圖書資訊學刊　第22卷第1期　（2024.6）　頁133-164
https://doi.org/10.6182/jlis.202406_22(1).133

深度學習方法在中國佛教經典目錄分類中的應用

Application of Deep-learning Methods in the Classification of Chinese Buddhist Canonical Catalogs

黃淑齡[1]　王昱鈞[2]　洪振洲[3]

Shu-Ling Huang[1], Yu-Chun Wang[2], Jen-Jou Hung[3]

摘　要

本研究採用深度學習方法，自動對中國佛教經典目錄進行分類。這項研究旨在解決如何將新增的佛教文獻納入傳統分類架構，以及手動分類耗時且難以達成共識等問題。我們透過將CBETA電子佛典集成的新增文獻整合到傳統目錄結構中，提高了分類準確性，並深入探討自動分類錯誤的原因。同時，我們還發現不同經典類別之間的歷史聯繫，未來可用於調整和優化目錄結構，以滿足現代研究需求。本研究成果已獲CBETA研究小組採納，成為未來新文獻編目的參考工具，以確保佛典目錄系統與學術領域變化同步並保持實用性。

關鍵字：CBETA電子佛典集成、大藏經、經錄、深度學習方法、文件分類

Abstract

This research focuses on the classification of Chinese Buddhist scripture catalogs, employing advanced deep learning methods to develop an automatic classification mechanism. Chinese Buddhist scripture catalogs serve as vital tools for organizing and retrieving Buddhist literature, facilitating research in the field. However, with the continuous addition of new texts and the need to adapt to modern academic research requirements, traditional manual classification methods have become time-consuming and less effective.

In this study, we aim to address this challenge by harnessing the power of deep learning techniques. Our research not only involves the integration of new literature into existing catalog structures but also explores the reasons behind misclassifications in these catalogs. Additionally, we examine the inherent connections between different categories of scriptures to provide a comprehensive understanding of the catalog structure. Our contributions in this research encompass:

1. Automatic Classification: We pioneer the use of deep learning methods for the classification of Buddhist scripture catalogs, allowing for faster and more accurate categorization of texts. This automated approach significantly enhances the efficiency of catalog management.
2. Error Analysis: We delve into the reasons behind misclassifications in the catalogs, shedding light on common pitfalls and misconceptions in traditional classification methods.
3. Interconnections: We uncover the original interrelationships between different categories of scriptures, offering valuable insights for adjusting and optimizing the catalog structure to align with contemporary research needs.
4. Practical Application: The research findings are adopted by the CBETA research group as a reference tool for the cataloging of new literature in future editions of the Tripitaka, ensuring that the catalog remains up-to-date and relevant.

This research is a significant step towards modernizing the management of Chinese Buddhist scripture catalogs, making them more efficient, accurate, and adaptable to the evolving landscape of Buddhist literature.

Keywords: CBETA; Tripitaka; Scripture Catalog; Deep-learning Methods; Text Classification

[1, 2, 3] 法鼓文理學院佛教學系

Department of Buddhist Studies, Dharma Drum Institute of Liberal Arts, New Taipei, Taiwan

*　通訊作者Corresponding Author: 王昱鈞Yu-Chun Wang, E-mail: ycwang@dila.edu.tw

Extended Abstract

The Catalog of Buddhist Scriptures, known as Jing Lu (經錄) in Chinese, played a crucial role in documenting and preserving the titles and content of Buddhist texts from ancient China. With the advent of the Comprehensive Buddhist Electronic Text Archive (CBETA), an online database of the catalog comprising over 2.3 billion characters, researchers are finally able to thoroughly scrutinize the entire content of Jing Lu. CBETA facilitates access to the entire Buddhist canon and establishes a standardized framework that enables comparative studies and critical research. However, the digitized catalog poses challenges that must be addressed. First, contradictions exist within the framework due to differences in traditional genre-based and topic-based classifications. This inconsistency hinders the accurate categorization of texts and creates difficulties for researchers. Second, the continual addition of new texts to the collection necessitates temporary categorization solutions that may later require revision. These challenges highlight the necessity of a flexible and definitive means of classification to ensure the utility and relevance of the catalog as Buddhist studies evolve. However, traditional manual classification of Buddhist texts is time-consuming, laborious, and often arbitrary. Thus, the present study proposes an automatic classification system that uses deep-learning models.

This study applied two deep-learning models to reclassify Buddhist texts in the last four comparatively modern categories of CBETA. Scriptures in these categories do not conform to the traditional thematic classification system and must be manually integrated into the 19 traditional categories. Moreover, this study assessed the suitability of the current text classification system, addressed ambiguities between categories, and identified the reasons behind misclassifications. By conducting automatic classification of all texts in accordance with the traditional 19 categories, the results of this study's analysis provide crucial reference data for manual revisions of the CBETA catalog.

This study utilized two machine-learning models: bidirectional long short-term memory (BiLSTM) and bidirectional encoder representations from transformers (BERT). Support vector machines (SVM) was also used as a comparison baseline. BiLSTM was selected for its ability to generate contextual word embedding and handle variable-length inputs, which renders it advantageous in text classification tasks. BERT,

Note. To cite this article in APA format: Huang, S.-L., Wang, Y.-C., & Hung, J.-J. (2024). Application of deep-learning methods in the classification of Chinese Buddhist canonical catalogs. *Journal of Library and Information Studies, 22*(1), 133-164. https://doi.org/10.6182/jlis.202406_22(1).133 [Text in Chinese].

To cite this article in Chicago format: Shu-Ling Huang, Yu-Chun Wang, and Jen-Jou Hung, "Application of deep-learning methods in the classification of Chinese Buddhist canonical catalogs," *Journal of Library and Information Studies* 22, no. 1 (2024): 133-164. https://doi.org/10.6182/jlis.202406_22(1).133 [Text in Chinese].

more adept at comprehending contextual nuances and dependencies, outperforms BiLSTM models in capturing semantic and syntactic relationships, whereas SVM's proven effectiveness and interpretability made it an ideal choice as a baseline model for comparison with BiLSTM and BERT. As expected, experimental results demonstrated the superiority of BERT to the other models. When trained and tested on the traditional 19 categories of texts, its test accuracy was 82.4%, surpassing the accuracy of SVM and BiLSTM by 2.8% and 0.5%, respectively.

Error analysis provided the following valuable insights into the classification process:

1. The models were most accurate in classifying specific categories of texts, including the Pure Land School Section (淨土宗部), the Chinese Chan School Section (禪宗部), and the Esoteric Section (密教部). By contrast, the models were less able to correctly identify the scriptures in the Ratnakūṭa Section (寶積部) and Nirvāṇa Section (涅槃部). However, the varying quantity of texts in different CBETA categories did not significantly affect the automatic classification performance.
2. The Avataṃsaka Section (華嚴部), Abhidharma-saṃnipāta Section (論集部), Suttanipāta Section (經集部), Historical Biography Section (史傳部), and Encyclopedia Section (事彙部) were especially likely to be misclassified due to unclear thematic boundaries and mixed categorization. The ease with which these scriptures can be misclassified emphasizes the necessity for a robust, flexible, and accurate classification model.
3. The models tended to cluster scriptures from specific categories together, such as the Nirvāṇa Section (涅槃部) with the Lotus Sutra Section (法華部) and the Avataṃsaka Section (華嚴部) with the Chinese Chan School Section (禪宗部). This clustering may have been due to similarities in topics. Subsequent investigations should be conducted in the future to explore these similarities.
4. After more modern texts were incorporated into the model, the performance of the BERT model decreased due to these texts' substantial differences in language style compared with the majority of the training data. By contrast, the accuracy of the SVM model, which did not consider style, increased. Therefore, in addition to incorporating modern texts with previous expert-defined categories into the training corpus to adjust the model, best practice should involve the combination of predictions from all three models as a reference for human judgment.
5. Cross-tagging is another best practice that can be applied to books that may span multiple categories. This approach allows for a more comprehensive representation of content and facilitates efficient retrieval and exploration of texts with overlapping themes and subject matter. By exploring cross-tagging, the CBETA catalog can better accommodate complex texts and cater to the diverse needs of researchers and scholars. For example, the comparatively modern scripture titled "Biography of Master Tsongkhapa" (宗喀巴大師傳) should

be classified under traditional category No. 18, the Historical Biography Section (史傳部) and was correctly so classified by the BiLSTM model. However, due to Tsongkhapa's connection with esoteric Buddhism, the BERT model placed this work under No. 10, the Esoteric Section (密教部), which, thematically, is also a valid category.

This study makes several contributions to the field. First, the study pioneers the use of deep-learning methods for the automatic and highly efficient classification of Buddhist scripture catalogs, allowing for faster and more accurate categorization of texts. Second, this study explores the reasons behind misclassifications in the catalog, shedding light on the frequently overlapping traditional categories that resist easy classification by any one scheme. Third, this study uncovers interrelationships between various traditional categories of scriptures, offering valuable insights for adjusting and optimizing the catalog structure to align with contemporary research needs. Fourth, the study's findings have been adopted by the CBETA research group as a reference tool for cataloging new literature in future editions of the Tripitaka, ensuring that the catalog remains up-to-date and relevant.

The utilization of deep-learning methods—particularly BERT—in the automatic classification of Chinese Buddhist texts can substantially improve efficiency and accuracy compared with laborious and subjective manual human classifications. The research findings refine and enhance the CBETA catalog, bridging the gap between traditional practices and modern needs. The proposed automatic classification system can serve as a valuable tool for researchers, enabling them to navigate and explore the vast collection of Buddhist texts. Future research may expand upon these findings to advance the fields of Chinese Buddhist studies and the digital humanities.

壹、緒論

佛教自東漢末年傳入中國後，透過譯著和論述而廣為流傳。隨著傳入經典越來越多，各種搜集、記錄、保存佛典或辨別其真偽的書籍也相繼問世，這批著作為早期佛教增添了客觀考察及學術研究的視野，一般稱之為經錄（註一）。經錄雖以目錄為名，但早期是以單書形式出現，這種書籍型態發展至隋唐，有按主題區分的「入藏目」或「入藏錄」出現，著重於區分大小乘和經、律、論「三藏」。同時，以這些入藏錄為依據的「寫本大藏經」隨後形成，使得經錄成為名副其實的叢書目錄（註二）。到了宋代初年，第一部官修「刻本大藏經」《開寶藏》大致依據釋智昇（730）《開元釋教錄．入藏錄》的分類刊刻問世。由於雕版印刷方便流傳，此後陸續有不同版本的大藏經出現，如傳入朝鮮的《高麗藏》、明代的《嘉興藏》、清代的《龍藏》等。這些大藏經所收錄的典籍各有不同，在經目的編排上也有差異，顯示出經錄因應時空環境的需求而不斷演變的過程（李富華、何梅，2003）。

目前，最重要的漢文佛典經錄是財團法人佛教電子佛典基金會（Comprehensive Buddhist Electronic Text Archive Foundation，簡稱CBETA基金會）的藏經目錄，主要是依據日本《大正新修大藏經》（簡稱《大正藏》）訂定

（侯坤宏、卓遵宏，2014），收錄超過2.3億字。《大正藏》的經錄分為「阿含、本緣、般若、法華、華嚴、寶積、涅槃、大集、經集、密教、律部、釋經論、毘曇、中觀、瑜伽、論集、經疏、律疏、論疏、諸宗、史傳、事彙、外教、目錄、古逸、疑似」等26部——不同於傳統大藏經先區分大小乘的作法，而是依佛典產生和流傳時代之順序，並以論述主題作為分類的依據——這種改變是為了因應現代學術研究的需求（註三）。CBETA在此基礎上有所微調，主要是加強主題性，將依體裁劃分的注疏，歸入相應的經、律、論中；此外又新增了現代蒐集的文本，成為23部類（註四）。CBETA目錄的優點是：系統性地整理佛教文獻，並蒐集和維護這些文本；由於同時具有標準化和一致性的特點，有助於研究上的徵引和比較；加上其使用電子載體，能夠更靈活的擴編，使新編或新發現的文獻得以納入收藏，對佛學研究或佛教保存有極大的貢獻。

CBETA的分類方式雖使研究者取用相關資料時有更大的方便，但仍有一些可改進之處，例如：一、CBETA經錄的前19類大致根據主題劃分，但後面新增的4類並非建置在相同的分類架構下，較像是新增資料的暫存類別。二、《大正藏》包含一些集成式的類別，如經集部、論集部（註五），依然為CBETA繼承，這些集成式類別收錄難以分類的文本，形同「其他」部類，導致分類上常出現兩可的情況。三、CBETA特定類別間存在連結關係，促使多重分類的產生，如釋惠敏（2005）提到寶積部、淨土宗部類和大集部類間存在重複分類的文本。然而，這也會使分類問題複雜化，例如：當新增的A類文本被分到B類時，難以確認是分類錯誤，抑或是多重分類的緣故——因為多重分類使得部類界線不夠分明。因此，本文的目的有三：首先，立足於CBETA分類的基礎，探討如何依據CBETA原有的分類架構，將現代新增的後4類文本，以自動分類方式併入前19類，以因應未來更多佛教文獻納入共同分類體系的需求；再者，釐清傳統分類架構中，部類界線不明的可能原因，以期提出改善錯誤分類的建議；最後，我們也希望找出哪些類別具有連結關係，其背後原因何在，以解決多重分類和錯誤分類間的疑義。

由於傳統的分類方法依賴文獻探討、人工閱讀及專家判斷，在巨量資料、類別眾多且界線不明的情況下，為大量佛典分類不僅費時耗力，也很難達成共識。故本文擬採用深度學習方法進行文本自動分類，再提供結果作為人工分類之參考，期能更有效率且客觀地提出經錄重整的建議。我們使用的方法包括：基於變換器的雙向編碼器表示技術（Bidirectional Encoder Representations from Transformers，簡稱BERT），以及雙向長短期記憶模型（Bidirectional Long Short-term Memory，簡稱BiLSTM），由於它們在文件自動分類任務上得到很好的分類結果（LeCun et al., 2015），因此我們首次將其應用於CBETA文本分類上。具體做法是以CBETA前19部類的文本作為訓練資料，讓

不同分類模型學習其分類準則，再用來判斷後4部類文本的類別。由於之前沒有類似的佛典分類研究，因此為評估其效能，我們用長久以來在文件分類領域表現良好的機器學習方法支援向量機分類器（Support Vector Machine，簡稱SVM）作為基線，來比較BERT和BiLSTM分類的成效。

綜合上述，本文的貢獻包括以下幾點：一、突破傳統人工分類方式，首次以人工智慧技術進行佛典文本分類並驗證其效度。二、提出將CBETA新增的敦煌寫本、國圖善本、南傳大藏經及新編部類等4類文本歸併入前19類的分類建議，為未來大藏經的新增文獻找到編目的有效方法。三、根據佛典內容解析分類錯誤的原因，探討集成式部類存在的問題及特定部類間存在的連結關係，提供未來修改經錄結構或進行多重分類之參考。四、在實際應用上，本研究成果獲CBETA研究團隊採納，作為未來經錄分類的參考工具之一。

貳、文獻探討

文件分類（text categorization / classification）從60年代開始發展，是一種自動將文件分派到預先指定類別的技術，當資料量大、類別眾多或不易掌類別分界時，往往能發揮很大的效能（Joachims, 1998; Sebastiani, 2002）。隨著各種電子文件的快速增長，在資訊檢索、內容管理及資訊過濾等應用上，文件自動分類已成為具有實用價值的關鍵技術（代六玲等，2004）。文件分類常基於機器學習，也就是從已知類別的例子中學習類別特徵，然後為新文件判定類別，而特徵通常是以文本表示法（text representation）表達（Sebastiani, 2002）。以下介紹本文使用的三種文件自動分類法的原理及研究成果。

一、支援向量機（SVM）模型

支援向量機是Cortes與Vapnik（1995）發表的二元分類器，是一種監督式學習模型，可以將二維平面的資料以分隔線區分開；若二維平面不能做到，則將資料映射到高維特徵空間中以平面切割，成為一種非線性的分類法。Joachims（1998）最早利用SVM進行文件分類，首先，他指出SVM的優勢在於可以處理較大的特徵空間，例如超過10,000以上的詞。其次，由於文件分類少有不相關的特徵，也就是說排名最低的特徵仍然包含相當多的訊息，所以越能結合眾多特徵的分類器表現會越好。再者，文件向量是稀疏的，而SVM很適合處理具有密集概念和稀疏實例的問題。最後，大多數文件分類都是線性分類問題，而SVM的設計理念就是要找到這樣的線性分離器。由於具有上述優勢，SVM雖然屬於早期的方法，但在文件分類上的效果仍然有突出之處。如Mohammad等人（2016）使用SVM、Naïve Bayes和神經網絡模型進行阿拉伯文的文本分類，結果顯示SVM給出了最好的成果。

由於文本是由一連串的詞彙組成，因此詞彙就成為文本的表徵，也是文件分類的

核心（Jin et al., 2016）。「文本表示」指的是將文本轉成數字或向量的方法（Naseem et al., 2021），主要有兩種類型：其一是離散式的表示（discrete / symbol-based text representations）：將每個詞視為獨立單元來表達，如獨熱編碼（one-hot encoding）、詞袋模型（bag-of-words model）、詞頻－逆文件頻率（Term Frequency-Inverse Document Frequency, TF-IDF）等均是；其二則是分散式表示（distributed text representations）：考慮文本中詞與詞之間的依存關係，又可分兩類：其一，連續型表示（continuous word representations），如詞嵌入（word embedding）；其二，考慮語境的表示（contextual word representations），如語言模型中的嵌入法（Embedding from language Models，簡稱ELMo）、基於變換器的預訓練語言模型（Transformer-based Pre-trained Language Models）等（Cartuyvels et al., 2021; Naseem et al., 2021）。本文用以建立基線的文本表示法，就是離散式中的詞袋模型。

詞袋模型是指不考慮文件中的文法及詞的順序，把所有詞獨立地裝進袋子以代表此文件的方法。透過計算詞袋中單詞出現的次數作為該詞權重（Qader et al., 2019），使每個文件皆由一個高維稀疏向量來表示（Xu et al., 2012）。然而，僅以詞頻選詞會使高頻但不具主題性的功能詞居前；因此，為了找出每個文件中最能區別該文件的詞彙，常進一步以詞袋模型中的TF-IDF方法來表示單詞向量。其計算公式如下：

$$tfidf_{i,j} = tf_{i,j} \times idf_i \quad (1)$$
$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$
$$idf_i = \log \frac{|D|}{|j:t_i \in d_j|}$$

其中，$tf_{i,j}$表示詞頻，分子$n_{i,j}$表示第i詞在第j個檔案中出現的次數，分母$\sum_k n_{k,j}$是第j個檔案中所有詞彙出現的次數總和。idf_i表示逆向的檔案詞頻，$|D|$表示語料庫中的檔案總數，$|j:t_i \in d_j|$是指包含詞彙t_i的所有檔案數。TF-IDF的算法即是將每個單詞的詞頻乘以逆向的檔案詞頻，然後將結果取對數。其意義是利用單詞頻率（即TF）來評估字詞對於文件集或語料庫其中一份文件的重要程度；IDF則把語料庫中的總文件數除以該詞出現的文件數，即對主題性較強的單詞給予更多的權重，使得字詞的重要性隨著在文件中出現的次數成正比增加，但主題性同時會隨著在語料庫中分布的均勻度成反比下降。TF-IDF可以防止較長文件造成偏頗，因此我們可以得到該詞的主題性強度，而不僅僅是原始計數（Eklund, 2018）；但由公式中也可看出，這種方法無法體現單詞在文件中的位置訊息。值得注意的是，傳統文件分類技術在分類前，通常需要刪除停用詞以及出現次數較少的單詞，以增加各類別的鑑別度。本研究刪除了頻率低於5的詞彙；停用詞表部分，由於在不同任務中的使用效果不一，故先進行實驗以評估其效能。由於中文現有的停用詞表多是針對現代漢語所訂，不適用於本研究的佛教文本；因此我們以CBETA文本為材料，採用Hvitfeldt與Sigle（2021）

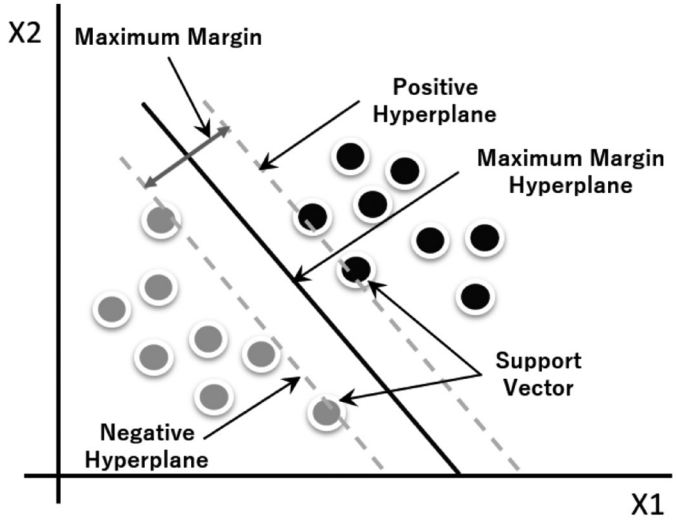
的建議，以前述計算公式中IDF分數最低的1,000筆詞彙作為停用詞——因為它們是跨越不同主題文本的詞彙，即在語料庫中分布最均勻的詞，也就是最不具主題性的關鍵詞。實驗結果顯示，刪除停用詞後，正確率反而下降了1.1個百分點，因此本研究不擬採用停用詞表。

在以TF-IDF計算出文本單詞向量後，即以此作為文本特徵來訓練文件的類別，然後建立SVM分類器，以判斷一個新的文件屬於哪種已知的類別。作法是以詞向量將文件表示為空間中的若干點，並使某一類別的文件盡可能明顯地與其他類別的文件區分開。如圖一中黑色的實心分隔線可以拉開等距的兩條灰色虛線，若虛線間的間距大於任何其他分隔線所能拉出的虛線間距，則判定以黑色實線為黑色文件和灰色文件的最佳分隔線。然後，將新的文件對映到同一空間，根據它們落在哪一類別的範圍，來預測所屬類別。

二、雙向長短期記憶（BiLSTM）模型

近幾年深度學習方法崛起，在大多數自然語言處理任務上皆得到比機器學習更好的分類結果（LeCun et al., 2015），例如：循環神經網路（Recurrent Neural Network, RNN）具有記憶的功能，可以藉著當前詞、前文或預測詞的向量，透過學習到的組合機率來猜測即將出現的詞，因此能處理上下文問題；但是它難以捕捉較長句子的時間關聯，於是又發展成為長短期記憶模型（Long Short-term Memory, LSTM）（Hochreiter & Schmidhuber, 1997）。LSTM可以處理時間序列中間隔或延遲很長的事件，因為具有記憶核，即使在有干擾、不可壓縮的輸入序列情況下，也能學習彌合超過1,000步的時間間隔，並藉由忽略（ignoring）、遺忘／記憶（forgetting / memory）、選擇（selection）匣門來決定哪些前文的字詞可

圖一　SVM分類示意圖



用以預測下文，哪些則被捨棄。Schuster與Paliwal（1997）發表了雙向的循環神經網路（Bi-directional Recurrent Neural Network, BRNN），他們透過正負時間方向同時訓練模型，可以納入過去和未來的背景訊息。Graves等人（2005）則結合LSTM和BRNN，提出雙向的長短期記憶（BiLSTM）模型，用於語音識別及分類上。

近年來，Liu與Guo（2019）將具有注意力（attention）機制和卷積層（convolutional layer）的BiLSTM應用於文件分類上。黃賢英等人（2019）結合了Word2vec和BiLSTM進行中文文本情感分析。Jang等人（2020）則利用CNN（Convolutional Neural Network）抽取相鄰單詞間的關係，得到更高層次的特徵，再以BiLSTM以順序模式處理單詞，捕捉單詞序列之間的更長的依存關係，同時也使用了大型語料庫的Word2vec預訓練權重，以確保模型具有更高的準確性。

在文本表示法上，大多數神經模型的輸入資料為詞嵌入向量（Jin et al., 2016），BiLSTM也不例外。詞嵌入屬於分散式連續詞的文本表示法（continuous words representation），可以在多維空間內，捕捉上下文單詞與單詞之間的關係；然後依據詞間關係，將文本中的每個單詞轉化為分散的詞向量表徵（distributed vector representations）（Mikolov et al., 2013）。詞嵌入最大的優勢在於，能通過保持上下文的詞的相似性和低維向量，提供更有效的向量表示；但由於每個單詞都對應到一個固定長度的向量，因此無法處理一詞多義，也無法保留文件中完整的上下文語義（Naseem et al., 2021）。因為每個詞的嵌入機率都是單獨計算的，故為「非關語境的詞嵌入（non-contextual embeddings）」（Naseem et al., 2021），舉例而言：輸入包含「蘋果」一詞的10個句子，其中5句中的「蘋果」為可食用的果實，另外5句則指手機品牌，詞嵌入將根據此10句訓練出單一的向量作為「蘋果」一詞的表徵。由於無法解歧，與詞袋不管詞序的情況類似，所以也有學者稱它為嵌入袋模型（bag-of-embeddings model）（Jin et al., 2016）。然而，詞嵌入模型和詞袋模型仍有很大差異，主要是前者能表達詞與詞間的關係，例如：在詞袋中，「貓」、「狗」與「蘋果」的不同僅是文本頻率的差異；而詞嵌入演算的結果會使「貓」與「狗」更接近，「蘋果」和它們距離較遠。

本文以Mikolov等人（2013）提出之Word2vec連續型詞袋模型（Continuous Bag of words, CBOW）作為詞嵌入之方法。黃淑齡與王昱鈞（2023）曾以偵測同義詞及干擾詞兩種評估實驗取得模型優化基線，針對CBETA語料調校出最佳模型參數組：CBOW、Dimension 400、window 10、epoch 10，故本文即以此作為BiLSTM實驗中的詞嵌入預訓練模型。CBOW模型與前饋式類神經網路（feedforward neural network）類似，不同之處在於CBOW將非線性隱藏層（non-

linear hidden layer）移除，並且在輸入層的所有單詞皆共享隱藏層。如圖二所示，此模型包含三層，分別為輸入層、投影層、輸出層。

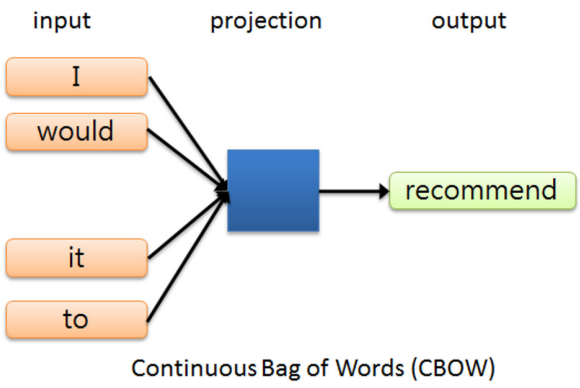
圖二表示在已知當前詞w_t的上下文（w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}）的情況下預測當前詞w_t出現的機率。從形式上看，CBOW模型是由前n個詞語和後n個詞語去預測當前詞的模型。

以CBOW取得詞向量後即用它來表示文本，然後進行BiLSTM建模，依序建立(一)輸入層，最大長度（maxlen）為500；(二) Embedding層，詞向量維度（vector size）是400，這部分藉由載入預訓練好的CBETA W2V模型（黃淑齡、王昱鈞，2023）來達成；(三)然後建立BiLSTM層和(四)密集層；最後是(五)輸出層，即每篇類別的分數。如圖三。

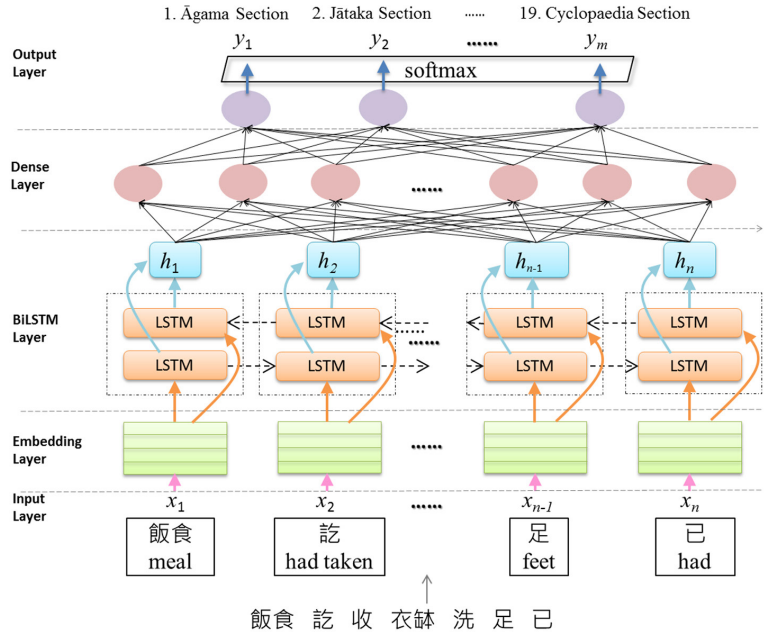
三、基於變換器的雙向編碼器表示（BERT）模型

然而，目前文件分類還是以深度學習中的BERT效果最好（Yu et al., 2021）。BERT是Devlin等人（2018）利用大量無標記文本預訓練出的Transformer的編碼器（encoder）。而Transformer是一種具有注意力機制（self-attention）的深度學習模型，與RNN一樣可用來處理輸入的文句。但是，注意力機制使它不僅能處理前一時間點的訊息，也能為輸入字串的任意位置提供上下文訊息；其優勢在於：(一)字串中的每個字詞向量都把字串裡其他字詞的相關性納入考慮，因此能做到詞嵌入模型做不到的語義解歧，進而提供考慮語境的單詞向量（contextualized word embedding）來作為特徵；(二)能將字詞位置的編碼（positional encoding）加入計算；(三)允許平行計算以減少訓練時間，而不像RNN一次只能處理一個字詞。BERT的執行步驟分為預訓練（pre-training）和微調（fine-tuning）兩部分；它

圖二　連續型詞袋模型



圖三　BiLSTM示意圖



飯食　訖　收　衣缽　洗　足　已

(When He) had taken (His) meal, (He) put away (His) robe and bowl, washed (His) feet.

改變了過去自然語言處理常以一個模型處理一項任務的做法，而在預訓練的階段以大量無標記文本訓練模型去理解一種語言，再用此模型去處理各種任務，期間只要先以預訓練的參數作為預設值，再輸入下游任務有標記的文本，針對所有參數進行微調，即可得到符合下游任務的適當輸出（Devlin et al., 2018）。

BERT不使用傳統的文本表示方法，而是在自我監督學習（self-supervised learning）和遷移學習（transfer learning）的基礎上訓練其雙向編碼學習能力，其採用的兩個自監督任務為：(一)遮罩語言模型（Masked Language Modeling, MLM）：其中15%的標記被任意遮蔽（即用[MASK]標記替換），再訓練模型來預測被遮蔽的標記。(二)下一句預測（Next Sentence Prediction, NSP）任務：同時輸入兩句話到模型中，然後預測第二句話是不是第一句話的下一句（Naseem et al., 2021）。此訓練使BERT能產生考慮到單字或單詞在句子中的位置，及其與上下文關係的向量表示；亦即它能根據語境提供不同的詞向量，使每個單字或單詞向量與句子表達的語義更相關（Subakti et al., 2022）。例如：上文提到的10個包含「蘋果」的句子在BERT中會輸出10個「蘋果」的嵌入分數，而這些分數會根據不同意義自然形成兩大群聚。Yu等人（2021）進一步驗證了用BERT模型代替傳統的詞嵌入模型來表示單詞向量，

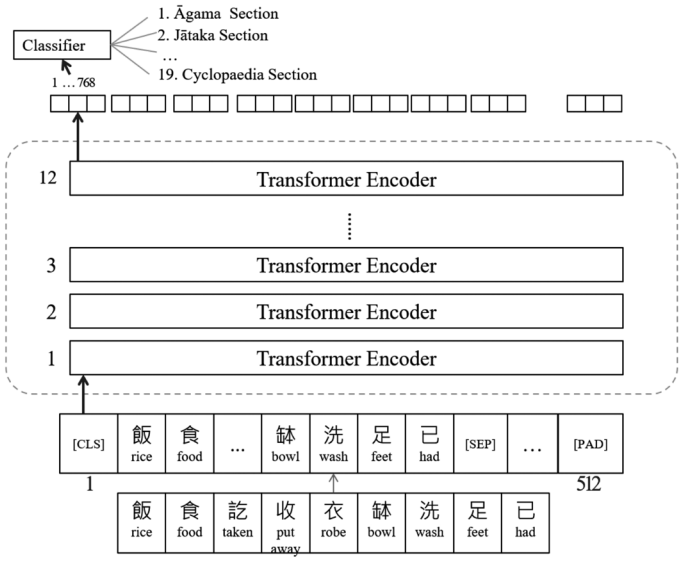
在進行文件分類的任務時，能得到更好的效果。Subakti等人（2022）也驗證了在非監督式的文件分群任務中，使用BERT作為文本表示方法，效果優於傳統的TF-IDF方法。

至於BERT的分類機制，是先將輸入的句子輸出為不同位置的相應嵌入向量，其中，[CLS]位置表示一句的開始，其輸出的嵌入向量代表整句的語義；[SEP]表示一句的結束；[PAD]則是完整輸入的結束位置，最大值是512。BERT模型會在每個輸入的字中輸出一個大小為768的嵌入向量。來自[CLS]標記的嵌入向量會被作為分類器的輸入，然後輸出一個分類任務中類別數量的向量。亦即是，BERT不同於一般分類器，它沒有獨立的分類架構，而是直接在語言模型上進行分類任務，如圖四。

此外，由於BERT僅適用於長度小於（或等於）512個單字的字串，而佛典文本大多長於這個範圍，可能使分類效果受限。因此，我們採用了Pappagari等人（2019）建議的BERT長文件分類方法，也就是將輸入分成較小的區塊，並將每個區塊送入模型；再統合各區塊分數得到最終的分類結果。

目前以BERT進行文件分類的文獻非常多，例如Li等人（2021）以中國COVID-19疫情期間的微博文本作為BERT微調材料進行輿情分析，結果顯示，與其他類似的模型相比，他們所提出的模型性能有了明顯的提高；Abdelgwad等人（2021）以阿拉伯語文本進行BERT微調，在三個不同阿拉伯語資料集上的情感分析結果超越了之前最好的分類結果；Nguyen與Ji（2022）則是透過以生

圖四　BERT的輸入及輸出與分類法



(When He) had taken (His) meal, (He) put away (His) robe and bowl, washed (His) feet.

物醫學文本微調BERT模型，得到目前最好的公共醫療資料分類結果。

本研究採用BERT提供的中文版本（chinese_L-12_H-768_A-12）做為預訓練模型，再以CBETA前19類文本進行微調。由於BERT預訓練過程採用中文字進行訓練，因此我們微調及實驗亦以字為單位。

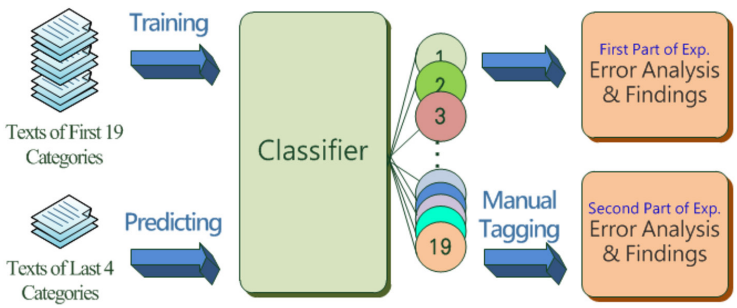
參、研究設計與實作

本文以神經網路模型BiLSTM及預訓練語言模型BERT作為佛典自動分類的主要方法，再以機器學習模型SVM驗證其成效。實驗的流程如圖五。我們先利用有類別標記的CBETA前19類文本訓練出三種模型的分類器，進行錯誤分析後，再利用訓練好的分類器預測CBETA後4類文本的類別，然後與人工分類結果進行比較，驗證其成效及分析錯誤原因。

一、實驗資料集

目前CBETA Online依據部類所列的經錄共計有23類，類名依序是：1阿含、2本緣、3般若、4法華、5華嚴、6寶積、7涅槃、8大集、9經集、10密教、11律部、12毘曇、13中觀、14瑜伽、15論集、16淨土宗、17禪宗、18史傳、19事彙、20敦煌寫本、21國圖善本、22南傳大藏經及23新編。我們從CBETA Online取得經的「部類」、「經號」、「經名」等後設資料（如表一）後，依照「經號」讀入已斷好詞的文檔（註六）「內容」作為訓練的材料，如表二。

圖五　本研究實驗流程圖



表一　CBETA後設資料舉例

部類	經號	經名	卷數	作者
事彙	A091n1057	新譯大方廣佛華嚴經音義	2.0	慧苑
事彙	A097n1267	大唐開元釋教廣品歷章	17.0	玄逸
事彙	A110n1490	天聖釋教總錄	2.0	惟淨
事彙	A111n1493	大中祥符法寶錄	16.0	楊億
事彙	A112n1494	景祐新修法寶錄	14.0	呂夷簡

表二　前處理完成後的資料舉例

經號	卷號	部類	部類號	經名	作者	內容（分詞）	內容（未分詞）
G084n2084	G084n2084_001	禪宗	17	坐禪箴	道元	坐禪 箴\n永平 承陽 大師 道元 撰\n佛 佛 要機 。 祖 祖 機要 。 不思量 而 現...	坐禪箴永平承陽大師道元撰佛佛要機。祖祖機要。不思量而現。不回互而成。不思量而現...
G052n1222	G052n1222_001	密教	10	一字頂輪王瑜伽經	不空	一 字 頂輪王 瑜伽 經\n大興善寺 三藏 沙門 大廣智 不空 奉 詔 譯\n...	一字頂輪王瑜伽經大興善寺三藏沙門大廣智不空奉詔譯...

前19部類共有3,881部經，每一部類所包含的經數、區塊數及詞數如表三。由於每部經長短差異很大（最短1卷，最長600卷），如果以「經」為單位進行資料訓練，對深度學習的方法而言，不僅計算量太大難以負擔，結果也較不精確。因此，我們透過參數最佳化調校，把三種資料集的「經」都拆為以500詞或500字為單位的「區塊」來進行運算（註七），以符合BiLSTM和BERT的輸入需求。其中，前後區塊中各有50字重疊（Pappagari et al., 2019）。不同文本單元的實驗資料數量詳見表四。

二、模型訓練及評估

再來，我們以相同資料用SVM、BiLSTM及BERT等方法進行訓練及驗證：前二者讀入表二已分詞內容，後者讀入未分詞內容作為訓練及測試材料（註八），並皆以「部類號」作為答案，找出這三種方法的最佳模型及驗證和測試正確率。

值得注意的是，進行模型訓練時，由於第一種方法採用的文本單位是「經」，後2種方法採用的是字或詞的「區塊」（註九）；因此，最後比較正確率時，必須制定一套規則來統一單位。本文以「經」作為共同比較對象，採多數決的投票機制來判定類別，規則是：同一經以其區塊中多數判定的類別為該經的類別。例如：某經有6區塊，其中4區塊判定為第9類，2區塊判定為第7類，由於4 > 2，因此該經判定為第9類；唯當判定的類別數相當，無法決定歸屬哪一類時，如上例其中3區塊判定為第9類，3區塊判定為第7類，3 = 3時，則隨機決定為兩類其中的一類。最後，以*Acc*表示分類正確率，計算公式如下：

表三　各部類的經數、區塊數及詞數

部號	部名	經數	詞區塊數	字區塊數	詞數
1	阿含部類	162	5,490	6,543	5,466,104
2	本緣部類	77	5,549	6,628	5,585,994
3	般若部類	217	22,303	28,553	23,084,258
4	法華部類	231	31,807	37,224	31,750,883
5	華嚴部類	157	26,431	31,853	26,993,325
6	寶積部類	57	2,326	2,785	2,327,549
7	涅槃部類	49	8,022	9,403	8,127,579
8	大集部類	35	3,108	3,958	3,231,954
9	經集部類	568	26,851	32,587	27,273,423
10	密教部類	729	25,482	33,270	26,773,333
11	律部類	249	32,139	37,466	31,794,673
12	毘曇部類	38	14,897	17,131	14,510,071
13	中觀部類	41	3,734	4,183	3,595,587
14	瑜伽部類	123	20,523	23,886	20,133,579
15	論集部類	113	6,213	7,130	6,105,358
16	淨土宗部類	134	7,012	8,281	6,942,052
17	禪宗部類	549	57,942	65,921	56,106,675
18	史傳部類	240	37,296	43,736	36,878,348
19	事彙部	112	20,108	23,794	19,962,781

$$Acc = \frac{\text{Number of correct predications}}{\text{Total number of predications}} \quad (2)$$

三、人工標記

第一階段實驗分別為訓練三種分類器、比較不同方法的表現，並觀察可能的錯誤原因。接著，為進一步了解佛典自動分類的有效性，我們也進行第二階段的人工標記實驗。方法是找兩組專業人士針對後4部類的1,001篇文本依19類的分類架構進行類別標記，允許多重分類。然後我們取出兩組標記完全一致的614篇文本，與三種自動分類方法進行比較，觀察它們在相同文本中的答案差異。

肆、研究成果與討論

本節分別說明CBETA前19類文本採用三種分類器的預測效能，以及後4類文本以人工分類為答案的人機比較結果；並討論可能的錯誤原因及提出改進的方案。

表四　實驗資料集檔案數一覽表

文本單位	1至19部類			20至23部類
經	3,881			1,001
	訓練資料（75%）	驗證資料（13%）	測試資料（12%）	
	2,902	513	466	
卷	16,265			3,639
	訓練資料	驗證資料	測試資料	
	12,000	2,421	1,844	
詞區塊（BiLSTM採用）	357,233			94,529
	訓練資料	驗證資料	測試資料	
	266,849	53,978	36,406	
字區塊（BERT採用）	424,332			112,747
	訓練資料	驗證資料	測試資料	
	315,926	64,786	43,620	

一、三種分類器的測試結果

本研究前19類文本經SVM、BiLSTM及BERT模型訓練後，採投票機制選擇出以「經」為單位的最大共識類別，再驗證及測試其預測正確率。測試結果如表五。

第一種方法SVM使用TF-IDF，採計全文的詞彙，並利用詞頻來統計向量表徵；這種方法的優點在於均衡且全面的掌握主題詞，故在採樣均勻度上略勝一籌。然而，這種詞袋型的文本表示法也有侷限：無法表達一詞多義或多詞同義，也不能考量詞彙相對位置等細緻語義關係；因此跟內容相關的特徵（如組織、風格等），都無法在建模時納入考慮，導致SVM分類僅能根據詞彙代表的主題來進行。然而，在CBETA文檔中，某些文本主題類似，但文言文與白話文之行文風格差異很大，則詞袋模型剛好能忽略風格性，單從主題維度上精確分類，使其缺點成為優勢。例如：印順法師的《中觀今論》（釋印順，1947a）以現代白話文書寫，SVM分到「中觀部」，BERT分到「經集部」，BiLSTM分到「禪宗部」，後兩者均考慮了語境、搭配、風格等；但以專業人士的標準來看，SVM的類別較為正確，這可

表五　三種分類方法進行CBETA前19類文本分類的正確率

模型	測試正確率
SVM	0.796
Bi-LSTM	0.819
BERT	0.824

能是因為此文本和訓練語料的行文風格差異太大，導致BERT及BiLSTM反而表現較差。由於CBETA後4類文本中，近現代的文本偏多，此時能凸顯主題的詞袋模型也更具分類優勢；故在第二階段的人機比較中，SVM的正確率反倒優於後兩種方法，詳見後文。

第二種方法BiLSTM以詞嵌入方法來計算單詞向量表徵，能夠捕捉文本中的序列資訊和上下文關係，然而，它不能全面考慮詞在句中的語義，且每一詞僅有一個向量，無法反映一詞多義或多詞同義的情形。第三種方法是預訓練的語言模型BERT，其單詞向量表徵考慮句子的位置資訊，能解決詞彙歧義的問題，如前文曾提到10個含有「蘋果」的句子會計算出10種「蘋果」的嵌入向量，並形成兩種語義（可食用的果實及3C產品品牌）的群聚。由於這兩種方法都考慮上下文關係，對於內容風格的掌握度優於SVM；尤其是BERT，它在大規模語料庫上進行預訓練，具有強大的泛化能力，通過微調可以適應不同的文本分類問題，是目前解決下游任務最好的選擇——我們的實驗結果亦顯示BERT的表現最好。

二、最佳及最差的預測類別

由方法的正確率來看，已知有8成以上的類別都能正確預測；至於各別類別的表現，則以表六的混淆矩陣（confusion matrix）呈現。其中，第一列（row）顯示19種預測類別，第一行（column）則是它們對應的19個實際類別。表中每個預測類別的最高百分比，大多對應到實際類別中的本類——即斜角線灰色部分的數值，通常是該行佔比最高的欄位——這顯示多數預測是正確的。進一步觀察，這些類別中，又以淨土宗、禪宗和密教類的正確率最高，接近100%；相對而言，寶積類和涅槃類則表現最差，因為此二類的多數文本都被預測為經集類，原因於下段討論。

此外，關於語料量對預測結果的影響，由圖六可知，雖然禪宗類的語料最多，寶積類的語料最少，看似與正確率正相關；然而，涅槃類的語料量又大於正確率100%的淨土宗類——可見語料量的多寡並不直接影響預測正確率。

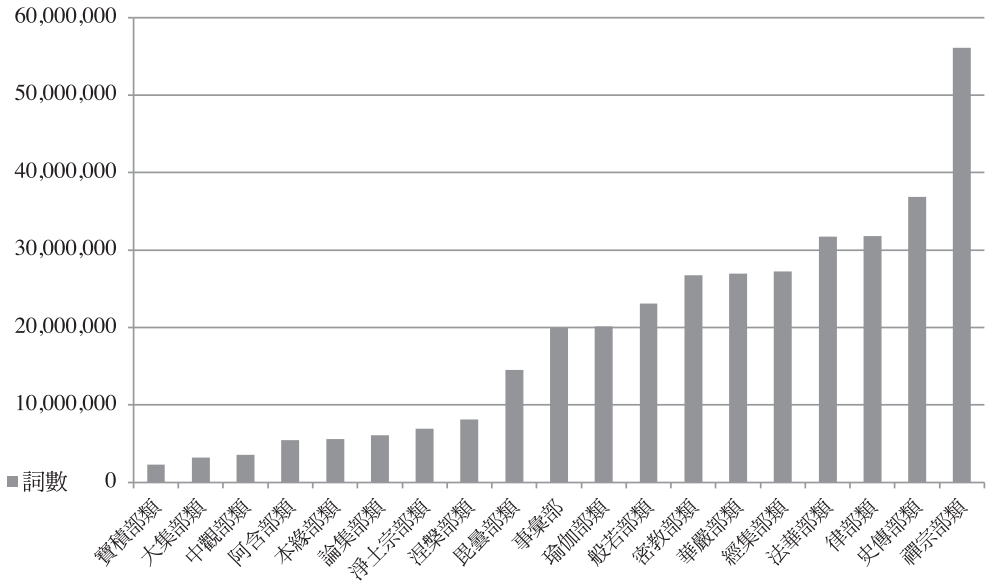
三、最容易引起混淆的類別

在混淆矩陣中，除了從斜角線看出預測正確率外，也可縱觀哪些類別最具混淆性。表中行代表的是前19類文本的實際類別，有值的欄位代表該預測類別實際上應分屬哪些類別，例如：表七行3下的每一格數值表示被BERT預測為本緣類的文本比例，僅有68.75%確為本緣類，其他6.67%實為阿含類，9.21%實為經集類（註十）。由於有些預測類別雖然實際上分屬多種類別，但除了本類之外，其他錯置類別的數值均很低，有可能只是模型的誤差所致。因此，我們嘗試以「正確率（即預測類落在本類）等於或低於80%，且不正確預測（即預測類落在非本類）的數值高於10%」者作為門檻，用來劃定混淆類別的界線，並以表七的灰色區塊顯

表六　從BERT模型計算得出的19類預測混淆矩陣來觀察預測正確率

實際＼預測	1. 阿含	2. 本緣	3. 般若	4. 法華	5. 華嚴	6. 寶積	7. 涅槃	8. 大集	9. 經集	10. 密教	11. 律部	12. 毘曇	13. 中觀	14. 瑜伽	15. 論集	16. 淨土	17. 禪宗	18. 史傳	19. 事彙
阿含	73.33	6.67							13.33		6.67								
本緣		68.75		6.25					25										
般若			87.5	4.17						4.17				4.17					
法華				89.66		3.45			3.45									3.45	
華嚴			5.88		70.59				5.88		5.88						11.76		
寶積						**14.29**			**71.43**									14.29	
涅槃				33.33			**16.67**		**50**										
大集								75	25										
經集		9.21	1.32		2.63	1.32	1.32		80.26	2.63				1.32					
密教									2.3	**97.7**									
律部						3.33	3.33			3.33	73.33	3.33					6.67	3.33	3.33
毘曇												85.71			14.29				
中觀									25				50		25				
瑜伽											8.33		8.33	75	8.33				
論集			6.67		13.33				20	6.67					53.33				
淨土																**100**			
禪宗																	**96.88**	1.56	1.56
史傳											3.7						11.11	74.07	11.11
事彙									7.14	7.14							7.14		78.57

圖六　前19部類總詞數統計



示符合此條件的欄位——可看出混淆類別包括華嚴類、經集類、論集類、史傳類和事彙類等5類。

為評估此判準，我們參照了前人的研究，其提到的混淆類別至少有三種類型：

(一) 經論或部類雜揉：如方廣錩（1997）說「本緣部」經和論雜揉，「經集部」則是主題不明的雜燴類別。《佛光大辭典》「佛教教典」條文下也提到：

> 經集部所收錄之經典，與大集部、寶積部等性質相同，屬集成式者……寶積、大集、經集等三部所收經典，與淨土宗有關者，有無量壽經、觀無量壽經、阿彌陀經；與禪宗有關者，有楞伽經、維摩經；與法相宗有關者，有解深密經；描寫女性之崇高理想者，有勝鬘經；祈求國家繁榮者，有金光明經（釋慈怡，1988，頁3365）。

(二) 單一部類中有不一致的主題：如《佛光大辭典》同一條文下說「論集部收錄上記諸部未輯錄之諸派論書」（釋慈怡，1988，頁3365）。

(三) 相同主題而分屬不同部類：如滿紀（2021）指出「以唯識『六經十一論』來說：六經中的《解深密經》、《楞伽經》在編號9的經集部，十一論在編號15的瑜伽部」（頁205）等。

由此可知，CBETA文本自動分類的錯誤，可能源於本來已存在的類別混淆。這些類別除了文獻提到的本緣部、經集部、大集部、寶積部和論集部，還有文獻未提到，而

表七　從BERT模型計算得出的19類預測混淆矩陣來觀察集成式類別

預測 實際	1. 阿含	2. 本緣	3. 般若	4. 法華	5. 華嚴	6. 寶積	7. 涅槃	8. 大集	9. 經集	10. 密教	11. 律部	12. 毘曇	13. 中觀	14. 瑜伽	15. 論集	16. 淨土	17. 禪宗	18. 史傳	19. 事彙
阿含	73.33	6.67							13.33		6.67								
本緣		68.75		6.25					25										
般若			87.5	4.17						4.17				4.17					
法華				89.66		3.45			3.45									3.45	
華嚴			5.88		**70.59**				5.88		5.88						11.76		
寶積						14.29			71.43									14.29	
涅槃				33.33			16.67		50										
大集								75	25										
經集		9.21	1.32		2.63	1.32	1.32		**80.26**	2.63				1.32					
密教									2.3	97.7									
律部						3.33	3.33			3.33	73.33	3.33					6.67	3.33	3.33
毘曇												85.71			14.29				
中觀									25				50		25				
瑜伽											8.33		8.33	75	8.33				
論集			6.67		13.33				20	6.67					**53.33**				
淨土																100			
禪宗																	96.88	1.56	1.56
史傳											3.7						11.11	**74.07**	11.11
事彙									7.14	7.14							7.14		**78.57**

我們觀察到的華嚴部、史傳部和事彙部。其中，華嚴部混淆的類別為論集部，可以先排除。至於史傳部和事彙部，前者進行人物描述時，往往會引入不同主題的思想，故於分類時容易分到他類，例如：史傳部與禪宗部的歸類混淆多，即是因為禪宗人物的傳記佔大多數；而事彙部包含各種百科辭書、古逸文及外教文獻等，屬於體裁上的「其他」類，所以也是一種集成類別。

四、特定類別間的連結關係

最後，我們從橫向的列來觀察混淆矩陣。每一列中有值的欄位，代表該類為被預測類別之一，例如：表八第2列實際類別為阿含類，其文本中有73.33%被BERT預測為本類阿含，6.67%預測為本緣類，13.33%預測為經集類，及6.67%預測為律部類（註十一）。在排除本類後，我們以「預測類別佔比高於10%，且不屬於上述提到的易混淆類別」者，作為決定哪些特定類別間存有較強的關聯性的判準（註十二）。符合此條件的類別為涅槃類與法華類，以及華嚴類與禪宗類，如表八。它們的關係也可從文獻中得到佐證。如下所述：

(一) 涅槃類與法華類

涅槃類是大乘五大部之一，以《大般涅槃經》為主，此經的要義是講「真我」而標舉「妙有」，一般認為是延伸般若、法華、華嚴等大品類經的思想，以對治消極的涅槃觀（張曼濤，1972）。吉藏曾判定《法華經》和《涅槃經》為佛陀入滅前最後所說之法，他更宣示《涅槃經》的佛陀常住和佛性思想同樣見於《法華經》，亦即《法華經》和《涅槃經》討論的主題很接近（廖明活，2000）。因此，從我們的實驗結果中，也可看出涅槃類之所以分類成效差，除了一半的文本分到經集類外，也有三分之一的文本分到法華類，高於其本類甚多。

(二) 華嚴類與禪宗類

釋聖嚴（1980）認為，宋代以後，華嚴宗已與禪宗合流；郭朝順（2018）則指出，華嚴宗的判教思想將佛陀教法分判為「小、始、終、頓、圓」五教，其中頓教內涵被解讀為等同於禪宗。吳言生（2001）主張，華嚴經的禪悟內涵，包含「夢幻泡影的大乘空觀」、「消除分別的不二法門」，及「絕言離相的禪悟智慧」，對禪宗產生重大的影響。這些均是兩類在主題上近似的可能原因，導致華嚴類文本中有七成落在本類，但也有一成以上被歸入禪宗類。

五、人工分類與機器預測的比較結果及分析

為了進一步確認自動分類的效能，在後4類的1,001篇文本中，我們請兩組專業人士進行獨立標記；取答案完全一致的614篇文本，與三種自動分類方法的預測結果進行比較，結果如表九所示。這個結果和第一階段實驗的測試結果有兩大差異：一是SVM表現最好，BERT表現最差；二是正確率大幅下降，如BERT正確率由82%降至47%。

關於第一個差異，如前所述，SVM表現優於BERT，可能是後4類文本的現代白話

表八　從BERT模型計算得出的19類預測混淆矩陣來觀察特定關係類別

實際＼預測	1. 阿含	2. 本緣	3. 般若	4. 法華	5. 華嚴	6. 寶積	7. 涅槃	8. 大集	9. 經集	10. 密教	11. 律部	12. 毘曇	13. 中觀	14. 瑜伽	15. 論集	16. 淨土	17. 禪宗	18. 史傳	19. 事彙
阿含	73.33	6.67							13.33		6.67								
本緣		68.75		6.25					25										
般若			87.5	4.17						4.17				4.17					
法華				89.66		3.45			3.45									3.45	
華嚴			5.88		70.59				5.88		5.88						**11.76**		
寶積						14.29			71.43									14.29	
涅槃				**33.33**			16.67		50										
大集								75	25										
經集		9.21	1.32		2.63	1.32	1.32		80.26	2.63				1.32					
密教									2.3	97.7									
律部						3.33	3.33			3.33	73.33	3.33					6.67	3.33	3.33
毘曇												85.71			14.29				
中觀									25				50		25				
瑜伽											8.33		8.33	75	8.33				
論集			6.67		13.33				20	6.67					53.33				
淨土																100			
禪宗																	96.88	1.56	1.56
史傳											3.7						11.11	74.07	11.11
事彙									7.14	7.14							7.14		78.57

表九　三種分類方法進行CBETA後4類文本分類的正確率

模型	人工標記正確率
SVM	0.632
Bi-LSTM	0.541
BERT	0.472

文比例較高所致。根據CBETA後設資料所提供的年份統計，後4類文本中有17%是西元1912年後出版的現代白話文文獻，而前19類中並無1912年後出版的文獻；由於現代文本的語言風格與訓練資料中的多數文本差異較大，以致BERT表現不好。至於正確率大幅下降的原因，我們觀察後發現，後4類文本包含大量前面提到的「混淆類別」（見表十）：其中，史傳類、事彙類和經集類加總後，佔總文本約45%（註十三）。進一步分析BERT預測結果，可發現其中一致性的錯誤有高達31%都跟這些混淆類別相關（見表十一）。

由於將來CBETA新增的文獻極有可能包含大量現代文本，未來除了將已由專家判定類別的現代文本加入訓練語料調整模型外，現階段最好的作法可能是兼採三種模型的預測結果，作為人工判斷參考。具體作法不應採用分數相加或是多數決來給定一個最終判斷類別，而應依正確率高低，將BERT、BiLSTM及SVM的預測類別分別標示為Top1、Top2、Top3選項以供專家判斷。此外，也應針對每一部經，列出三種模型的預測一致性，如：三種都相同、兩種一樣、

表十　CBETA目錄中後4類的文本分布

史傳部	27.9%
事彙部	12.2%
瑜伽部	5.5%
般若部	5.2%
禪宗部	5.1%
經集部	5.1%
……	

三種皆異等。根據後4部類人工標記結果，所有文本中，三種都相同的預測數佔總文本的34%，而其預測的類別有83%與專家選定的類別一致。由此可見，此參考資料確實具備輔助效能。我們發現如果把三種模型的預測結合起來，同時作為後4部類文本的候選答案，和人工標記結果對比，正確率可達79%。

伍、結論

對佛教發展而言，如何彙整古今文獻及不同地區的藏經文本，進而建立佛教文獻資料庫，無疑是首要工作之一。其中，典藏這些資料的藏經目錄如何編排至為關鍵，因為它不僅具有承先啟後的意義，也必須因應現代需求而有多元的發展。漢文大藏經相較於藏文大藏及巴利文大藏，具有更古老的經錄傳承歷史及分類系統，也亟需善加保存及廣泛應用。CBETA雖已進一步往主題分類發展，以利學術研究，卻不易完善；因為傳統的分類方法依賴人工閱讀、專家判斷及文獻討論，在巨量資料、類別眾多且界線不明的

表十一　BERT模型預測的前三名一致性錯誤百分比

人工分類_史傳部 -> BERT預測_禪宗部	19.89%
人工分類_事彙部 -> BERT預測_禪宗部	6.16%
人工分類_寶積部 -> BERT預測_經集部	5.32%
總計	31.37%

情況下，重新分類不僅費時耗力，也很難達成共識。因此，為有效解決上述問題，本研究採用數位方法，即先結合資訊計算的自動文本分類法，再提供結果作為人工分類之參考，以取代傳統的專家分類及文獻討論方式，期能更有效且客觀地提出經錄重整的建議。

目前的成果是運用自然語言處理技術的文本自動分類法，為CBETA經錄中新增的4部類進行以主題為主的重新歸類，以符合前19類的分類原則。表現最好的分類器是BERT，在針對前19類文本進行機器訓練測試時，可達到單經分類正確率0.824。然而，考量其他兩種分類器也有各自的分類優勢，如在對後4類文本進行人機比較時，由於文本中加入大量現代文本，這些文本的語言風格與訓練資料中的多數文本差異較大，以致對語言風格敏銳的BERT反而表現不好；而詞袋模型剛好能忽略風格性，單從主題維度上精確分類，故SVM的表現就比BERT模型更傑出。因此，就作為人工判斷的參考資料而言，將三種模型的預測結果結合以供參考，應該是最好的方式。除此之外，我們的研究還提出幾項發現：一、最佳的預測類別是淨土宗類、禪宗類和密教類，最差的預測類別是寶積類和涅槃類，表現差的主要原因是混淆類別和關聯類別所造成；二、最容易引起混淆的類別包括華嚴類、經集類、論集類、史傳類和事彙類等5個類別，因為它們分別是內容或體裁上的集成類別；三、涅槃類與法華類、華嚴類與禪宗類之間存在關聯性，這源於它們思想上的相似性；四、機器預測結果在新增文本中表現較差，可能是因為現代白話文本不包含在訓練資料中，且新增文本包含大量混淆類別所致。

未來，我們將持續探索多元的經錄呈現方式，例如：利用文本主題所產生的文字雲來視覺化佛典的內容特徵；或進一步在語義空間中呈現部類間的關聯性。使得在傳統條列式分類之外，還能有其他顯示經錄關聯性的分類架構。事實上，多重分類的需求由來已久，如前面所提及寶積類、淨土宗類和大集類間存在重複分類的文本，這是因為《寶積經》、《大集經》之名原本就含有「法寶總集」、「諸法聚集」等意義（註十四），故以這些經典為主的部類本身就具有「集成」概念。而現代的多重分類需求又遠超於此，主要包括以下兩種考量：首先，應能兼容同一部佛典之多重屬性，例如：「敦煌石窟文本」一方面可保留此來源屬性，又能依據主題將這些文本歸到傳統架構中。又如

《宗喀巴大師傳》一書，BERT預測為密教類，BiLSTM預測為史傳類；因為宗喀巴大師為密宗的創始人之一，他的生平傳記多圍繞密宗思想而論，故這兩種預測乃根據不同屬性得出，都是正確的。其次，應包含比現行的「一部經」更小的分類單位（如卷、章或節等），例如：印順法師《佛法概論》（釋印順，1947b）中，各章有不同主題，〈中道泛論〉可能屬於中觀部類，而〈正覺與解脫〉可能偏向般若部類。由以上考量，進而產生為文本段落加標籤的分類概念，例如：根據需求，可以在同一文本附加地點屬性或類別屬性的標籤；或於同一文本的不同段落附加不同主題標籤，打破傳統單一書目編排的規則，且能進一步以視覺化的方式呈現這些標籤間的語義關聯，形成有別於以往的圖書編目方式。而機器學習中的分類（classification）、類聚（clustering）、降維（dimensionality reduction）及主題建模（topic modeling）等方法，都有助於在文本上附加屬性或類別標籤。目前我們在實驗中觀察到的部類混淆（如禪宗類與史傳類間的疊合），也許在拓展多重分類的範疇及方法後，就不成問題了。

總體而言，我們的研究證明，深度學習模型在文本分類任務中表現出色。但未來應進一步探討類別混淆和多重分類等複雜情況，以進一步提高準確性，並開展佛典經錄的分類架構及呈現方式，以因應學術研究的需求。

註釋

註一：梁代釋僧祐（510）在現存最早的經錄《出三藏記集》序中提到：「昔安法師以鴻才淵鑒，爰撰經錄，訂正聞見，炳然區分。」（頁1a29-b2）

註二：早期經錄多是根據譯作時代進行分類，隋代法經的《眾經目錄》最早提出先區分大小乘，再在其下區分經、律、論三類的分類體系。隋代費長房（597）《歷代三寶記》中有「入藏目」，唐代釋智昇（730）《開元釋教錄》有「入藏錄」，兩者都繼承了法經的分類體系，再加以完善。《開元釋教錄》的入藏錄更確定了數千卷大小乘經典在一切經中的位次，後來漢地的官寫本大藏經以及各大寺院的寫本大藏經，大多依據《開元釋教錄》書寫，成為當時國家敕准的寫本大藏經的標準（李富華、何梅，2003）。

註三：呂澂（1963）說：「日本從1923至1928年編印《大正新修大藏經》，對漢文大藏的編次再度作了改訂。它以清新圓到的編纂為目標，要在學術基礎上，一新從來經本以混雜排列而使其系統組織明確整齊，這樣就在分類上有顯明的特點。」（頁91a8-11）方廣錩（1997）則說：「《大正藏》則完全拋棄傳統的『重大（乘）輕小（乘）』的分類原則，力圖依據思想的發展與典籍的演變這樣的歷史線索來安排大藏經的結構，以期給研究者

最大的方便。」並指出《大正藏》的編纂原則之一是「打破傳統的大藏經結構體例，按照學術原則重新分類，以反映佛典思想的發展與文獻的變遷」（頁235-246）。

註四：釋惠敏（2002）說：「《大正藏》原來是分為26部，我們（按：指CBETA）則整編成20部類，將經律論的注疏（包含敦煌文獻，也即古逸部）歸併到相關的經律論。（譬如：阿含的部類中，除了原有的『阿含部』（T01-02），也將阿含的論（T25）、阿含的疏（T33）歸併為阿含部類）；宗派（諸宗部）也是一樣，歸併到相關的經律論。淨土宗跟禪宗是中國獨特發展出來的，且不易歸併到某特定的經律論部類，故將它獨立成類。接下來是史傳部（T49-52）與事彙部類（T53-55, 85），最後是疑似部（T85）。」（頁20-21）文中說的20類，之後又刪去最後的「疑似部」，成為現在所見到的19類。後來又新增了「敦煌寫本、國圖善本、南傳大藏經及新編部類」等4類，共計23部類。

註五：如《佛光大辭典》（釋慈怡，1988）「佛教教典」條文下提到「經集部所收錄之經典，與大集部、寶積部等性質相同，屬集成式者」，又說「論集部收錄上記諸部未輯錄之諸派論書」（頁3365）。

註六：我們採用法鼓文理學院針對CBETA開發的自動分詞系統進行分詞，其F-Score為93.96%（Wang, 2020）。分詞後的資料存放於https://github.com/DILA-edu/word-segment/tree/main/word-segmented-cbeta

註七：在區塊數設定方面，我們分別採用了maxlen = 300 / 400 / 500等三種參數針對BiLSTM和BERT進行實驗，皆以maxlen = 500的表現最好。

註八：也就是SVM及BiLSTM以詞彙為訓練及測試單元，BERT以字為訓練及測試單元。我們也曾以字為單位評估SVM的文件分類效能，結果比以詞為單位的模型正確率下降了2個百分點。這是因為中文單字的歧義性太高，根據所查找的文獻，SVM和BiLSTM模型大多都採用詞為單位，理由即是詞彙能更精確地捕捉語義訊息，故與SVM或BiLSTM搭配的文本表示法，如「詞袋模型」或「詞嵌入模型」，多採用詞作為單位。

註九：由於BERT和BiLSTM這種序列模型原本就設計為僅能讀取固定長度的輸入，所以會發生每部經僅能採用前面的數百字進行分類的問題。因此，根據文獻，將文件切分為區塊進行分類，是處理序列模型遇到長文本或大型文件的最佳解方。然而，SVM不同於序列模型，它可以取用全域的關鍵詞，不受限於文本的長短，故不會

有結果較不精確的問題。再者，根據我們的實驗結果，將SVM切分成區塊，反而導致測試正確率由0.796降為0.672。換句話說，SVM在處理長文本時，保持文本的完整性對於其性能是重要的，文件的分割可能會導致SVM失去原本文本的完整性，進而影響其性能。

註十：由於行（column）的總和代表實際類別中每個類別的樣本數量，故其加總值不是100%。

註十一：由於列（row）的總和代表模型的預測結果中每個類別的樣本比例，故每一列加總均為100%。

註十二：本研究決定「特定類別關聯性」的門檻值，如同判定「混淆部類」一樣，是我們自行設定的條件，經調整後可得到不同結果。就判定「混淆部類」而言，我們發現目前所設定的門檻與文獻所述最為一致，例如：接近而未達條件的「本緣部」，確實也是文獻所提及的混淆部類。至於「特定類別關聯性」的判定方式，除了目前以混淆矩陣來觀察外，未來我們也將採主題模型分析的方式，進一步找出其他可能的高關聯部類。

註十三：統計後4類文本的文本分布時，其類別乃參照人工標記的結果得到。

註十四：印順法師的《寶積經講記》說：「《大寶積經》被作為多種經典的編集，在玄奘法師時代，早就如此了。」（釋印順，1962，頁2a4-5）《中華佛教百科全書》（藍吉富，1994）「大方等大集經」條目下則提到，《大集經》之名有「大眾會集」及「諸法聚集」二義。

誌謝

本論文最初口頭發表於第12屆日本數位人文協會年會（JADH 2023），我們誠摯地感謝該協會以及本期刊的匿名審查者所提出的寶貴意見。同時，我們也感謝財團法人佛教電子佛典基金會（CBETA Foundation）校訂組同仁為本研究進行佛典類別的人工標記，並給予豐富的研究建議，謹此一併致以深深的謝意。

參考文獻References

方廣錩（1997）。大正新修大藏經評述。在南京金陵刻經處（編），*聞思：金陵刻經處130周年紀念專輯*（頁230-253）。華文。【[Fang, Guang-Chang] (1997). [Da Zheng xin xiu Da Zang Jing ping shu]. In [Nan Jing Jin Ling Ke Jing Chu] (Ed.), *[Wen si: Jin Ling Ke Jing Chu 130 zhou nian ji nian zhuan ji]* (pp. 230-253). Sino-Culture. (in Chinese)】

代六玲、黃河燕、陳肇雄（2004）。中文文本分類中特徵抽取方法的比較研究。*中文信息學報，18*(1)，26-32。【Dai, Liu-Ling, Huang, He-Yan, & Chen, Zhao-Xiong (2004). A comparative

study on feature selection in Chinese text categorization. *Journal of Chinese Information Processing, 18*(1), 26-32. (in Chinese)】

吳言生（2001）。*禪宗思想淵源*。中華書局。【[Wu, Yan-Sheng] (2001). *[Chan Zong si xiang yuan yuan]*. Zhonghua Book. (in Chinese)】

呂澂（1963）。*新編漢文大藏經目錄*（2023.Q4, LC06, no. 6）。CBETA。https://cbetaonline.dila.edu.tw/zh/LC0006_012【[Lu, Cheng] (1963). *[Xin bian Han wen Da Zang Jing mu lu]* (2023.Q4, LC06, no. 6). CBETA. https://cbetaonline.dila.edu.tw/zh/LC0006_012 (in Chinese)】

李富華、何梅（2003）。*漢文佛教大藏經研究*。宗教文化。【[Li, Fu-Hua], & [He, Mei] (2003). *[Han wen Fo Jiao Da Zang Jing yan jiu]*. China Religious Culture. (in Chinese)】

侯坤宏、卓遵宏（2014）。*六十感恩紀：惠敏法師訪談錄*。國史館。https://doi.org/10.978.98604/18194【[Hou, Kun-Hong], & [Zhuo, Zun-Hong] (2014). *[Liu shi gan en ji: Hui Min Fa Shi fang tan lu]*. Academia Historica. https://doi.org/10.978.98604/18194 (in Chinese)】

張曼濤（1972）。大般涅槃經中的涅槃思想。*華岡佛學學報，2*，5-44。https://www.chibs.edu.tw/ch_html/hkbj/02/hkbj0202.htm【[Zhang, Man-Tao] (1972). [Da Bo Nie Pan Jing zhong de nie pan si xiang]. *Hwakang Buddhist Journal, 2*, 5-44. https://www.chibs.edu.tw/ch_html/hkbj/02/hkbj0202.htm (in Chinese)】

郭朝順（2018年9月14日）。*華嚴宗*。華文哲學百科。https://pse.is/5mjx5d【[Guo, Chao-Shun] (2018, September 14). *Hua-Yen Buddhism*. Encyclopedia of Philosophy. https://pse.is/5mjx5d (in Chinese)】

費長房（597）。*歷代三寶紀*（2022.Q4, T49, no. 2034）。CBETA。https://cbetaonline.dila.edu.tw/zh/T2034_001【[Fei, Zhang-Fang] (597). *[Li dai san bao ji]* (2022.Q4, T49, no. 2034). CBETA. https://cbetaonline.dila.edu.tw/zh/T2034_001 (in Chinese)】

黃淑齡、王昱鈞（2023）。詞嵌入應用於佛學研究—兼論詞嵌入模型評估。*數位典藏與數位人文，12*，43-82。https://doi.org/10.6853/DADH.202310_(12).0003【Huang, Shu-Ling, & Wang Yu-Chun (2023). Word embedding in Buddhist studies: On the basis of evaluation of word embedding models. *Journal of Digital Archives & Digital Humanities, 12*, 43-82. https://doi.org/10.6853/DADH.202310_(12).0003 (in Chinese)】

黃賢英、劉廣峰、劉小洋、陽安志（2019）。基於word2vec和雙向LSTM的情感分類深度模型。*計算機應用研究，36*(12)，3583-3587, 3596。https://doi.org/10.19734/j.issn.1001-3695.2018.08.0599【Huang, Xianying, Liu, Guangfeng, Liu, Xiaoyang, & Yang, Anzhi (2019). Sentiment classification depth model based on word2vec and bi-directional LSTM. *Application*

Research of Computers, 36(12), 3583-3587, 3596. https://doi.org/10.19734/j.issn.1001-3695.2018.08.0599 (in Chinese)】

廖明活（2000）。吉藏與大乘《涅槃經》。*佛學研究中心學報，6*，111-137。【Liu, Ming-Wood (2000). Jizang and the Mahayana Mahaparinirvana-sutra. *Journal of the Center for Buddhist Studies, 6*, 111-137. (in Chinese)】

滿紀（2021）。《佛光大藏經・入藏目錄》考。*佛光學報，7*(1)，189-221。【Man, Ji (2021). ["Fo Guang Da Zang Jing Ru Zang Mu Lu" kao]. *Fo Guang Journal of Buddhist Studies, 7*(1), 189-221. (in Chinese)】

藍吉富（1994）。*中華佛教百科全書*。上海古籍。http://cbeta.buddhism.org.hk/bkqs/2【Lan, Ji-Fu (1994). *[Zhong Hua Fo Jiao bai ke quan shu]*. Shanghai Classics. http://cbeta.buddhism.org.hk/bkqs/2 (in Chinese)】

釋印順（1947a）。*中觀今論*（2023.Q3, Y09, no. 9）。CBETA。https://cbetaonline.dila.edu.tw/zh/Y09n0009【Shih, Yin-Shun (1947a). *[Zhong guan jin lun]* (2023.Q3, Y09, no. 9). CBETA. https://cbetaonline.dila.edu.tw/zh/Y09n0009 (in Chinese)】

釋印順（1947b）。*佛法概論*（2023.Q3, Y08, no. 8）。CBETA。https://cbetaonline.dila.edu.tw/zh/Y08n0008【Shih, Yin-Shun (1947b). *[Fo fa gai lun]* (2023.Q3, Y08, no. 8). CBETA. https://cbetaonline.dila.edu.tw/zh/Y08n0008 (in Chinese)】

釋印順（1962）。*寶積經講記*（2023.Q3, Y02, no. 2）。CBETA。https://cbetaonline.dila.edu.tw/zh/Y0002【Shih, Yin-Shun (1962). *[Bao Ji Jing jiang ji]* (2023.Q3, Y02, no. 2). CBETA. https://cbetaonline.dila.edu.tw/zh/Y0002 (in Chinese)】

釋惠敏（2002）。CBETA版電子佛典集成部類目錄（CBETA版經錄）簡介。*佛教圖書館館訊，32*，18-25。http://www.gaya.org.tw/journal/m32/32-main2.htm【Bhikshu, Huimin (2002). [CBETA ban dian zi Fo dian ji cheng bu lei mu lu (CBETA ban jing lu) jian jie]. *Information Management for Buddhist Libraries, 32*, 18-25. http://www.gaya.org.tw/journal/m32/32-main2.htm (in Chinese)】

釋惠敏（2005）。資訊時代的佛典目錄初探——以CBETA版電子佛典部類目錄為例。在釋惠敏（主編），*佛教與二十一世紀：第四屆中華國際佛學會議中文論文集*（頁427-444）。法鼓文化。【Bhikshu, Huimin (2005). [Zi xun shi dai de Fo dian mu lu chu tan: Yi CBETA ban dian zi Fo dian bu lei mu lu wei li]. In H. Bhikshu (Ed.), *[Fo jiao yu er shi yi shi ji: Di si jie Zhong Hua guo ji Fo xue hui yi Zhong Wen lun wen ji]* (pp. 427-444). Dharma Drum. (in Chinese)】

釋智昇（730）。*開元釋教錄*（2022.Q4, T55, no. 2154）。CBETA。https://cbetaonline.dila.edu.tw/zh/T2154_001【[Shi, Zhi-Sheng] (730). *[Kai Yuan Shi Jiao lu]* (2022.Q4, T55, no. 2154). CBETA. https://cbetaonline.dila.edu.tw/zh/T2154_001 (in Chinese)】

釋慈怡（主編）（1988）。*佛光大辭典*。佛光文化。【[Shi, Ci-Yi] (Ed.). (1988). *[Fo Guang da ci dian]*. Fo Guang Culture. (in Chinese)】

釋聖嚴（1980）。中國佛教的特色——禪與禪宗。*華岡佛學學報，4*，5-20。https://buddhism.lib.ntu.edu.tw/FULLTEXT/JR-BJ007/bj74_1.htm【Shih, Sheng-yen (1980). [Zhong Guo Fo Jiao de te se: Chan yu Chan Zong]. *Hwakang Buddhist Journal, 4*, 5-20. https://buddhism.lib.ntu.edu.tw/FULLTEXT/JR-BJ007/bj74_1.htm (in Chinese)】

釋僧祐（510）。*出三藏記集*（2022.Q4, T55, no. 2145）。CBETA。https://cbetaonline.dila.edu.tw/zh/T2145_001【[Shi, Seng-You] (510). *[Chu san zang ji ji]* (2022.Q4, T55, no. 2145). CBETA. https://cbetaonline.dila.edu.tw/zh/T2145_001 (in Chinese)】

Abdelgwad, M. M., Soliman, T. H. A., Taloba, A. I., & Farghaly, M. F. (2021). Arabic aspect based sentiment classification using BERT. *Journal of Big Data, 9*, Article 115.

Cartuyvels, R., Spinks, G., & Moens, M.-F. (2021). Discrete and continuous representations and processing in deep learning: Looking forward. *AI Open, 2*, 143-159. https://doi.org/10.1016/j.aiopen.2021.07.002

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273-297. https://doi.org/10.1007/BF00994018

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies* (Vol. 1, pp. 4171-4186). Association for Computational Linguistics.

Eklund, M. (2018). *Comparing feature extraction methods and effects of pre-processing methods for multi-label classification of textual data* [Unpublished master's thesis]. KTH Royal Institute of Technology. http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-231438

Graves, A., Fernández, S., & Schmidhuber, J. (2005). Bidirectional LSTM networks for improved phoneme classification and recognition. In W. Duch, J. Kacprzyk, E. Oja, & S. Zadrożny (Eds.), *Artificial neural networks: Formal models and their applications—ICANN 2005* (pp. 799-804). Springer. https://doi.org/10.1007/11550907_126

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hvitfeldt, E., & Sigle, J. (2021). *Supervised machine learning for text analysis in R*. Chapman & Hall/CRC. https://doi.org/10.1201/9781003093459

Jang, B., Kim, M., Harerimana, G., Kang, S., & Kim, J. W. (2020). Bi-LSTM model to increase accuracy in text classification:

Combining Word2vec CNN and attention mechanism. *Applied Sciences, 10*(17), Article 5841. https://doi.org/10.3390/app10175841

Jin, P., Zhang, Y., Chen, X., & Xia, Y. (2016). Bag-of-embeddings for text classification. In G. Brewka (Ed.), *Proceedings of the twenty-fifth international joint conference on artificial intelligence (IJCAI-16)* (pp. 2824-2830). AAAI.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In C. Nédellec & C. Rouveirol (Eds.), *Machine learning: ECML-98* (pp. 137-142). Springer. https://doi.org/10.1007/BFb0026683

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*, 436-444. https://doi.org/10.1038/nature14539

Li, H., Ma, Y., Ma, Z., & Zhu, H. (2021). Weibo text sentiment analysis based on BERT and deep learning. *Applied Sciences, 11*(22), Article 10774. https://doi.org/10.3390/app112210774

Liu, G., & Guo, J. (2019). Bidirectional LSTM with attention mechanism and convolutional layer for text classification. *Neurocomputing, 337*, 325-338. https://doi.org/10.1016/j.neucom.2019.01.078

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 26 (NIPS 2013)* (pp. 3111-3119). Curran Associates.

Mohammad, A. H., Alwada'n, T., & Al-Momani, O. (2016). Arabic text categorization using support vector machine, naïve Bayes and neural network. *GSTF Journal on Computing, 5*(1), 108-115. https://doi.org/10.7603/s40601-016-0016-9

Naseem, U., Razzak, I., Khan, S. K., & Prasad, M. (2021). A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models. *ACM Transactions on Asian & Low-Resource Language Information Processing, 20*(5), Article 74. https://doi.org/10.1145/3434237

Nguyen, B., & Ji, S. X. (2022). Fine-tuning pretrained language models with label attention for biomedical text classification. *arXiv:2108.11809 [cs.CL]*. https://doi.org/10.48550/arXiv.2108.11809

Pappagari, R., Żelasko, P., Villalba, J., Carmiel, Y., & Dehak, N. (2019). Hierarchical transformers for long document classification. In R. K. Das (Ed.), *2019 IEEE automatic speech recognition and understanding workshop (ASRU)* (pp. 838-844). IEEE. https://doi.org/10.1109/ASRU46091.2019.9003958

Qader, W. A., Ameen, M. M., & Ahmed, B. I. (2019). An overview of bag of words: Importance, implementation, applications, and challenges. In S. M. Dauda (Ed.), *2019 international engineering conference*

(IEC) (pp. 200-204). IEEE. https://doi.org/10.1109/IEC47844.2019.8950616

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing, 45*(11), 2673-2681. https://doi.org/10.1109/78.650093

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34*(1), 1-47. https://doi.org/10.1145/505282.505283

Subakti, A., Murfi, H., & Hariadi, N. (2022). The performance of BERT as data representation of text clustering. *Journal of Big Data, 9*, Article 15. https://doi.org/10.1186/s40537-022-00564-9

Wang, Y.-C. (2020). Word segmentation for classical Chinese Buddhist literature. *Journal of the Japanese Association for Digital Humanities, 5*(2), 154-172. https://doi.org/10.17928/jjadh.5.2_154

Xu, Z. E., Chen, M., Weinberger, K. Q., & Sha, F. (2012). From sBoW to dCoT marginalized encoders for text representation. In X. Chen (Ed.), *CIKM'12: Proceedings of the 21st ACM international conference on information and knowledge management* (pp. 1879-1884). Association for Computing Machinery. https://doi.org/10.1145/2396761.2398536

Yu, Q., Wang, Z., & Jiang, K. (2021). Research on text classification based on bert-bigru model. *Journal of Physics: Conference Series, 1746*(1), Article 012019. https://doi.org/10.1088/1742-6596/1746/1/012019

（投稿日期Received: 2023/10/30　接受日期Accepted: 2024/1/18）