人格特質與社群媒體文章分析

Introduction to Information Retrieval and Text Mining
Prof. Chien Chin Chen
Information Management Dept. at National Taiwan Univeristy

陳姵如†
資訊管理學系三年級
國立台灣大學
B06406009@ntu.edu.tw

林語萱
資訊管理學系三年級
國立台灣大學
B06705026@ntu.edu.tw

黃冠文
圖書館資料學系四年級
國立台灣大學
B05106010@ntu.edu.tw

趙禹誠
資訊管理學系三年級
國立台灣大學
B05502058@ntu.edu.tw

壹. 摘要

本報告旨在訓練出一項人格分析模型，利用使用者於Instagram上發表之文章內容推測出其在社群媒體所展現出的人格特質。本報告以 Jones 和 George (2019) 的課本中提供 Costa 和 McCrae (1986) 所分類出的五大人格特質其中四項之測驗問卷，包含開放性 (Openness)、勤勉正直性 (Conscientiousness)、外向性 (Extraversion)、親和性 (Agreeableness)，蒐集 78 位使用者的測驗結果以及其Instagram貼文作為訓練資料 (Training Data) 建模，並以政治、娛樂、運動、文學、Youtuber 五大類領域中各二十位名人的Instagram貼文做為測試資料，希望能準確預測並分析五類領域名人在社群媒體所展現出的人格特質。

貳. 關鍵字

text mining, big five personality traits, instagram, social media

參. 前言

人格心理學是一項心理學十分發達的分支，具有豐富的理論和應用研究。而分析人格特質有一種最新方法，是依賴於「機器學習」和使用者生成的文本，讓應用程式逐漸「學習」從文本中估測使用者的人格特質 (Dutta 等，2017)。而 Instagram 是目前全球增長最快的社群網站（Wagner，2015），因此我們十分好奇是否可訓練出一種模型，只要利用使用者於Instagram上發表之文章內容即可推測出其在社群媒體所展現的人格特質；我們以Jones和George (2019) 的課本中提供 Costa 和 McCrae(1986) 所分類出的五大人格特質其中四項之測驗問卷，包含開放性 (Openness)、勤勉正直性 (Conscientiousness)、外向性 (Extraversion)、親和性 (Agreeableness)，蒐集一共七十八位使用者的測驗結果以及其Instagram 貼文作為訓練資料 (Training Data)製作出我們的模型，並以政治、娛樂、運動、文學、Youtubers 五大類領域中各二十位名人的Instagram貼文做為測試資料，預測並分析五類領域名人在社群媒體所展現的人格特質。

肆. 文獻回顧

Big Five Personality Traits

大五人格特質模型 (Big Five Personality Trait Model) 被許多學者認可為能夠明確解釋大多數人的人格差異，是性格因素分析研究的產物 (Zhang, 2002)。根據 Taylor 與 MacDonald (1990) 的說法，該模型最初是由 Galton (1884) 提出，並隨後由 Allport 和 Odbert (1936) 和 Norman (1963) 等人進行實

證。其可以理解為一般人格特徵的描述性分類學理論，由五項相互獨立的主要維度組成，分別為神經質（Neuroticism, N）、外向性（Extraversion, E）、開放性（Openness to Experience, O）、可親性（Agreeableness, A）和勤勉正直性（Conscientiousness, C）。

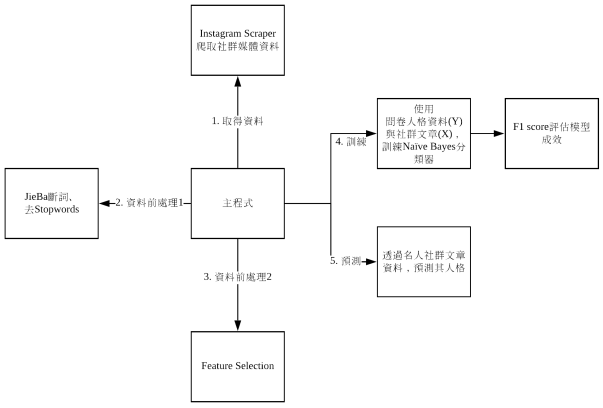
在 Costa 和 McCrae (1992) 的研究中顯示，神經質等級高的人往往會有如情緒反覆、難堪、罪惡感、悲觀和自卑等等的負面情緒；而外向性分數高的人往往善於交際且行事果斷、樂於合作；開放性高者富含想像力，偏愛多樣變化且往往具有到見解，他們通常相對不那麼保守傳統；至於親和性高的人基本上講求利他主義、樂於助人，他們重視並尊重他人的信仰和習慣；勤勉正直性高的人有明確目標、堅強意志，他們負責任也十分值得信賴。五項維度的人格具有各自明確的特質，相互獨立。

而社群網絡 (Social Network) 對於將網際網路從眾所皆知的資訊傳播空間轉變為使用者參與創建和生成自己個人內容文章的地方這件事而言至關重要 (UK, 2018)，而後者正是我們所關注到的，人們開始傾向於在網路上產出與分享自身的心情與想法。因此我們想從社群軟體 Instagram 上使用者發表的個人文章中分析出其較為明顯的人格特質，將這些實證數據嘗試以五大人格特質模型進行歸類解釋。

Instagram 目前是全球增長最快的社交網站 (Wagner,2015)。但是，尚未有許多學術研究對其使用者進行行為分析，也並未深究其軟體本身與營運模式 (Sheldon and Bryant, 2016)。而台灣資策會創新應用服務研究所 FIND 團隊 (2017) 進行了國人社群網站使用行為的調查分析，發現擁有最多台灣用戶的軟體前五名為 Facebook（90.9%）、LINE（87.1%）、YouTube（60.4%）、PTT（37.8%）、Instagram（32.7%）；而每週造訪網站頻率依序則為 Line (86.7%)、Facebook (86.6%)、YouTube（44.8%）、PTT（34.6%）以及 Instagram（20.1%）。

然而這前五名的社群軟體之中，Line 主要功能為訊息聯絡，Youtube 則為影音傳播，兩者皆不主打發表文章的功能，並不適於本報告進行資料蒐集與分析。PTT 雖主打文章發布與討論，然而並不是主打個人版面上「情緒抒發」之文章發表，因此也不適於本告報所要進行的人格分析研究。Facebook 未提供可以獲取使用者文章內容的 API，技術上受限。而 Instagram 在資料獲取上有合適的 API 可使用，且其在創市際市場研究顧問公司社群網站使用報告 (2018) 中也顯示其台灣月活躍用戶數達到 740 萬人，約佔台灣 31% 人口，在年齡上使用人口介於 15～34 歲間便占了 6 成以上。綜合上述，本報告最終決定以 Instagram 用戶的文章內容進行分析研究。

伍. 方法架構



圖表 1 系統設計架構圖

1. 使用 Instagram Scraper[1] 程式爬取使用者 IG 文章資料。
2. 使用結巴斷詞對社群文章做斷詞，去除表情符號、英文字、數字、停止詞（如：代名詞、時間複詞、高頻詞。
3. 嘗試使用 likelihood 與 chi-square 兩種 feature selection 的方法，每種人格抽取 120 個特徵詞，最後整合成共 463 個非重複的詞當作模型的語料庫。
4. 選擇 Naïve Bayes 分類器進行模型訓練，使用問卷人格資料與社群媒體文章作作為訓練資料。以 F1 score 作為模型成效評估指標。
5. 使用名人在社群媒體上發布的文章，預測其人格

陸. 資料集

訓練資料

根據下表，由 78 份有效回收問卷中調查出來，我們分別測驗出 9 位開放性人格、22 位親和性人格、10 位外向性人格，以及 14 位勤勉正直性人格。而我們也發現在社群活躍度上，例如追蹤關係的建立、發布文章、按讚與留言，可親性人格皆為活躍度最高，勤勉正直性人格則皆較其餘三者低。

表格 1 各類人格資料統計表

人格	users	Follower	Following	Posts	Likes	Comments
	Count	Average	Average	Average	Average	Average
外向性	10	315	309	66	51	3.6
親和性	22	450	416	128	76	5.8
勤勉正直性	14	165	220	9	28	2.8
開放性	9	178	228	44	39	3.5

我們去蒐集了以上 78 位使用者的 Instagram 發文，作為我們的訓練資料，訓練模型找到不同人格使用者的用詞習慣，希望進而使其可以利用發布文章之用詞推測使用者所屬人格。

預測資料

我們蒐集了政治、娛樂、運動、文學、Youtuber 五大類領域中，各二十位名人的 Instagram 公開帳號，以裡面的貼文做為訓練後模型的測試資料。

柒. 資料處理方法

Extract Terms

首先針對每篇貼文過濾掉英文字、數字、標點符號與表情符號(emoji)，利用 JieBa 套件進行斷詞，並過濾掉 stopwords。由於斷詞結果仍然包含一些常見的、對分類沒有幫助的詞，因此又用人工方式將代名詞、時間複詞、高頻詞等等加進 stopwords。最終從 176355 個 tokens 中留下 34265 個 terms。

Feature Selection

針對每個分類選 top120 字詞，最終整合成 463 個字詞作為字典) 嘗試兩種方法 likelihood 與 chi-square，並以最終 validation 資料集的分類結果決定用哪種方法來建構模型。

[1] Instagram Scraper: https://github.com/rarcega/instagram-scraper

Likelihood

Likelihood 模型假定每個詞為二項性分佈，我們使用 -2log λ 分數高低，每個類別取前 120 個詞相加作為分類標準。

Chi square（ruby）

chi-square 是利用獨立假設去驗證兩者是否獨立、無關聯。故我所使用公式為把字典裡的每個字拿出來，字典數乘該詞在此分類的機率乘該詞在出現的機率若結果，若與字典數承該詞在此分類出現的機率相近，則代表這個詞與此類別為獨立集合，兩者無關係。以上敘述之公式所得之期望值，每個類別取前 120 個詞相加作為分類標準。

表格 2 各類人格高頻特徵詞

人格	高頻特徵詞
外向性	「不錯」、「同學」、「心情」、「成發」、「跳」、「合照」、「窩」、「哈哈哈哈」
親和性	「生活」、「跳」、「同學」、「盃」、「跳舞」、「陪」、「幫忙」、「當初」、「喝」、「搭」
勤勉正直性	「決定」、「直接」、「心情」、「室友」、「當初」、「慢慢」、「目標」、「願意」、「真正」
經驗開放性	「故事」、「新」、「歷史」、「城市」、「展」、「重新」、「一路」、「同學」、「小隊」

Naïve Bayes 分類模型

Naïve Bayes(NB)貝式分類器為透過聯合機率分布的計算來進行分類預測，將文本歸類於 MAP Probability 最高的類別。根據其機率計算方式不同而有 BernouliNB 與 MutinomialNB 兩種模型：只考慮字詞在文本中出現與否的 BernouliNB，以及考慮字詞在文本中的出現頻率的 MutinomialNB。而本報告使用 MutinomialNB 來建構分類預測模型。

Validation 準確度評估

將資料集以 8：2 分為 training data 以及 validation data。我們使用 F1 分數來作為模型評斷標準。我們每個類別接以八比二的比例分為 training data 跟 validation data，每個類別得到一個 F1 分數，最後將五類 F1 分數加總並平均，作為評斷模型之最後 F1 分數。

捌. 模型成效評估

Feature Selection 成效評估

表格 3 不同 feature selection 下的模型成效

Models	F1 score
NB+Chi-square	0.7595
NB+Likelihood	0.8375

以 NB 為基礎，分別以 Chi-square 與 Likelihood 作為 feature selection 所建構出的兩個模型，所得的 F1 score 分別為 0.7595 與 0.8375，因此最終選擇 NB+Likelyhood 來訓練分類模型。

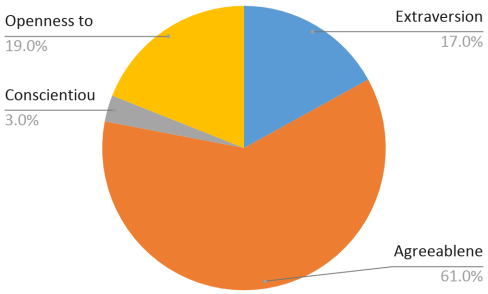
玖. 名人預測結果

使用上述建出的分類模型，對政治、娛樂、運動、文學、Youtuber 等領域名人傾向於在社群媒體展現的人格特質做預測，其中我們得到兩個結論：名人整體親和性個性預測較多、作家經驗開放性個性比較多。

category\人數	athlete	celebrity	politician	writer	youtuber
Extraversion	3	2	5	2	5
Agreeableness	14	15	13	5	14
Conscientiousness	2	0	0	1	0
Openness to Experience	1	3	2	12	1

圖表 2 五大類名人的預測人格分佈

結論一名人整體親和性個性預測較多



圖表 3 名人之預測四大人格分佈

從名人預測中所得人格以親和性預測為多，高頻用詞有「趕快」、「支持」、「很爽」、「嗨」、「打球」、「盃」，這些詞彙較多表現大方、營造親民形象的用詞，符合親和性中大方、樂於助人的特徵，符合我們對公眾人物的想像。

結論二：作家經驗開放性個性比較多

表格 4 各類名人與各人格的 Chi-Square 相關性計算

category 人數	athlete	celebrity	politician	writer	youtuber
外向性	0.05	0.58	0.75	0.58	0.75
親和性	0.27	0.64	0.05	4.25	0.27
勤勉正直性	3.27	0.60	0.60	0.27	0.60
經驗開放性	2.06	0.17	0.85	17.69	2.06

我們使用 Chi-square 計算發現，作家類別名人與經驗開放性個性有顯著性相關，細部探究會此類人格的高頻詞發現，有較多體驗、經驗的描述的詞，如：「城市」、「美國」、「重新」、「地鐵」、「歷史」、「山」、「教育」，符合經驗開放性中對知識好奇的特徵，也符合我們對作家的想像。

壹拾. 結論

在本報告中，我們建立了社群軟體使用者的人格預估模型，並使用五項不同領域中各二十位名人於 Instagram 發表之文章預測其在社群媒體傾向於展現出五大人格特質中的哪一類人格特質。本報告證實了某些領域中人士常見的共同性格特質。

壹拾壹.技術改進

受限於資料量的取得，本報告在建模後僅進行了簡要的訓練與預測，旨在提供對於模型的想法，並期望在可以取得更大量資料的情況下可以提高準確率、加強各人格差異區分，使模型功能更趨完整。

壹拾貳.未來應用

利用使用者於社群軟體上發表之文章內容進行人格分析、區分各用戶個性差異的概念，除了可以設計類似心理測驗的應用軟體，達到娛樂效果以外，也可推薦軟體後台應用於廣告投放等相關行銷策略之中，預測每位使用者之人格特質，並對其投放相應偏好的廣告商品內容，以期在不因置入廣告造成使用者排斥感的同時，甚至可能提升廣告點擊率。

壹拾參.參考文獻

Allport, G. W., & Odbert, H. S. (1936). Trait-names: A psycho-lexical study. *Psychological Monographs, 47*(1), i-171. https://doi.org/10.1037/h0093360

Costa, P. T., & McCrae, R. R. (1986). Cross-sectional studies of personality in a national sample: I. Development and validation of

survey measures. *Psychology and Aging, 1*(2), 140-143. https://doi.org/10.1037/0882-7974.1.2.140

Dutta, K., Singh, V. K., Chakraborty, P., Sidhardhan, S. K., Krishna, B. S., & Dash, C. (2017). Analyzing Big-Five personality traits of Indian celebrities using online social media. *Psychological Studies, 62*(2), 113-124. doi: https://doi.org/10.1007/s12646-017-0408-8

Essays, UK. (2018). *Impact of Instagram on Social Networks*. Retrieved from https://www.ukessays.com/essays/sociology/proposal-how-instagram-makes-us-sociable.php?vref=1

Galton, F. (1884). Measurement of character. *Fortnightly Review, 36*, 179-185. Retrieved from http://galton.org/essays/1880-1889/galton-1884-fort-rev-measurement-character.pdf

Jones, G., George, J. (2019). Values, attitudes, emotions, and culture: the manager as a person. In *Essentials of Contemporary Management* (pp.53). New York, NY: McGraw-Hill Education.

Kircaburun, K., & Griffiths, M. D. (2018). Instagram addiction and the Big Five of personality: The mediating role of self-liking. *Journal of Behavioral Addictions, 7*(1), 158-170.Retrieved from https://doi.org/10.1556/2006.7.2018.15

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The Journal of Abnormal and Social Psychology, 66*(6), 574-583. https://doi.org/10.1037/h0040291

Sheldon, P., Bryant, K. (2016). Instagram: Motives for its use and relationship to narcissism and contextual age. *Computers in Human Behavior, 58*, 89-97, doi: https://doi.org/10.1016/j.chb.2015.12.059

Wagner, W., & Mathison, P. (2015). Connecting to communities: Powerful pedagogies for leading for social change. *Innovative Learning for Leadership Development. New Directions for Student Leadership, 145*, 85-96. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1002/yd.20126

Zhang. (2002). Thinking styles and the big five personality traits. *Educational Psychology, 22*(1), 17-31. doi: https://doi.org/10.1080/01443410120101224

蘇怡文（民國 106 年 5 月 1 日）。**八成以上台灣人愛用 Facebook、Line 坐穩社群網站龍頭 1 人平均擁 4 個社群帳號　年輕人更愛 YouTube 和 IG** 部落格文字資料。取自 https://www.iii.org.tw/Press/NewsDtl.aspx?nsp_sqno=1934&fm_sqno=14&fbclid=IwAR0iwJU_OkutnE8W1MbJMd5tN832zerGpVdW6qSXbZBpULUSi2X-Cktvhxs

蘇思云（民國 107 年 10 月 2 日）。**台灣 IG 月活躍用戶數已近三分之一人口！亞瑞特數位社群行銷執行長黃逸旻：用這 5 招提高粉絲數【部落格文字資料】**。取自 https://www.cheers.com.tw/article/article.action?id=5092097