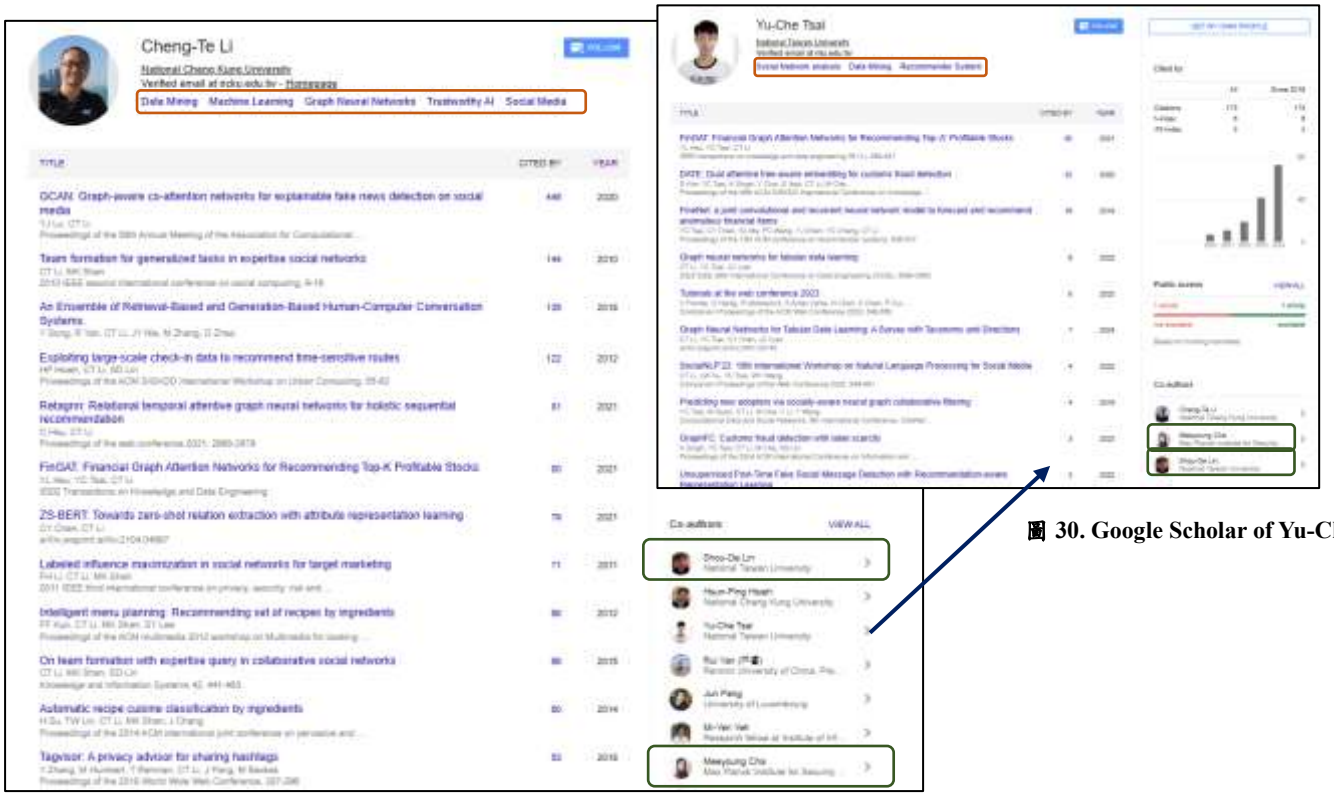

圖 30. Google Scholar of Yu-Che Tsai

圖 29. Google Scholar of Li Professor

III. Possible Work

從上面的觀察，我們可以發現，優秀的學者通常合作對象也是高水準的學者，例如:李教授目前任教於成功大學，而 Yu-Che Tasi 的機構也是登記在台灣大學。

利用觀察到的特性並結合我們看的相關論文，我們認為可以利用 Google Scholar 去建立其 Social Network，並結合各學者的經歷，例如:大學、研究所的實驗室、指導教授、後研究學者經歷等，建立完整的 Social Network。此網路旨在形成優秀的學術研究圈，讓初學者能夠從一個優秀學者找到另一個優秀學者，慢慢擴展對研究領域的了解。

我們將提出初步簡單的 Work Flow 作為探討此主題的可行性。

1. 資料蒐集

從 Google Scholar 提供的資料作為核心輸入，並搭配此學者的一些經歷。因此，輸入可能包含:學者姓名、研究領域、機構、碩博士學校、實驗室、指導教授、博士後研究機構、論文合作對象、合作頻率、論文相關資訊。

2. 資料前處理

可能可以對研究機構進行排名，以找出較為優秀論文品質較為穩定的機構。

3. 建立學術社交網路圖

以學者或是研究機構作為節點，並以邊作為其關聯性。

4. 資料探勘

- 關聯分析:挖掘初學者與學者間的關係
- 研究領域分類:將各研究領域挖掘出子社群，例如是以某幾個影響力的教授主導的圈子
- 影響力分析:研究較為活躍或是影響力大的研究學者，以中心化方式分析。

5. 機器學習- 建立推薦模型

- 輸入:學生感興趣的研究領域或是學者
- 輸出:推薦相關高影響力的學者、論文、研究方向
- 模型:使用我們前面看的論文提出的方法-圖嵌入(Graph Embedding)，並搭配其技術

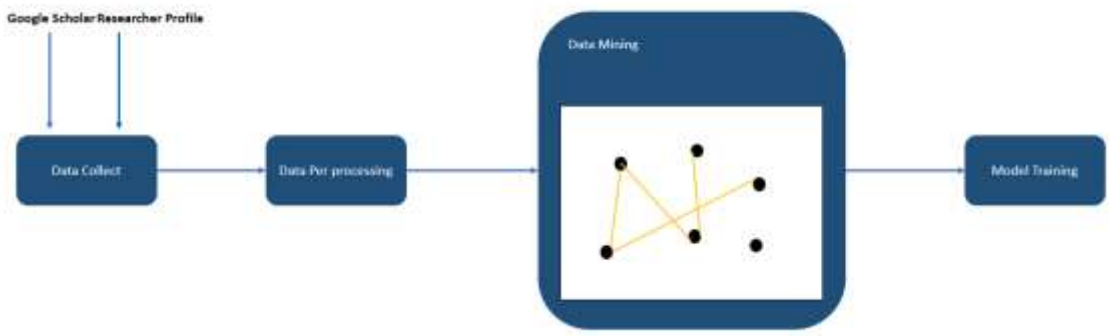

圖 31. Possible Work Flow

此份作業總結

Social Network 中蘊藏著豐富的信息，關鍵在於如何有效挖掘並提取其中的有用資訊。