運用NLP領域中多種特徵提取工程與分類模型進行多重組合分析並建置簡易真假新聞判斷系統

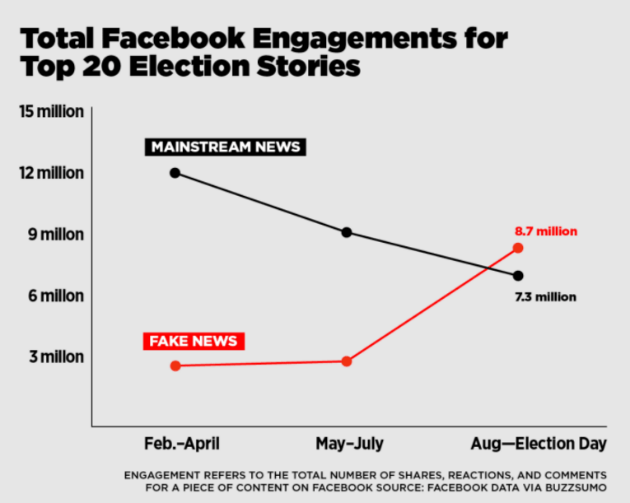
指導教授：葉向原

學生：巨資四A黃奕鈞、巨資四B李冠暶

1. **Overviews**

台灣正式進入2019尾聲，面對即將來臨的2020總統大選，選民們，準備好了嗎？

西元2016年，美國總統大選，川普大爆冷門當選美國總統，其中，假新聞被認為是川普當選的一大推手。38歲的保羅霍納(Paul Horner)就是位假新聞寫手，他說：「現在根本沒有人在確認新聞真假」。為什麼要製造假新聞呢？因為假新聞愈寫愈多，每一則都帶來大筆流量，也增加進帳。華盛頓郵報記者凱特琳杜威：「假新聞問題的癥結，還有保羅霍納如此成功斂財的原因，在於所謂的確認偏差，也就是大家想要看和分享的新聞，是能夠符合他們原本認知的新聞。」假新聞竄起的一大原因就是主流媒體大多有政治立場，選邊站，選民不再信任。



圖(1)：據BuzzFeed分析，2016大選後期主流媒體與假消息。

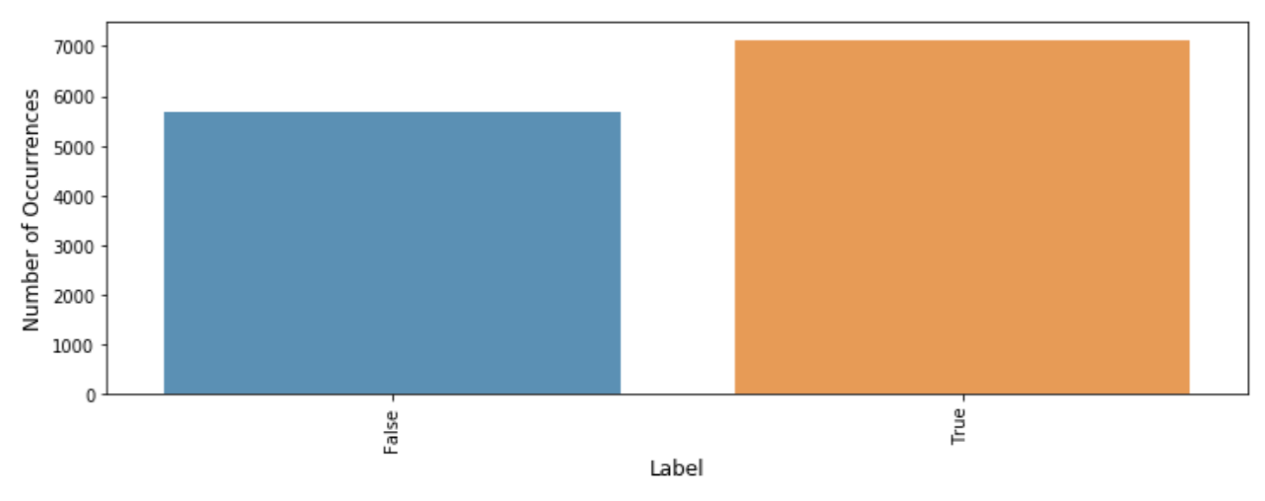
在台灣，主流新聞媒體大多也存在政治立場，而中國與台灣的兩岸關係更是各位媒體識讀人最關心的議題之一。2019年6月23日，由「館長」陳之漢及時代力量立法委員黃國昌號召遊行，又稱「623凱道集結」，遊行目的是反對中國共產黨、拒絕親陸媒體（又稱「紅色媒體」）滲透、守護台灣的民主與自由，同時聲援香港反對《逃犯條例》修訂草案的推動。其中，由蔡衍明主持的旺旺中時媒體集團、及其旗下的《中國時報》報系、中視與中天電視報導，被認為是親陸的立場。

「網軍」，一個網路用語，多被台灣網友使用，大致可以歸納出兩種不同的定義，第一種，是受僱於特定政治背景的個人或組織，並為其刺探網路情報、輿論顛覆、帶動輿論風向者。第二種，則是受僱於特定的民間企業，透過各種方式行銷、推廣特定人、事、物者。而在2020總統大選更是被多次指控網路上充斥著發布假消息、帶風向的網軍，面對種種的消息，選民又該如何培養正確的媒體識讀能力呢？

本篇研究運用自然語言處理(NLP)，透過不同的文字向量以及分類模型，找出最佳的組合並對新聞描述進行真假預測。

1. **Research methodology**

本次所使用真假新聞資料集來自LIAR資料集中「LIAR: A BENCHMARK DATASET FOR FAKE NEWS DETECTION」共12791筆資料，包含真新聞資料7134筆、假新聞資料5657筆。資料散佈呈現於圖(2)。



圖(2)：本次使用資料包含7134筆真新聞、5657筆假新聞。

首先，我們先對新聞描述做資料預處理，包含：

（一）Remove stop words：為節省存儲空間和提高搜索效率，在處理自然語言數據之前或之後會自動過濾掉某些字或詞，這些字或詞即被稱為Stop Words(停用詞)。停用詞大致分為兩類。一類是人類語言中包含的功能詞，這些功能詞極其普遍，與其他詞相比，功能詞沒有什麼實際含義，比如’the’、’is’、’on’等。另一類詞包括詞彙詞，比如’want’等，這些詞應用十分廣泛，但是對這樣的詞搜尋引擎無法保證能夠給出真正相關的搜索結果，難以幫助所小搜索範圍，同時還會降低搜索的效率，所以通常會把這些詞從問題中移去，從而提高搜索性能。本次研究使用NLTK的Stop Words資料集進行移除停用詞。

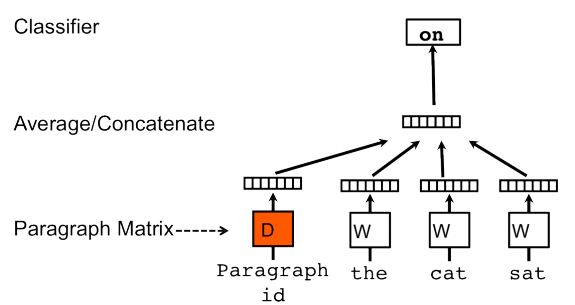
（二）Stemming：在英語中，一個單詞常常是另一個單詞的’變種’，如：happy=>happiness，這裡happy叫做happiness的詞幹(stem)。提取詞幹(stemming)，即除去英文單詞分詞變換形式的結尾。本次研究使用NLTK的Snowball函數庫進行詞幹提取(stemming)。

其次，將處理過後的資料進行特徵提取，此次研究所使用的特徵提取有以下三種方法：

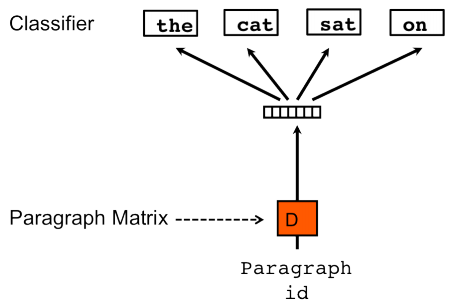
（一）Sklearn中的CountVectorizer和TFIDFTransformer：CountVectorizer(詞頻向量化)會將文本中的詞語轉換為詞頻矩陣，例如矩陣中包含一個元素a[i][j]，它表示j詞在i類文本下的詞頻。TF-IDF是一種用於資訊檢索與文本挖掘的常用加權技術。用以評估一個字詞對於一個文件集或一個語料庫中的其中一份文件的重要程度。字詞的重要性隨著它在文件中出現的次數成正比增加，但同時會隨著它在語料庫中出現的頻率成反比下滑。TF-IDF的主要思想是：如果某個詞或短語在一篇文章中出現的頻率TF高，並且在其他文章中很少出現，則認為此詞或者短語具有很好的類別區分能力，適合用來分類。

（二）Sklearn中的TFIDF\_N-Gram：N-Gram的基本思想是將文本裡面的內容按照字節進行大小為N的滑動窗口操作，形成了長度是N的字節片段序列。每一個字節片段稱為gram，第N個詞的出現只與前面N-1個詞相關，而與其他任何詞都不相關，整句的概率就是各個詞出現概率的乘積。這些概率可以通過直接從語料中統計N個詞同時出現的次數得到。

（三）Gensim中的Doc2Vec：Doc2Vec分兩個模型PV-DM與DBOW



PV-DM模型



DBOW模型

第三，針對真假新聞進行分類，本次研究使用的分類模型有以下五種：

（一）Naive bayes。

（二）Logistic Regression。

（三）SVM。

（四）SGD。

（五）Random Forest。

第四，將三種特徵提取方法與五種分類模型進行多重比較，用F1-Score評估各項組合優劣，篩選出分數最高的模型作後續的真假新聞系統建置。

最後，製作簡易的真假新聞判斷系統，輸入想要判斷的新聞描述，系統會自動判斷描述是真是假，以及提供F1-Score參考。

1. **Results**

本次研究結果呈現於表(1)，此次評估模型好壞運用指標為F1-Score，表中可看到最成功的組合為：TFIDF\_N-gram，其Testing F1-Score達到0.613。

|  |  |  |  |
| --- | --- | --- | --- |
| F1-Score | CountV+TFIDF | TFIDF\_N-gram | Doc2Vec |
| Naïve bayes | 0.580 | 0.580 | NULL |
| Logistic Regression | 0.607 | 0.607 | 0.541 |
| SVM | 0.583 | 0.583 | 0.543 |
| SGD | 0.507 | 0.507 | 0.425 |
| Random Forest | 0.613 | 0.613 | 0.539 |

表(1)特徵提取與分類模型多重組合分數比較

1. **Conclusion**

隨著台灣2020總統大選將至，本次研究提供各位媒體識讀人一套簡易的真假新聞辨識系統，可將獲得的新聞資訊輸入至系統，系統則會判斷此新聞資訊是否為真，以及提供背後運算的統計數據。

本次研究將NLP領域中流行的多種特徵提取工程以及分類模型整合並做多重組合比較分析，並透過最優秀的組合建置一套簡易的真假新聞判斷系統。

本次研究

1. **References**

Github假新聞範例草考<https://github.com/tychen5/NLP_FakeNewsDetection>

Github假新聞範例參考<https://github.com/nishitpatel01/Fake_News_Detection>

假新聞之王報導<https://buzzorange.com/techorange/2016/11/29/fake-news-makes-isa-president/>

圖(1)來源<https://www.buzzfeed.com>

A list of contractions from <http://stackoverflow.com/questions/19790188/expanding-english-language-contractions-in-python>

623凱道集結https://zh.wikipedia.org/wiki/「拒絕紅色媒體、守護臺灣民主」遊行

Dov2Vec模型示意圖<https://medium.com/@ddoo8059/讀paper之心得-word2vec-與-doc2vec-5c8b8daa7f12>