以機器學習與規則方法辨識中文民事裁判書結構
Using Machine Learning and Pattern-Based Methods for Identifying Elements in Chinese Judgment Documents of Civil Cases

林泓任 劉威志 劉昭麟 楊婕
Hong-Ren Lin Wei-Zhi Liu Chao-Lin Liu Chieh Yang
國立政治大學資訊科學系
Department of Computer Science, National Chengchi University
{109753156, 109753157, chaolin}@g.nccu.edu.tw, 05141343@gm.scu.edu.tw

摘要

在建構法學資訊相關的分類模型或是推薦系統時，提供模型關於語料中與任務相關的輔助訊息有助於提升模型的效能及運算結果的可解釋性。在本研究中，我們選擇以人工智慧應用於法學資訊中較少見的民事訴訟案件作為研究對象。我們將「給付扶養費」相關的判決書依照其結構的功能性，找出聲請人所提出的主張、相對人所提出的答辯、法院對案件的見解及法院所引用的法條這四個部分，並從中抽取出該案件訴訟兩造的主要爭點。

Abstract

Providing structural information about civil cases for judgement prediction systems or recommendation systems can enhance the efficiency of the inference procedures and the justifiability of produced results. In this research, we focus on the civil cases about alimony, which is a relatively uncommon choice in current applications of artificial intelligence in law. We attempt to identify the statements for four types of legal functions in judgement documents, i.e., the pleadings of the applicants, the responses of the opposite parties, the opinions of the courts, and uses of laws to reach the final decisions. In addition, we also try to identify the conflicting issues between the plaintiffs and the defendants in the judgement documents.

關鍵字：民事案件、給付扶養費、文章結構分類

Keywords: civil cases, the issues of alimony, legal element identification

1 緒論

要讓使用者能接受演算法所提供的推薦項目，提供推薦理由是很重要的一件事。Mumford et al. (2021) 曾在研究中提到 *"without an explanation of why the case was so classified, the adjudicator has no reason to follow."*，意即如果無法說明案件結果是如何被分類出來，那法官也沒有理由去採信模型所提出來的決策。為了讓電腦產生的案件預測結果能被使用者接受或是提升電腦的預測能力，建立讓電腦理解及能夠解釋法律文件細節的核心能力是有必要的。

因此，我們的研究主旨在提供法律文件中各段落、句子在裁判書中的作用。我們選擇將裁判書內文依照文章結構的功能性分類為四個大類，分別是： 1.聲請人提出的主張、2. 相對人提出的答辯、3. 法院在判決中所引用的法條、4.法院對該案件的見解與裁判，在往後的文章中會簡稱為 C1、C2、C3、C4。除了將裁判書依其文章結構中的功能性分出四大類外我們也希望能找出案情的爭議點，在完成四大類後我們會找出裁判書中的爭點，其為法院對於案情爭點的說明以及裁判，並將這類型的句子簡稱為 C5。

不過這些標籤在我們所取得的裁判書資料中是不存在的，因此本研究中會先建構一套使用規則與正規表示法(regular expressions)的技術來對特定範圍內的裁判書以段落為單位達成初步的自動標記，並透過段落的自動標記轉出換成對句子的預標記，以此訓練模型將C1、C2、C3、C4的標記推廣到其他無法直接使用規則標記方法的段落上。如此一來可以針對各別部分建立推薦系統使推薦系統更

有說服力，也可應用於裁判結果預測或其他法學資訊的相關應用研究上。

使用以規則自動標記的預標記建立對裁判書文章架構的分類系統後，我們會透過藉由法學領域專家的標記來驗證我們的分類系統成效，並對比以專家人工標記訓練的分類模型與自動標記系統建立的分類模型對前述四分類任務的差距。

完成C1、C2、C3、C4四分類系統後，我們將文章架構分類器推廣到第5個分類，也就是爭點的部分。此分類法官會描述聲請人與相對人所提出的主張中的爭議點，並加以說明、裁判。其都存在於C4類別中，但C5中會描述到聲請人或相對人所提出的主張，因此也容易與C1、C2的分類混淆。

後續內容會介紹我們針對台灣民事案件裁判書所做的前處理、如何切割段落及斷句，並說明如何轉換成我們模型的輸入資料並探討以不同前後短句句數訓練的成效差異及若將模型加深、複雜化能對分類效果帶來的提升；之後以一開始針對處理的裁判書類型以外的裁判書搭配專家人工標記的結果驗證模型。確認完能在不同段落數的裁判書生效後將分類工作推廣到更進一步的案件爭點分類、搜尋工作上。此外我們也會比較一些既有模型設定上的差異以及基於前面C1、C2這種易與C5混淆段落的驗證。

2 相關研究

過往，將人工智慧技術應用於法學資訊在國外已經有很長的一段歷史，最早的研討會可追溯到1987年 (Bench-Capon et al., 2012) 。而常見的法學資訊相關研究可分為判決書結果及刑度預測、類似案件推薦、法律問答系統等 (Zhong et al., 2020) 。

其中， Lin et al. (2012) 的研究中透過 CRF (conditional-random field)模型對刑事案件判決書中的量刑因子進行自動標記並應用於判決結果及刑度的預測上。建構命名實體辨識(NER, named-entity recognition)系統時，根據Chen et al.(2020)的研究指出，提供模型相關判決書額外的法律資訊也是有助於提升模型的。

綜合 Zhong et al. (2020)的研究整理及先前提及的研究，為了讓電腦產生的案件預測結果能被使用者接受或是提升電腦的預測能力，建立讓電腦理解及能夠解釋法律文件細節的能力是有必要的。因此在 Chalkidis et al. (2021)的研究中，透過已經標註好段落與法院判決間的關係的資料，讓模型抽取到與法院作出判斷相關聯的段落時就給予額外的獎勵來建構可解釋性的法律預測模型。

但我們所使用來自於台灣司法院開放平台[1]所公開的裁判書資料，其並不具有這些額外的法律相關資訊的標記。而且在民事案件的裁判書中，也不像刑事案件有較為明確的量刑因子與法律用語，台灣相關研究的數量也較為少見，不過近期也有針對將 AI 技術應用於民事案件可行性的相關研究 (Ho, 2021) 。

因此我們又參考了 de Buy Wenniger et al. (2020) 的研究，其研究中讓文章輸入模型時將句子附加上依照其在文章節中的功能所給予的標記提升了預測成效；說明了文章結構影響分類結果，也啟發我們開始從各個段落及文句在裁判書中所具有的功能性開始著手。而 Li et al. (2019) 的研究中，將中國的刑事訴訟裁判書分成三個部分，分別為描述被告過往是否有其他犯罪紀錄的段落、該次案件的犯罪事實、法院對於案件做出的裁定，並透過注意力機制將關於被告犯罪紀錄的描述以及該次案件犯罪事實結合起來得到了能更好預測法院裁定結果的模型。

因此我們將初步目標放在區分出民事案件中的段落及句子在文章結構上的功能性，依此建構一套透過規則與正規表示法來自動找出裁判書中的C1、C2、C3、C4標記並以此為基礎訓練以句為單位的分類模型，並驗證這個方法的可行性。

在 C5，也就是爭點類別的搜尋中，我們參考了 Xu et al. (2021) 所提出的研究。他們其中一項研究是將案件全文中的句子分類為IRC與non IRC兩種，其中 IRC 的部分包含有 I，是法院在該案件中所要裁決的問題、還有C是法院對I所做出的裁決，而最後的R是法院說明如何得出結論的句子；此一部分恰好為我們緒論中提到要尋找的爭點與爭點的說明，也就是被我們所定義為 C5 這個類別的句子，因

[1] https://opendata.judicial.gov.tw

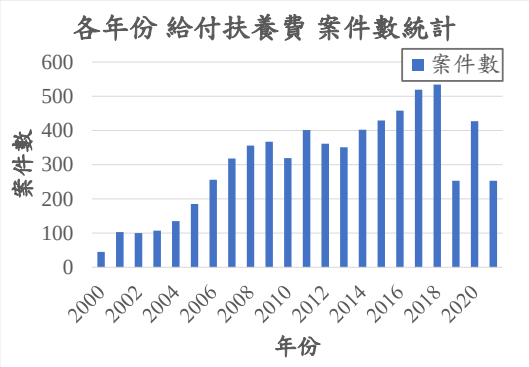

圖 1. 各年分給付扶養費案件數統計

此我們同樣會使用分類模型的方式來嘗試找到我們所定義的 C5 類別。

3 問題定義與假設

我們研究的目標是處理台灣司法院開放平台上關於民事訴訟，且案由為給付扶養費的案件，將裁判書中的句子及段落依照其在文章結構中的功能性區分出 C1、C2、C3、C4 這四種類別，因此我們嘗試將其視作針對裁判書中各句子的分類任務。

除了上面四種類別外，我們也希望能找出案件的爭點，與 Xu et al. (2021) 提出研究不同的是，他們將在整篇裁判書中挑出少量的 IRC 段落並將其他段落視為 non-IRC，而在我們的觀察中，這類段落會包含在 C4 類別的句子之中，因此我們透過 5 分類的模型從裁判書中找出本次案件的 C5 類別，也就是爭點，並且同時給予模型目標句子的前後句做為參考。

在我們所定義的 C5 相關的句子中會包含法官對爭點的說明，也就是聲請人與相對人所提出的事證中有所衝突的部分，因此這類句子會與前述聲請人提出的主張或相對人的答辯容易產生混淆，所以我們也會嘗試從 C4 類別中被分類器錯誤分類為 C1、C2 的句子中找出 C5。

4 資料前處理

我們所使用的資料主要來自於台灣的司法院開放平台上所公開，擷取其中自 2000 年到 2021 年的裁判書。因台灣司法院開放平台網站每月會更新公開的裁判書，所以相關文件總數會是浮動的數值。自 2000 年至 2021 年份間，案由為給付扶養費的裁判書數量統計如圖 1，21 年間總數共 6679 篇裁判書。

4.1 語料挑選

台灣司法院對於所公開的裁判書，在撰寫上並沒有嚴格限制格式。因此我們會透過一些規則及正規表示法來濾除多餘的部分取出我們所需要的段落。

裁判書中裁判字號、日期、案由會在固定的位置，可以藉由判斷開頭行數關鍵字對其進行整理。其中，我們透過找出裁判書上的案由並篩選出其案由為給付扶養費的案件。除此之外，還有許多單純筆錄、上訴的案件不會包含緒論中所定義的四項結構標簽，因此在前處理中也會將這方面的裁判書濾除。

4.2 語料清理

篩選完給付扶養費相關裁判書後，我們會進一步清理裁判書的內文。在裁判書內文中除了裁判字號、日期、案由、聲請人和相對人姓名外，便是裁判書內文。

以我們主要研究對象，案由為給付扶養費的案件為例；內文會包含有主文及理由 (或是事實、事實與理由) 段落。其中主文是紀錄法院最後的裁判結果，而事實與理由段落則會包含有聲請人在訴訟中提出的聲明、相對人所提出的抗辯、還有法院對裁判結果的說明及所引用的法條，也就是緒論中所定義的 C1、C2、C3、C4 四種分類。

因此我們主要分析的語料便是裁判書中的事實與理由段落。我們會透過前面裁判書固定段落中的聲請人、相對人姓名做紀錄並屏蔽掉事實與理由段落中的聲請人、相對人姓名部分。之後透過正規表示法找出裁判書中數字相關的用語，如金額、年分、生日，將其取代為某金額、某年、某月這種較為模糊的用語。除此之外，我們也會藉由中研院開源的工具 CKIP 對初步處理過後的文句以其中的 NER 工具找出各段落的人名、地名、組織部分，將其替換為某人 1、某人 2 這種形式後才應用於後續分類任務。

4.3 段落裁切與斷句

進一步分析後發現，裁判書中會固定使用幾種章節標號來區分段落，我們整理其使用的章節標號用來切割裁判書大段落的依據。

壹、, 貳、, 參、, 肆、…
㈠, ㈡, ㈢, ㈣, ㈤ …
（一）, （二）, （三）, （四）, …
一、, 二、, 三、, 四、…
㊀, ㊁, ㊂, ㊃…

表 1. 裁判書中常見章節符號

透過如上表 1 所示的章節符號切割出大段落後，圖 2 是我們統計從 2000 年至 2021 年底為止，案由為給付扶養費的裁判書，各種不同大段落數所包含的裁判書數量。此處的裁判書並未濾除語料清理段落所說不合要求的裁判書。

除了切割大段落外，我們藉由「，、；。：」五種標點符號作為斷句，將大段落切割成數句短句，並保留每句短句的標點符號。其短句長度分布區間如圖 3 所示，大部分的短句長度會在 32 個字以內。

5 實驗設計

在我們的實驗中，挑選前述語料中大段落數為 4 或 5 的 1629 篇裁判書作為初步的研究對象，因我們觀察各種不同段落數的裁判書後發現在這兩種大段落數的裁判書，撰寫上是較有規律的。透過語料清理段落所提到的清理過程後剩餘 814 篇裁判書。

5.1 資料標記

我們分別使用兩種方法來對資料進行標記。在初期實驗，我們並不具有任何語料的標記，因此藉由規則與正規表示法來做為針對特定段落數裁判書自動標記的手段。

以前處理段落所述的大段落來做為分界，首先，C4 的段落可能出現在裁判書事實與理由段落的開頭，而且在 C4 的段落中可以同時找到“(本|法)院”及“(判斷|心證|據)”這些強烈暗示該段落為法院用語的詞組。

而描述聲請人主張與相對人抗辯內容的 C1、C2 段落，C1 的段落總是會安排在 C2 的段落之前，且開頭會分別提到“聲請人聲起略以”、“相對人答辯略以”這類直接提及聲請人及相對人的詞彙，可藉此區分出 C1、C2 的段落。

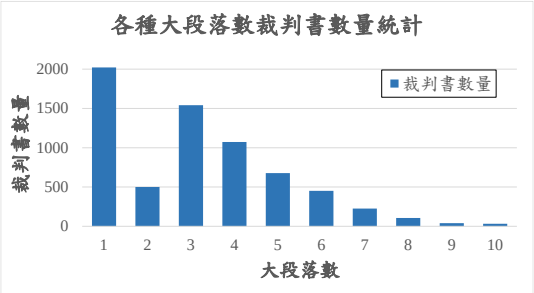

圖 2. 各種不同大段落數裁判書數量統計

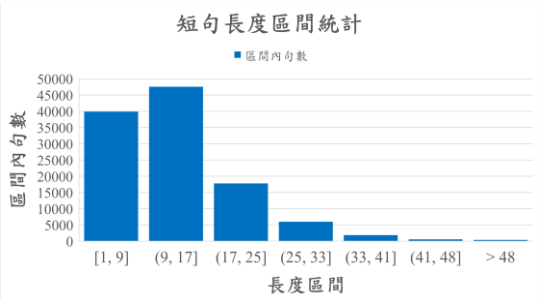

圖 3. 短句長度區間統計

當法官要描述所引用的法條，也就是 C3 的段落，會以“按”作為開頭。但 C3 的段落可能混雜在 C4 中的其中一個小段落中，因此需要針對 C4、C3 的小段落額外去做排除與區別。這種依靠規則與正規表示法對裁判書中段落進行標記的方式後續會簡稱為規則標記。在得到大段落的規則標記資料後，我們會藉由前面段落裁切與斷句章節中所提到的方法，將大段落中的文字切割為短句，並將每句短句視為與該大段落相同的標籤做為規則標記資料使用。

除了上述規則標記的方法外，我們也聘請法學院畢業的專任助理對這些裁判書的句子進行標記。這些經過專家人工判斷與標記後的方式，我們往後會簡稱為人工標記。

5.2 分類模型與參數設定

本次研究主軸並非針對已有的資料集提出更進一步的改善，而是提出一個新的、對於法律資訊相關研究有所幫助的成果，並非改善既有的分類任務，因此分類器會應用既有的工具及模型來組成。

其中我們使用了 TensorFlow[2]的框架，並以其下 TensorFlow Hub[3]所包裝好的，由 Google

[2] https://www.tensorflow.org/
[3] https://www.tensorflow.org/hub

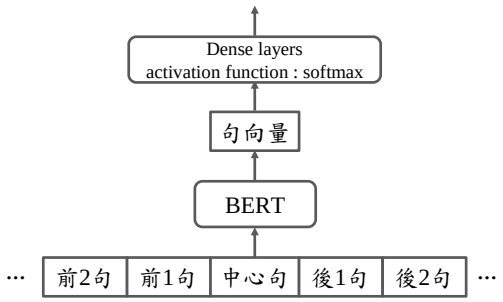

圖 4. 分類模型示意圖

所開源的預訓練模型 BERT[4]針對中文做預訓練的 bert_zh_L-12_H-768_A-12 接上一層的全連接層，並且在訓練中會對 BERT 進行 fine tune。而訓練時會以 TensorFlow 內建的 adamw 作為優化器 (optimizer)，將起始的學習率設定為 5e-5。並使用 TensorFlow 框架內建的 sparse categorical cross entropy 作為損失函數 (loss function)，此外因為硬體限制所以 batch size 設定為 16；若驗證資料的 loss 連續四次沒有下降則停止訓練。

輸入資料除目標句子外，也就是中心句外，也會提供模型額外的前後各 n 句句子，我們會嘗試設定 n 值，從最小為 1 到最大為 5。輸入 BERT 前會使用 BERT 內建的 token "[SEP]"做為前後各個輸入句的分隔及連接，模型設計如圖 4 所示；並取其中較好的結果嘗試 Rao et al. (2018) 提出的 sentence representations LSTM 概念。往後會將該模型簡稱為 SR-LSTM，此部分會在往後章節詳細介紹。

實驗時，使用 scikit-learn[5]套件中的工具，以裁判書為單位隨機挑選 20%的篇數做為測試資料，剩餘 80%的篇章中取 20%做為訓練過程的驗證資料，80%做為模型的訓練資料。

雖然僅 814 篇裁判書，但若以句子為單位，我們每次實驗會有 70000、18000、23000 筆以上的訓練、驗證及測試資料，這些資料的數量會受到 random seed 選取的影響而浮動。

為了避免數據洩露的問題，所有的資料都是以裁判書為單位切割出大段落後，以大段落為範圍來生成包含有前後不同句數的資料。也就是每次輸入的 n 句短句會以裁判書的大段落為範圍，不會輸入不同大段落的短句，若是前或後短句不足則會以 "[PAD] "替代之。

BERT + Dense					
	n 句前後文				
	1	2	3	4	5
C1	0.724	0.777	0.743	0.750	0.756
C2	0.545	0.594	0.524	0.542	0.468
C3	0.875	0.858	0.815	0.809	0.790
C4	0.833	0.852	0.815	0.797	0.761
macro F_1	0.744	0.770	0.724	0.724	0.694

表 2. 人工標記測試資料驗證規則標記訓練的模型

BERT + Dense					
	n 句前後文				
	1	2	3	4	5
C1	0.723	0.767	0.740	0.744	0.743
C2	0.529	0.565	0.518	0.523	0.440
C3	0.783	0.777	0.749	0.742	0.740
C4	0.772	0.785	0.766	0.744	0.708
macro F_1	0.702	0.723	0.694	0.688	0.658

表 3. 規則標記測試資料驗證規則標記訓練的模型

6 實驗結果

我們首先會實驗以實驗設計段落提出的規則標記方法來對裁判書進行 C1、C2、C3、C4 四分類，之後分別以規則標記及人工標記的資料進行測試，確認初步的規則標記結合深度學習方法可行之後以人工標記資料訓練模型並與我們規則標記訓練出來的模型進行對比。

除了四分類模型外我們也會拓展到五分類，透過五分類模型來找出裁判書中的爭點，以及依靠爭點句特性設想出的方法實驗結果。

6.1 以規則標記資料訓練句的四分類模型

這個章節我們會使用前面資料標記章節所述的規則標記資料做為訓練資料的模型，並且分別使用人工標記及規則標記的方式對測試資料進行標記後測試模型效果。

以表 2 與表 3 分別呈現測試資料以人工標記及規則標記後的 F_1 score。

[4] https://github.com/google-research/bert

[5] https://scikit-learn.org/

BERT + Dense					
	n 句前後文				
	1	2	3	4	5
C1	0.741	0.745	0.728	0.759	0.758
C2	0.559	0.562	0.504	0.538	0.490
C3	0.788	0.788	0.774	0.767	0.757
C4	0.783	0.788	0.776	0.779	0.778
macro F_1	0.718	0.721	0.696	0.711	0.696

表 4. 人工標記測試資料驗證人工標記訓練的模型

BERT + Dense					
	n 句前後文				
	1	2	3	4	5
C1	0.763	0.767	0.748	0.785	0.784
C2	0.581	0.592	0.531	0.577	0.534
C3	0.936	0.928	0.920	0.928	0.904
C4	0.874	0.878	0.868	0.883	0.875
macro F_1	0.788	0.791	0.766	0.793	0.774

表 5 規則標記測試資料驗證人工標記訓練的模型

令人意外的是，就算是以規則標記的訓練資料訓練模型，在人工標記的測試資料上整體都維持有比在規則標記的測試資料上更高的 F_1 score。推測是因為在大方向上規則標記的方式有成功找出與專家所給的標記相符合的意見，但其中仍有許多不完善的地方會造成規則標記的方式在少數場合無法得到正確的標記，使其中含有額外的雜訊使得規則標記的測試資料整體 F_1 score 都比人工標記的測試資料還略低 0.03 到 0.05 之間。

另外，在 n=2 時在 macro F_1 上有最佳的分數，之後隨著前後句句數增長反而略微下降，在此猜測或許是過多的不夠精確的標記配上前後文反而令分類模型混淆。

6.2 以人工標記資料訓練句的四分類模型

這個章節中我們透過使用經由法律領域專家標記資料做為訓練資料的模型。表 4 與表 5 分別為以人工標記的測試資料測試及以規則標記的測試資料測試的 F_1 score。

BERT + SR-LSTM				
	人工標記		規則標記	
	n 句前後文		n 句前後文	
	2	4	2	4
C1	0.752	0.806	0.729	0.770
C2	0.494	0.620	0.478	0.582
C3	0.933	0.938	0.786	0.783
C4	0.875	0.898	0.783	0.794
macro F_1	0.763	0.816	0.694	0.732

表 6. 人工標記訓練資料訓練 SR-LSTM

BERT + SR-LSTM				
	人工標記		規則標記	
	n 句前後文		n 句前後文	
	2	4	2	4
C1	0.742	0.778	0.735	0.778
C2	0.529	0.602	0.515	0.595
C3	0.867	0.865	0.782	0.788
C4	0.839	0.851	0.777	0.796
macro F_1	0.744	0.774	0.702	0.739

表 7. 規則標記訓練資料訓練 SR-LSTM

以規則標記的測試資料整體趨勢與先前實驗相似，皆是在 n=2 時 F_1 score 最高。但 n=4 時不管是規則標記或人工標記的測試資料皆有所上升，在人工標記的測試資料中 F_1 score 是分數最高的。

考量兩個實驗的結果，後續實驗會先以 n=2 與 n=4 的句數為主要實驗目標進行實驗與驗證。

6.3 SR-LSTM 模型四分類測試

表 6 與表 7 分別整理測試多次以人工標記或規則標記的訓練資料訓練 SR-LSTM 後各類別的 macro F_1 score 以及整體 4 項類別的 macro F_1 score。

除了前面實驗使用的直觀的 BERT 接上單層的全連接層進行分類外，我們模仿(Rao et al., 2018)所提出的 SR-LSTM 模型概念，以 BERT 輸出詞向量做為嵌入層後交由與輸入句數數量相同、彼此獨立的 LSTM 將詞向量轉換成句向量。此處每個短句會輸入同一個 BERT 內，但短句間的關係並不透過 BERT 處理，而是使用 BERT 下一層的 LSTM。BERT 下一層的每

SR-LSTM 5 分類實驗						
		人工標記				
		C1	C2	C3	C4	C5
預測結果	C1	3361	305	104	237	9
	C2	293	1434	4	319	35
	C3	1	1	4596	183	3
	C4	760	354	315	7991	302
	C5	94	36	13	1027	1096

表 8. SR-LSTM(n=4) 5 分類結果混淆矩陣

SR-LSTM 5 分類實驗					
	人工標記				
	C1	C2	C3	C4	C5
macro F_1	0.810	0.620	0.934	0.826	0.627

表 9. SR-LSTM(n=4) 5 分類 macro F_1 score

個 LSTM 彼此獨立，且其 unit 數皆會參考前面 BERT 所輸出的詞向量長度，設定為和句向量長度相同的 unit 數。之後交由另外一層獨立的 LSTM 處理分類問題，因這層 LSTM 主要處理前一層 LSTM 所整理出來的句向量，因此 unit 數設定與前一層獨立 LSTM 數量相同，意即若 n=9 的 SR-LSTM 模型，則第一層會有 9 個獨立的 LSTM 將詞向量處理為句向量，並由第二層 unit 數為 9 的 LSTM 接受處理過後的句向量來輸出分類問題的答案。這一 SR-LSTM 模型在訓練時也會同時使用訓練資料對 BERT 進行 fine tune。

與單純使用 BERT 加上一層全連接層不同的是，所有 SR-LSTM 的結果皆呈現 n=4 的時候有著比 n=2 時更好的 F_1 score 且 n=4 時有著目前最好的 F_1 score，因此後續延伸實驗會以 SR-LSTM 為主要的分類模型。

6.4 延伸實驗與假設驗證

除了嘗試依照句子在裁判書中的功能性將其分類為 C1、C2、C3、C4 四個類別外，我們也嘗試從裁判書中使用分類器及按照我們對徵點句的假設嘗試找出爭點。

6.4.1 五分類測試

文章開頭提到，我們除了希望完成裁判書中文章架構的功能性分類外，也想要從中整理出該次判決的爭點以幫助使用者能更快更輕鬆的理解裁判書，因此我們先使用最直觀的方式，將問題轉換成一個 5 分類問題，並選擇以先前有最高 macro F_1 score 的 SR-LSTM，並同樣設置前後句數 n=4 來做為五分類測試的實驗模型。

SR-LSTM (n=4)		
測試資料	人工標記	
訓練資料	人工	規則
C1	0.772	0.737
C2	0.654	0.622
C3	0.934	0.846
C4	0.888	0.840
macro F_1	0.812	0.761

表 10. 以大段落總數為 6 的人工標記資料驗證

我們目前尚未找出能以規則標記爭點的方式，因此此處的訓練資料、驗證資料、測試資料都會使用專家人工標記的資料為主。下表 8 顯示了其中最接近整體 10 次實驗平均的某次其混淆矩陣，另外表 9 則是與先前相同，進行 10 次實驗後整理 5 分類模型的 macro F_1 score。

其中 C1、C2、C3 的部分與原先 4 分類實驗中並無太大差異，而 C5 則與開頭的假設類似，有部分 C5 句子會與 C1、C2 混淆，而可能因為 C5 大多是參雜在 C4 類別句子之中的一段描述，因此極易與 C4 的句子混淆。

6.4.2 從易混淆句子中搜尋爭點

此外我們也嘗試使用先前的假設為概念，將各個大段落中的句子分為 C1、C2、C3、C4 後，將大段落的分類視為其中句子分類標記最多相同類型，並從被分為 C4 的大段落中檢查其中 C1、C2 的句子是否屬於爭點。

不幸的是，在初步的嘗試中準確率及召回率分別只有 0.16 與 0.04。

6.4.3 以大段落數為 6 的文章驗證 4 分類模型

如表 10 所見，我們挑選了相同案由不過大段落數為 6 的裁判書，這些裁判書通過相同的前處理後仍有 322 篇，並且依照相同的前處理方式處理大段落數為 6 的裁判書資料，以此驗證先前兩種訓練資料訓練出來的模型能應用於不同大段落數的裁判書上。

BERT + Dense 前後各 4 句				
訓練資料	人工標記		規則標記	
測試資料	人工	規則	人工	規則
C1	0.272	0.261	0.308	0.300
C2	0.223	0.205	0.166	0.163
C3	0.805	0.688	0.777	0.684
C4	0.751	0.663	0.739	0.656
macro F_1	0.513	0.454	0.497	0.451

表 11. BERT + Dense (n=4)模型，不對 BERT 進行 fine tune

6.4.4 驗證 BERT fine tune 效果

我們嘗試類似 (Zhang et al., 2021) 中所提出的實驗，比較不對 BERT 進行 fine tune 對預測結果的影響。如表 11 所示，若不讓 BERT 對訓練資料進行 fine tune，可以看到 macro F_1 從 0.79 降低到 0.51 左右。

7 結語

對於人工智慧應用於法律資訊的推薦系統亦或是判決預測系統來說，告訴使用者電腦做出該推薦或是預測的原因及正當性是有必要的。

本篇文章中我們比較了使用不同方式來將原始裁判書整理成依照文章架構的功能性上具有意義的段落及句子，並將這問題轉換成分類問題並比較其成效。以目前的成果看來初步的 C1、C2、C3、C4 這 4 大類別的分類上已有不錯的效果。但 C5，也就是爭點的抽取或分類仍有改善空間。

此外，也驗證了規則標記的概念可以應用在無法與專家配合標記資料時快速得到具有一定可靠度的標記；這方面標記是有能力提升下游更應用面的任務，如判決預測任務的準確率及可解釋性。

8 致謝

本研究承國科會研究計畫 107-2221-E-004-009-MY3 與 110-2221-E-004-008-MY3 與國立政治大學高教深耕校內補助計畫 111H124D-13 之部分補助，謹此致謝。

參考文獻

Trevor Bench-Capon, Michał Araszkiewicz, Kevin Ashley, et al.. 2012. A history of AI and Law in 50 papers: 25 years of the international conference on AI and Law. *Artif Intell Law 20*, pages 215–319. https://doi.org/10.1007/s10506-012-9131-x

Yanguang Chen, Yuanyuan Sun, Zhihao Yang, and Hongfei Lin. 2020. Join entity and relation extraction for legal documents with legal feature enhancement. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1561–1571. https://doi.org/10.18653/v1/2020.coling-main.137

Gideon Maillette de Buy Wenniger, Thomas van Dongen, Eleri Aedmaa, Herbert Teun Kruitbosch, Edwin A. Valentijn, and Lambert Schomaker. 2020. Structure-tags improve text classification for scholarly document quality prediction. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 158–167. https://aclanthology.org/2020.sdp-1.18/

Yi Feng, Chuanyi Li, and Vincent Ng. 2022. Legal judgment prediction via event extraction with constraints. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pages 648–664. https://doi.org/10.18653/v1/2022.acl-long.48

Jim-How Ho. 2021. AI 引入民事程序可行性之研究 (The feasibility research on introducing artificial intelligence into civil procedures) [In Chinese]. Doctoral Dissertation, Department of Information Management, National Taiwan University of Science and Technology. https://hdl.handle.net/11296/pkvh27

Wan-Chen Lin, Tsung-Ting Kuo, Tung-Jia Chang, Chueh-An Yen, Chao-Ju Chen, and Shou-de Lin. 2012. Exploiting machine learning models for Chinese legal documents labeling, case classification, and sentencing prediction. *International Journal of Computational Linguistics and Chinese Language Processing, vol. 17, no. 4,* pages 49–68. https://aclanthology.org/O12-5004.pdf

Shang Li, Hongli Zhang, Lin Ye, Xiaoding Guo, and Binxing Fang. 2019. Mann: A multichannel attentive neural network for legal judgment prediction. *IEEE Access*, pages 151144 – 151155. https://ieeexplore.ieee.org/document/8861054

Jack Mumford, Katie Atkinson, and Trevor Bench Capon. 2021. Machine learning and legal argument. In *Proceedings of the 21st Workshop on Computational Models of Natural Argument*, pages 47–56. https://core.ac.uk/display/477904310

Guozheng Rao, Weihang Huang, Zhiyong Feng, and Qiong Cong. 2018. LSTM with sentence representations for document-level sentiment classification. *Neurocomputing*, 308, pages 49–57. https://doi.org/10.1016/j.neucom.2018.04.045

Huihui Xu, Jaromir Savelka, and Kevin D Ashley. 2021. Accounting for sentence position and legal domain sentence embedding in learning to classify case sentences. In *Legal Knowledge and Information Systems*, pages 33–42. https://ebooks.iospress.nl/doi/10.3233/FAIA210314

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does NLP benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230. https://aclanthology.org/2020.acl-main.466/