運用機器語言模型預測與分析中文媒體之報導框架

姓名：王修佑

學校：國立政治大學附屬高級中學

摘要

隨著科技日新月異，新聞傳播媒體對現今社會造成的影響日漸增加，身為讀者如何在獲取大量資訊的同時具有媒體識讀的能力成為相當重要的課題；民主社會中每一位媒體都擁有發表自己的政治框架(political framing)或是不同的主觀框架(frame)立場的權利，大眾傳播領域中的「皮下注射理論」(Hypodermic needle model)提到當這些特定框架立場資訊透過媒體的傳播給「受眾」後，閱聽人經常會無條件地吸收媒體所要傳達的想法與資訊；而現今媒體為追求報導閱覽量，因此經常使用片面的框架撰寫報導來吸引讀者的矚目，在這樣長期不斷接收所有資訊的暴露之下，不少人受到「框架立場」、「不實訊息」、「文字情感」等因素淺移默化的影響，無法獲得完整正確的報導，造成價值觀偏頗、無法客觀處事、獲取與散播錯誤不實訊息，甚至可能造成社會恐慌。本次研究希望透過人工智慧自然語言處理技術(NLP, Natural Language Processing)以及多種不同語言模型，並且以近年重要議題的網路新聞作為學習文本，在比較出多種模型與最佳學習方法後，使機器具備可預測新聞報導之媒體框架的能力，並且進一步分析各媒體針對不同報導的切入方向夠不夠全面，將有效輔助讀者選擇較具公信力的媒體，吸收資訊時可以更快速、更完整正確的了解該議題全面事實真相，避免被媒體框架、政黨、他國政治因素干預或是具偏頗政治立場的媒體報導所影響、誤導。

一、前言

(一) 研究動機

當今天如果沒有社群媒體 LINE、Twitter、Instagram、Facebook，我們眼前的世界可說是一片黑暗，這時新聞媒體就是我們在世界各地角落的眼與耳，控制我們獲取的資訊；在民主自由的國家社會中，媒體有言論的權力與自由，但如果這些媒體為追求營利收入而用片面框架報導，讀者們又該如何看到每一個事件的全貌與真相呢？過去台灣也曾經是一個新聞媒體被獨裁政府綁架的社會，2003 年國家立法讓官方黨政軍勢力退出媒體，

轉而接手新聞媒體產業的變成國內各大財團與資本家，這些財團透過經營媒體傳達自身立場的同時時常運用片面的框架陳述議題，以吸引更多讀者點閱來平衡廣告收入；我們平時使用網路搜尋資料、閱聽新聞時，往往會發現在相同議題上，各家的媒體都有各自不同的陳述方式，多元面向的新聞框架可以將事實呈現，而那些不好的框架往往是將立場放在事實之前，一方面吸引讀者矚目，一面在自身支持的觀點上畫重點；過去中國央視在反送中事件上一系列的報導中，一味宣揚港警的英雄形象，不報導港民真正的訴求，並且使用各種手段塑造上街示威的香港民眾是社會中少數的暴力分子，進而使人民跟政府站在同一邊，拒絕政府以外不同的聲音。而台灣為了落實民主社會，在 1998 年依《公共電視法》成立台灣唯一公營媒體—公視，所謂公營媒體並非「官媒」，而是由國家人民納稅共同出資經營的媒體，具有完全的經營獨立性，保障不受外在勢力侵擾、干預報導內容，在同為公共媒體的英國 BBC 網站 Mission, values and public purposes 中很清楚可以看到這些公共媒體都以「為社會與人民向國際發聲並提供最理性公正的報導」為己任。

（二）問題與目的

從上述香港的事件經驗中，我們可以發現當社會中充斥著單一面向「選擇性曝光」的框架報導，不再有多元聲音時，錯誤的框架報導可能葬送一整個民主制度，顯示在媒體立場被控制時很可能造成國安危機。雖然有了公共媒體，但這時我們不禁想問公共媒體又真有如此公正嗎？在網路上的搜尋、瀏覽新聞報導往往在閱讀全文前只能看到記者經過選擇與重組的標題，即便直接閱讀全文，時常也很難讓一般讀者一眼看出報導中的框架立場去選擇適當的報導閱讀，為避免流竄在社會中不好的框架報導對民眾造成不良的影響，就要提升讀者的媒體識讀能力，因此本次研究希望能訓練出可增進、輔助讀者識讀媒體報導框架的機器模型，並比較不同模型與演算法所訓練出的成果差異，探討對於此研究領域最佳的機器訓練工具與方法，期望使機器精準預測報導中的框架後，進一步分析公營媒體、私營媒體的報導框架切入層面是否真有相異，增加此機器的務實可能性，也使讀者能更正確選擇適當的媒體。

三、研究方法與過程

（一）資料集(Datasets)

本研究將以近年國內中文主流網路與平面新聞媒體的報導，分別為自由電子報、中時電子報、ET Today 新聞雲、中央社、華視新聞網、民視新聞網等六家媒體，資料將鎖定「COVID-19 疫情」、「香港反送中」、「台灣多元成家」等熱門議題，並透過爬蟲(web

crawler)等方式擴充現有可用的資料集，共約四萬筆資料作為研究中使用。取得資料後進行資料前處理(Data pre-processing)，包括去除雜訊、整理純文字文件、轉檔等步驟，以方便程式讀取資料建構本研究之基礎。將使用中央研究院開發的中文斷詞系統—monpa斷詞系統，處理「能夠獨立運用，具有完整語意的最小語言成分」；資料文本將以「報導」為單位進行標記(label)，針對報導內容將框架將分類為：政治-民進黨、政治-國民黨、政治-無、經濟、民生等，進行標記。

（二）建構模型

本研究使用監督式學習領域方法，並且在實驗中比較多種分類器與模型對於分析中文新聞媒體報導框架成效之差異，將使用的分類器與方法包括：隨機森林分類器(RFC, Random Forest Classifier)、決策樹分類器(DTC, Decision Tree Classifier)、支援向量機(SVM, Support Vector Machine)、單純貝氏分類器(Naive Bayes classifier)中的高斯貝氏分類器(GNB, Gaussian Naive Bayes)、多項式貝氏分類器(MNB, Multinomial Naive Bayes)、伯努利貝氏分類器(BNB, Bernoulli Naive Bayes)、邏輯斯迴歸(Logistic regression)；將嘗試活用兩種機器語言模型：BERT (Bidirectional Encoder Representations from Transformers)、ELMo (Embeddings from Language Models)，進行遷移學習，試圖找出對於此學習資料相對最具成效的工具。研究初期將先使用現有 30,000 筆資料，針對此資料集以標籤完成的新聞分類(已完成的現有標籤為新聞網中現有的分類，例如：科技、教育、政治等)建構基本的模型；預計透過利用 Tensorflow 等開源軟體工具，作為本次研究的開發環境。

（三）效能評估

依據訓練結果計算準確率(Precision)和召回率(Recall)數據，從其調和平均數 F1-score (F-measure)和 ROC 曲線(Receiver Operating Characteristic curve)數據圖形樣貌呈現，進行使用不同模型訓練下成效之客觀評估比較，也將使用 Scikit-learn 來實現研究中多種訓練模型成效差異之比較，評估與探討較佳的學習模型作為可行的務實成果目標，整體流程期望以 Tensorflow 運用 Numpy 函式庫進行語言模型訓練的基礎想法，並嘗試結合使用較高階的神經網路 API，Keras 作為訓練工具之一，且在其中運用長短期記憶模型(LSTM, Long short-term memory)與遞迴神經網路(RNN, Recurrent neural network)的核心概念進行研究，進一步綜合性探討、尋找對分析與預測新聞媒體報導框架資料學習的最佳分類器與模型方法。最後，透過研究中最佳務實模型與分類器針對自由電子報、中時電子報、ET Today 新聞雲與公共電視面對「COVID-19 疫情」、「香港反送中」、「台灣多元成

家」新聞議題的報導進行框架預測，取得的預測結果後近一步整理、分析，製成圖表後希望嘗試從中找出不同性質的媒體在報導相同議題時是否有從不同面向、多元的框架角度切入報導，使受眾能公正的了解議題的全貌，進而輔助受眾能將機器預測結果作為自身參考的指標之一，以更快選擇適合自己或是值得信任的媒體，務實語言機器的輔助工作。

四、未來展望

現今社會中紙本平面刊物沒落下，各電子媒體為保有固定客源因此都有電子報的訂閱功能，這造成受眾經常閱讀單一媒體製作的議題報導，而在媒體使用不夠全面的框架報導下被誤導、無法看清事件的全貌，未來持續透過結合自然語言技術的情感分析、假新聞辨識、框架預測等功能，將可使臺灣在新的電子新聞媒體時代中，輔助每一位受眾都可以具有更好的媒體識讀能力，取用(access)新聞的同時具有更完善的分析(analyze)與評估(evaluate)行為，做好民眾監督媒體，媒體關心社會、監督政府的互利共榮媒體生態。

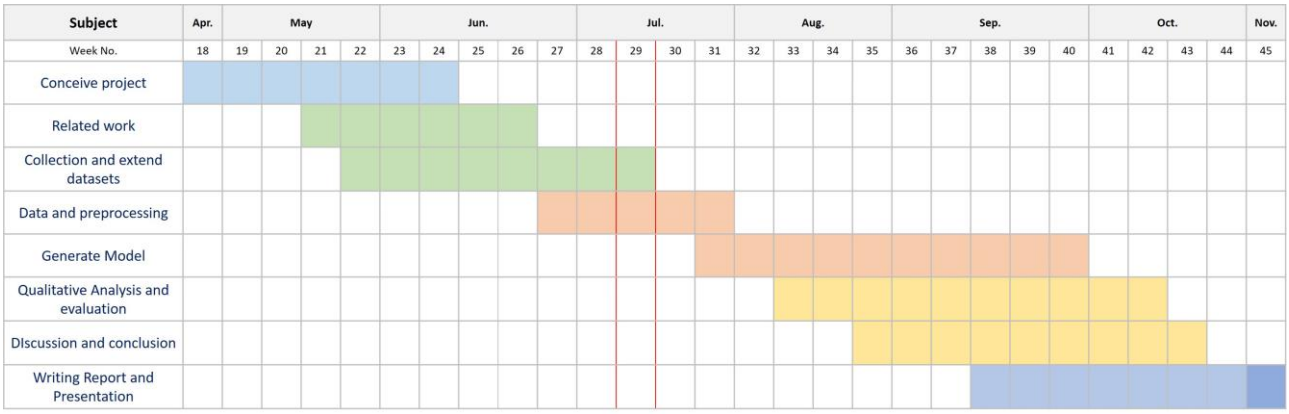

圖 1：研究計畫時程