特徵選取演算法對可讀性模型的影響
Impact of Feature Selection Algorithms on Readability Model

戴采寧 Tsai-Ning Tai[1], 曾厚強 Hou-Chiang Tseng[1], 宋曜廷 Yao-Ting Sung[2]

[1]Graduate Institute of Digital Learning and Education, National Taiwan University of Science and Technology
[2]Department of Educational Psychology and Counseling, National Taiwan Normal University
phoebeqoo@gmail.com, tsenghc@mail.ntust.edu.tw, sungtc@ntnu.edu.tw

摘要

閱讀是獲得知識的重要方式之一。學者指出，為了促進閱讀的成效，提供難易適中的材料是非常重要的。若是閱讀的材料太過簡單，讀者通常無法在閱讀過程中獲得新知；反之，材料若是太難，會造成讀者過重的認知負擔，進而影響其學習成效。因此，給予讀者適性閱讀的材料是一個重要的議題。針對這個問題，有許多學者開始研究可讀性模型，並發現「特徵選取」（Feature Selection）被認為是一個可以提升可讀性模型準確率的重要方式。然而，各種特徵選取演算法和分類器（Classifier）之間的交互作用在過去的研究中並沒有大量地被探討。因此，本研究將使用三種特徵選取演算法：Chi-squared test、ANOVA 及 Mutual Information 和 25 種分類器，應用於國文科 1-12 年級之可讀性模型準確率的比較。實驗結果將呈現準確率最高的模型之特徵選取演算法和分類器。本研究發現使用 ANOVA 做為特徵選取演算法來選取語言特徵並利用 LGBM 做為分類器時，只須採用累加 13 個特徵，在預測 1-12 年級的國文科課文就能達到準確率 48%、鄰近準確率 76%。

Abstract

Reading is one of the most important ways of acquiring knowledge. Researchers have pointed out that to promote the effectiveness of reading, it is very important to provide materials of the right level of difficulty. If the reading materials are too easy, readers usually cannot acquire new knowledge in the process of reading; on the other hand, if the materials are too difficult, it will cause excessive cognitive burden to the readers, affecting their learning effectiveness. Therefore, giving readers appropriate reading is an important issue. To address this issue, many scholars have begun to develop readability models and found that feature selection enhances the accuracy of readability models. However, the interaction between various feature algorithms and classifiers has yet to be much explored in past studies. Therefore, in this study, three feature selection algorithms, Chi-squared test, ANOVA, Mutual Information, and 25 classifiers, were applied to compare the accuracy of readability models for grades 1-12 in the textbooks of the Chinese language. The experimental results show the feature selection algorithm and the paired classifiers with the highest accuracy. This study found that using ANOVA as the feature selection algorithm and LGBM as the classifier can have 48% accuracy, 73% adjacent accuracy, and 85% reduction in the number of features.

關鍵字：中文文本可讀性、特徵選取、機器學習、分類器

Keywords: Chinese Readability, Feature Selection, Machine Learning, Classifier

1 緒論

一般而言，人們能透過閱讀書本、文章、網站等方式來獲得知識（De Clercq & Hoste, 2016）。讀者需要有適合的閱讀材料才能有較好的成效（Kuo et al., 2018）。現今在國際上，有舉辦測量學生閱讀能力的大型比賽，如：跨國評估學生能力計畫（Progress in International Reading Literacy Study, PISA）和國際閱讀素養調查（Programme for International Student Assessment, PIRLS），都將閱讀素養納入為重要指標。台灣 PISA 國家研究中心（2023）將閱讀素養定義為「實現個人目標、增長知識、發展個人潛能以及參

與社會活動，而理解、運用、評鑑、省思與參與文本的能力」。由此可知，國際上非常重視閱讀能力。想要提升閱讀能力，直覺而言，可以透過大量閱讀來提升（Liao, 2011）。然而，在閱讀中，若是材料太過簡單，讀者通常無法在其過程中獲得新知；反之，材料若是太難，則會造成讀者過重的認知負擔（Cambria, 2010）。若能先瞭解文本的難易程度，並給予讀者適合的閱讀材料，則能使讀者在閱讀中有較好的成效（Kuo et al., 2018）。為此，從古至今，有許多學者想了解如何評量文本的難度，而開始研究文本可讀性（Text Readability）（Dale & Chall, 1948; De Clercq & Hoste, 2016; DuBay, 2007; Feng et al., 2010; François & Fairon, 2012; Mc Laughlin, 1969; Si & Callan, 2001）。

文本可讀性是指文本可以被理解的程度。文本可讀性高，也可以有較高的被理解性（Dale & Chall, 1949）。文本有高的被理解性，文本中的資訊才能有效地被讀者吸收（DuBay, 2007）。為了評估文本的難度，許多國家的學者開始研究文本可讀性。如：在法國，François 和 Fairon（2012）研究以法語為第二外語的人工智慧可讀性公式（AI readability formula）。以 CEFR 的標準，將文本分類至各層級。比較專家選取特徵和使用 Spearman 來選取特徵之可讀性模型準確率的差異。在義大利，Dell’Orletta 等人（2014）對具有基本識字水平和輕度智障的成年人進行研究。使用 GRAFTING 做為排名演算法（Ranking Algorithm），發現在評估句子可讀性中，最有效的特徵是句法特徵（Syntactic）和句法形態配列特徵（Morphosyntactic Features）。在菲律賓，Imperial 和 Ong（2020）使用 Spearman correlation 和 Information Gain 做為特徵選取演算法（Feature Selection），在小學教材中，將語言學習模型特徵（Language Model Features）、傳統類的特徵（Traditional Features），如：字數（Word Count）、句子數（Sentence Count）等，以及詞彙類的特徵（Lexical Features），如：生詞率（Type-Token Ratios）、辭彙密度（Lexical Density）等結合。搭配邏輯迴歸（Logistic Regression）和支持向量機（Support Vector Machine，SVM）做為分類器（Classifier）。實驗結果發現，使用邏輯迴歸做為分類器的準確率較支持向量機高。而在邏輯迴歸的特徵選取中，發現使用單獨一類特徵的準確率較低，傳統類的特徵只有準確率 38%、詞彙類的特徵準確率 33%和語言模型特徵準確率 44%。然而，當使用三類特徵所訓練出的模型，能達到準確率 72%。

由上述研究可知，透過特徵選取演算法能提高模型的準確率（De Clercq & Hoste, 2016; Feng et al., 2010; François & Fairon, 2012）。有些研究對象為國小（Feng et al., 2010; Imperial & Ong, 2020）、有些是針對特定文本或特定對象，如：第二外語者（François & Fairon, 2012）或有殘疾者（Dell’Orletta et al., 2014; Feng et al., 2009）。在中文可讀性研究中，由於中文與其他外文本身有句法不對稱（Syntactic Asymmetry）等結構上的差異（Wang & Zou, 2018），因此，國外文本可讀性的研究結果是否與中文可讀性相符，這部分有待實證研究來探討。此外，Liu 等人（2015）研究在 1-9 年級國文科教科書和優良課外讀物中，發現在不同類型的特徵組合在逐步迴歸（Stepwise Regression）與支持向量機中，所訓練的可讀性模型之效能。受此研究的啟發，本研究認為若能在可讀性模型的研究中考慮到更多不同閱讀階段，例如：1-12 年級，將會是值得研究的議題。

有鑑於此，本研究將基於三種特徵選取演算法：Chi-squared test、ANOVA 和 Mutual Information，搭配 25 種分類器，訓練國文科 1-12 年級的可讀性模型，並比較不同特徵選取演算法和分類器的準確率。本研究的內容如下：第二節將描述特徵選取演算法之相關研究，第三節將呈現實驗資料，第四節將分析實驗數據，最後第五節將總結及未來研究展望。

2 相關研究

提升可讀性模型效能的方法有許多種，每位學者所使用的方法都不大相同，因此，開始有許多提升模型準確率的相關研究（Chen & Lin, 2014; Imperial & Ong, 2020）。以分類器而言，當預測項目不為連續時，會形成分類任務。分類器會從所獲得的數據特徵中，預測變項所屬的類別（Pereira, 2009）。分類器有很多種，每個分類器的效果也不相同。舉例來說，Karabulut 等人（2012）比較不同種特徵選取演算法搭配 3 種分類器：

Naïve Bayes、MLP 和 J48 所訓練模型的準確率。實驗結果發現，使用 MLP 做為分類器時，搭配特徵選取演算法，最多可以提升模型 15.6%的準確率。由此可知，當同一份數據放入不同分類器時，會產生不同的分類效果（Ibrahim, 2020）。Liu 等人（2015）使用逐步迴歸與支持向量機建立模型，比較二個分類器在可讀性模型的效能。發現逐步迴歸使用在預測國小文本的準確率較高；在支持向量機中，詞向量表示法做為特徵時，準確率較佳。

以非中文的可讀性而言，另一個常見的作法是使用特徵選取演算法。特徵選取演算法是機器學習（Machine Learning）中常使用的方法，能夠有效地除去冗餘和不相關的特徵（Aghdam et al., 2009; Eesa et al., 2015），並提升模型的準確率（De Clercq & Hoste, 2016; Feng et al., 2010; François & Fairon, 2012）。同時，能減少模型訓練的時間（Zebari, 2020），以及避免當特徵數量相對樣本數量較大時，模型過度擬合：模型在訓練集的預測效果極好，但因為訓練集資料過度擬合，導致測試集效果不佳（Sima & Dougherty, 2006）。Feng 等人（2010） 以邏輯迴歸和 LIBSVM 做為分類器，比較三種特徵選取方法的準確率：將特徵分組，透過貪婪演算法（Greedy Algorithm），選出各組前幾名的特徵-AddOneBest、基於 Weka 的特徵選取演算法選出特徵-WekaFS，以及所有特徵-Allfeatures。發現使用 122 個特徵的 AddOneBest 準確率最高，達到 74%的準確率；其次是使用 273 個特徵的 Allfeatures，有 72.2%的準確率，最後是使用 28 個特徵的 WekaFS，有 70.1%的準確率。雖然使用 WekaFS 的特徵選取演算法準確率最低，但他所使用的特徵數量比 AddOneBest 少 94 個，比 Allfeatures 少 245 個。

反觀在中文可讀性研究中，對於特徵選取的琢磨比較少。Chen 和 Lin（2014）使用特徵選取演算法和特徵提取（Feature Extraction）：Mutual Information、Chi-square test、Information Gain、Principal Components Analysis 和 Latent Semantic Analysis，研究 1-6 年級國文科、社會科、自然科和生命教育科的可讀性模型。從實驗結果發現，使用 Chi-square test 做為特徵選取演算法和 SVM 做為分類器時有最好的效果。由此可知，不同特徵選取演算法搭配不同分類器時，會影響可讀性模型預測結果的高低。然而，目前對於中文可讀性而言，很少將特徵選取演算法和不同分類器放在一起討論。另外，如果能將研究範圍由國小 1-6 年級擴大至 1-12 年級的文本，也將會是值得研究的議題。

本研究使用 Python，運用套件 sklearn.feature_selection（Pedregosa et al., 2011）的分類（Scikit-Learn, 2023）演算法：Chi-squared test、ANOVA 和 Mutual Information 做為特徵選取演算法；和套件 lazypredict.Supervised 的 LazyClassifier 做為分類器。研究在不同特徵選取演算法中，國文科 1-12 年級文本可讀性模型準確率之比較。

3 實驗設計

本研究使用台灣三大出版社：翰林出版社（翰林出版，2009）、康軒出版社（康軒，2009）和南一出版社（南一教師網，2009）98 學年度 1-12 年級國民基本教育（教育部全球資訊網，2023）國文科教科書，共 633 篇文章，其中 80%（507 篇）為訓練集資料，20%（126 篇）為測試集資料。各年級文章數量詳見表 1。

實驗資料有 633 篇文章，將資料分為訓練集資料和測試集資料，所有文章皆透過 CRIE（Chinese Readability Index Explorer，文本可讀性指標自動化分析系統）（Sung et al., 2016）計算出其計算語言學特徵，共計 86 個特徵。再使用特徵選取演算法評估每一個特徵的重要性，並對評估後的特徵進行排序。舉例來說，如表 2 所示，為 1-12 年級 Chi-squared test、ANOVA、Mutual Information 做為特徵選取演算法排序前 10 名的特徵。模型將依照特徵排序，將特徵累加入模型中訓練，再搭配分類器訓練可讀性模型。當可讀性模型訓練完成後，並會預測測試集資料中的文本可讀性，以確認利用不同特徵所訓練的可讀性模型的效能。舉例而言，當使用 Chi-squared test 做為特徵選取演算法時，第一個模型會使用第一個特徵「對應母體實詞頻變異數」做為訓練模型的特徵。第二個模型會累加第二個特徵，換句話說，即使用第一、二個特徵「對應母體實詞頻變異數」和「對應母體詞頻變異數」做為訓練可讀性模型的特徵。以此類推，第

三個模型會使用第一、二和三個特徵，對模型進行訓練。實驗流程圖詳見圖 1。

年級	總數
1	24
2	67
3	61
4	71
5	69
6	70
7	37
8	34
9	28
10	84
11	41
12	47

表 1. 1-12 年級文本數量

特徵排序	Chi-squared test	ANOVA	Mutual Information
1	對應母體實詞頻變異數	華語詞彙難度均方和	實詞種類數
2	對應母體詞頻變異數	文言文詞素	字數
3	領域詞頻變異數	高階級詞彙數	中筆劃字元數
4	領域實詞頻變異數	華語高難度詞數	低筆劃字元數
5	對應母體詞頻平均	華語詞彙難度平均	二字詞數
6	對應母體實詞頻平均	負向連接詞數	段落平均句數
7	字數	文言文辭素總詞數比	華語詞彙難度平均
8	詞數	單字詞比率	詞數
9	低筆劃字元數	流利級詞彙數	句數
10	實詞數	連接詞數	動詞數

表 2. 三種特徵選取演算法排名前 10 名的特徵

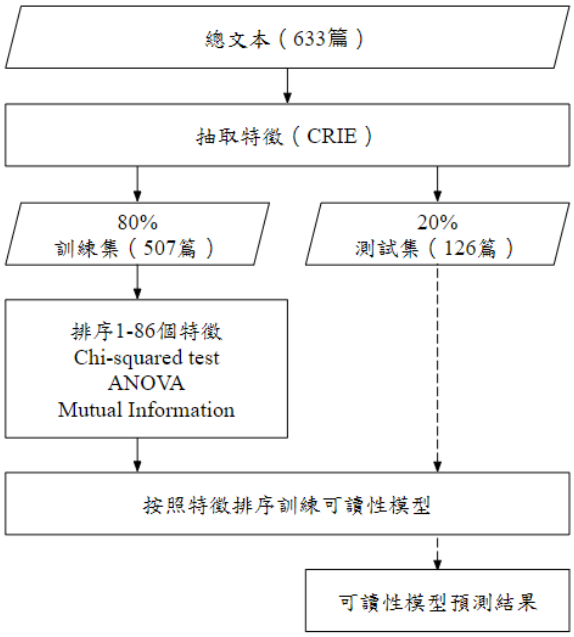

圖 1. 實驗流程圖

4 實驗結果

本研究分析使用三種特徵選取演算法選出的特徵搭配分類器所訓練出來的模型準確率。實驗結果顯示，在三種特徵選取演算法中，準確率最高的分類器以 RandomForest、LGBM 和 ExtraTrees 為主。

文本可讀性模型為分類議題，一般而言，若分類器將文本誤分至前、後一個年級，仍為可接受的範圍。舉例來說，若分類器將三年級的文本誤分為四年級，三年級的學生仍能夠在某個程度上理解四年級的文本；相反地，若將四年級的文本誤分為三年級，四年級的學生也可以閱讀三年級的文本。換句話說，一個三年級的文章，若預測成二、三、四年級，都在可以接受的範圍內。此計算的方法稱為鄰近準確率（Adjacent Accuracy Rate），即將模型所預測的標準放寬至前、後各一個年級。這個做法也可以觀察模型分類錯誤的嚴重性。舉例來說，若將三年級的文本，預測成四年級，仍在可以接受的範圍；但若將三年級的文本預測成十二年級，則與正確答案相去甚遠。由此可知，鄰近準確率越高，表示模型有一定的預測能力；倘若鄰近準確率不高，可以藉由觀察模型預測的答案，了解模型預測錯誤的嚴重性。

4.1 Chi-squared test

如圖 2、3、4 和表 3 所示，在 Chi-squared test 中，模型準確率最高的分類器為 RandomForest、LGBM 和 ExtraTrees。其中，準確率最高的分類器是 RandomForest，在累加 85 個特徵時，準確率 52%，鄰近準確率 79%。然而，在分類器為 ExtraTrees，排名 7，累加 15 個特徵時，就能達到準確率 47%，鄰近準確率 69%。雖然準確率下降 5%，但特徵使用的數量較前者減少 70 個。在使用 LGBM 和 RandomForest 做為分類器，排名 2 的模型，使用累加 30、39 個特徵，皆能夠達到準確率 51%，鄰近準確率 77%、76%。特徵數量較排名 1 分別減少 55、46 個。

因此，在 Chi-squared test 做為特徵選取演算法時，使用 LGBM、RandomForest 和 ExtraTrees 做為分類器，在累加 30、39 和 15 個特徵時，準確率分別能達到 51%、51%和 47%，鄰近準確率 77%、76%和 69%。整體來說，有達到減少特徵的效果。

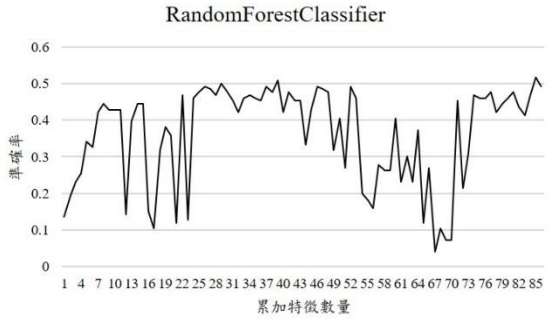

圖 2. 特徵選取演算法 Chi-squared test 搭配分類器 RandomForest 累加特徵之模型準確率趨勢圖

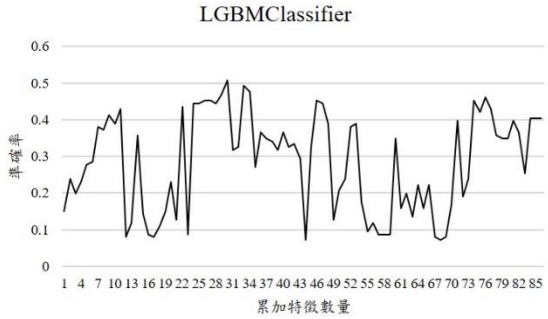

圖 3. 特徵選取演算法 Chi-squared test 搭配分類器 LGBM 累加特徵之模型準確率趨勢圖

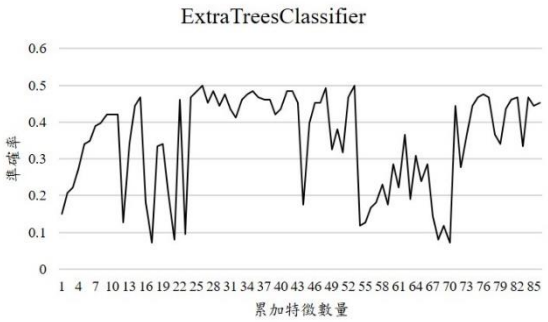

圖 4. 特徵選取演算法 Chi-squared test 搭配分類器 ExtraTrees 累加特徵之模型準確率趨勢圖

排名	分類器	準確率	鄰近準確率	特徵數量
1	RandomForest	52%	79%	85
2	LGBM	51%	77%	30
2	RandomForest	51%	76%	39
3	ExtraTrees	50%	76%	26
3	RandomForest	50%	76%	29
3	ExtraTrees	50%	75%	53
7	ExtraTrees	47%	69%	15

表 3. 特徵選取演算法 Chi-squared test 模型比較

4.2 ANOVA

如圖 5、6、7、8 和表 4 所示，在 ANOVA 中，模型準確率最高的分類器為 RandomForest、ExtraTrees、LGBM 和 Linear Discriminant Analysis。其中，準確率最高的分類器是 RandomForest，在累加 85 個特徵時，準確率 52%，鄰近準確率 79%。此結果與 Chi-squared test 做為特徵選取演算法排名 1 的結果相同。在分類器皆為 LGBM，排名 4、6 的模型中，分別累加 13、12 個特徵時，能達到準確率 48%、46%，鄰近準確率 76%、81%。此模型的準確率相較分類器為 ExtraTrees 和 RandomForest，排名 7，皆累加 76 個特徵，準

確率皆為45%，鄰近準確率分別為75%和72%的模型高。

因此，在 ANOVA 做為特徵選取演算法中，以 LGBM 做為分類器，在累加 13、12 個特徵時，能有準確率 48%、46%，鄰近準確率 76%、81%。由此可推知，該特徵選取演算法可以大量減少累加特徵的數量，且維持模型的準確率。

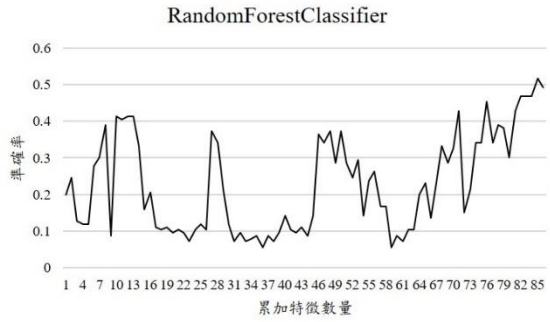

圖 5. 特徵選取演算法 ANOVA 搭配分類器 RandomForest 累加特徵之模型準確率趨勢圖

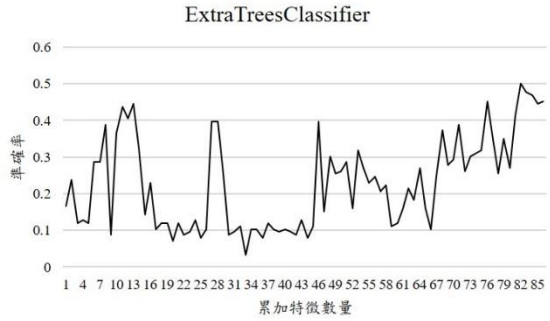

圖 6. 特徵選取演算法 ANOVA 搭配分類器 ExtraTrees 累加特徵之模型準確率趨勢圖

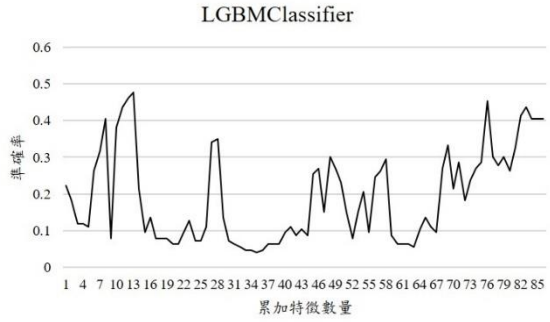

圖 7. 特徵選取演算法 ANOVA 搭配分類器 LGBM 累加特徵之模型準確率趨勢圖

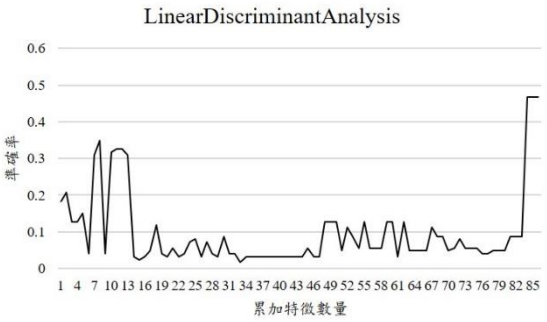

圖 8. 特徵選取演算法 ANOVA 搭配分類器 Linear Discriminant Analysis 累加特徵之模型準確率趨勢圖

排名	分類器	準確率	鄰近準確率	特徵數量
1	RandomForest	52%	79%	85
2	ExtraTrees	50%	79%	82
3	RandomForest	49%	82%	86
4	ExtraTrees	48%	77%	83
4	LGBM	48%	76%	13
5	ExtraTrees	47%	77%	83
5	RandomForest	47%	76%	82
5	RandomForest	47%	81%	84
5	RandomForest	47%	80%	83
5	Linear Discriminant Analysis	47%	74%	84
5	Linear Discriminant Analysis	47%	74%	85

5	Linear Discriminant Analysis	47%	74%	86
6	LGBM	46%	81%	12
7	ExtraTrees	45%	75%	76
7	RandomForest	45%	72%	76
7	LGBM	45%	71%	76

表 4. 特徵選取演算法 ANOVA 模型比較

4.3 Mutual Information

如圖 9、10 和表 5 所示，在 Mutual Information 中，模型準確率最高的分類器為 RandomForest 和 ExtraTrees。其中，有三個模型並列排名 1，準確率皆達 49%，分別是：在 RandomForest 時，使用全部特徵，鄰近準確率 82%；在 ExtraTrees 時，使用累加 80、79 個特徵，鄰近準確率 77%、76%。而在準確率排名 2，分類器為 ExtraTrees，累加 32 個特徵時，能達到準確率 47%，鄰近準確率 73%。在排名 3、二個並列排名 4 的模型中，分類器皆為 RandomForest 時，使用累加 31、32 和 20 個特徵時，分別能達到準確率 46%、45%和 45%，鄰近準確率 71%、74%和 66%。

因此，當 Mutual Information 做為特徵選取演算法時，排名 2、3 和二個並列排名 4 的模型較排名 1 的模型準確率分別低 2%、3%和 4%（二個並列排名 4 的模型）。和排名 1，分類器為 ExtraTrees，累加 79 個特徵時，能達到準確率 49%的模型比較，所使用的特徵數量在排名 2 減少 47 個、排名 3 減少 48 個，排名 4 各減少 47 和 59 個。由此可知，該特徵選取演算法能達到減少特徵數量且維持準確率的效果。

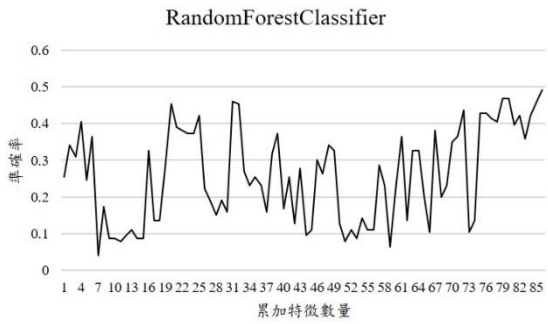

圖 9. 特徵選取演算法 Mutual Information 搭配分類器 RandomForest 累加特徵之模型準確率趨勢圖

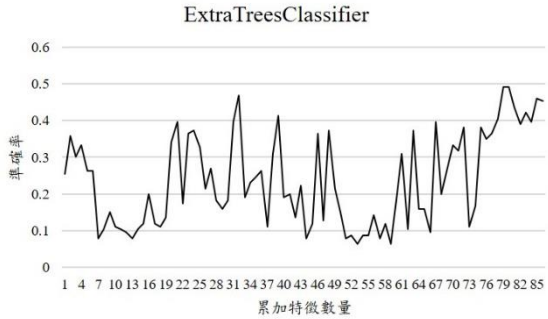

圖 10. 特徵選取演算法 Mutual Information 搭配分類器 ExtraTrees 累加特徵之模型準確率趨勢圖

排名	分類器	準確率	鄰近準確率	特徵數量
1	RandomForest	49%	82%	86
1	ExtraTrees	49%	77%	80
1	ExtraTrees	49%	76%	79
2	ExtraTrees	47%	73%	32
3	RandomForest	46%	71%	31
4	RandomForest	45%	74%	32
4	RandomForest	45%	66%	20

表 5. 特徵選取演算法 Mutual Information 模型比較

4.4 Chi-Square 、 ANOVA 、 Mutual Information 之比較

如表 6 所示，綜上所述的特徵選取演算法，皆能在模型中有效減少特徵數量。發現模型準確率最高的模型分類器為：RandomForest、LGBM 和 ExtraTrees。綜合三種特徵選取演算法準確率、鄰近準確率和特徵數量，整體來說，準確率最高的是 Chi-squared test。在排名 1，分類器為 LGBM、Random Forest ，累加 30、39 個特徵時，準確率皆能達到 51%，鄰近準確率 77%、76%。在排名 2，分類器為 ExtraTrees、Random Forest，累加 26、29 個特徵時，準確率皆能達到 50%，鄰近準確率皆為 76%。使用特徵數量最低的是 ANOVA，在排名 3、5，分類器皆為 LGBM 時，使用累加 13、12 個特徵時，即能達到 48%、46%的準確率，鄰近準確率 76%、81%。Mutual Information 整體的準確率較前面二種特徵選取演算法低。在排名 4、5 和二個並列排名 6 的模型中，排名 4 使用分類器 ExtraTrees，排名 5 和二個並列排名 6 皆使用 RandomForest 做為分類器，在分別累加 32、31、32 和 20 個特徵時，分別能達到準確率 47%、46%和 45%（二個並列排名 6 的模型），鄰近準確率 73%、71%、74%和 66%。雖然整體準確率較低，但仍能達到降低特徵數量的效果。

在 Chi-squared test、ANOVA、Mutual Information 排序前 25 名的特徵中，重複的共有 13 個，如表 7 所示。在 Chi-squared test、Mutual Information 排序前 25 名的特徵中，重複的共有 3 個，如表 8 所示。在 ANOVA、Mutual Information 排序前 25 名的特徵中，重複的共有 3 個，如表 9 所示。根據 Sung（2015），將特徵分為四大類：語意類、詞彙類、句法類、文章凝聚性。在表 8、表 9，排名前 25 名的特徵，皆為詞彙類特徵，例如：入門級詞彙數、高階級詞彙數等。在表 7 中，前 9 個特徵皆屬於詞彙類特徵，另外 4 個特徵中，除「複雜結構句數」屬於句法類特徵，其餘「實詞數」、「文言文詞素」和「實詞種類數」皆屬於語意類特徵。由此推測，表 7 中的詞彙類特徵屬於對各年級文本都很重要的基本特徵。除此之外，詞彙類的特徵，如：「實詞數」、「文言文詞素」等，以及句法類的特徵，如：「複雜結構句數」，是由於本研究的年段橫跨至高中 12 年級，高中 12 年級的文本因為有新的詞彙、文言文和修辭等，使文本難度被提升，文本結構也有所差異。因此，語意類和句法類的特徵可以提升高年級可讀性模型的效能。

排名	特徵選取演算法	分類器	準確率	鄰近準確率	特徵數量
1	Chi-Square	LGBM	51%	77%	30
1	Chi-Square	Random Forest	51%	76%	39
2	Chi-Square	Extra Trees	50%	76%	26
2	Chi-Square	Random Forest	50%	76%	29
3	ANOVA	LGBM	48%	76%	13
4	Mutual Information	Extra Trees	47%	73%	32
4	Chi-Square	Extra Trees	47%	69%	15
5	ANOVA	LGBM	46%	81%	12
5	Mutual Information	Random Forest	46%	71%	31
6	Mutual Information	Random Forest	45%	74%	32
6	Mutual Information	Random Forest	45%	66%	20

表 6. 三種特徵選取演算法之模型比較

特徵名稱	特徵種類
字數	詞彙類
詞數	詞彙類
名詞數	詞彙類
動詞數	詞彙類
進階級詞彙數	詞彙類
華語高難度詞數	詞彙類
難詞數	詞彙類
低筆劃字元數	詞彙類
中筆劃字元數	詞彙類
複雜結構句數	句法類
實詞數	語意類
文言文詞素	語意類
實詞種類數	其他類

表 7. Chi-squared test、ANOVA、Mutual Information 共同排序前 25 名的特徵

特徵名稱	特徵種類
入門級詞彙數	詞彙類
領域詞頻平均	語意類
二字詞數	詞彙類

表 8. Chi-squared test、Mutual Information 共同排序前 25 名的特徵

特徵名稱	特徵種類
高階級詞彙數	詞彙類
華語詞彙難度平均	詞彙類
華語詞彙難度均方和	詞彙類

表 9. ANOVA、Mutual Information 共同排序前 25 名的特徵

5 結論與未來發展

在可讀性研究中，中文的可讀性研究相對比較少，將多個特徵選取演算法搭配多個分類器的研究又更稍微少。因此，本研究在較大的年段之下：國文科一到十二年級，探討三種特徵選取演算法：Chi-squared test、ANOVA、Mutual Information 搭配不同分類器，在可讀性模型預測的準確率、鄰近準確率和選取累加特徵數量之比較。實驗結果顯示，三種特徵選取演算法皆能減少特徵數量並維持模型準確率，並且準確率較高的模型所搭配的分類器皆以 RandomForest、ExtraTrees 和 LGBM 為主。其中，使模型準確率最高的特徵選取演算法為 Chi-squared test，搭配 LGBM、RandomForest 和 ExtraTrees 做為分類器，分別能達到準確率 51%、51%和 50%，鄰近準確率 77%、76%和 76%，特徵數量較原本累加 86 個特徵降至累加 30、39 和 26 個特徵（詳見表 3）。最有效減少特徵且不失準確率的特徵選取演算法為 ANOVA，搭配分類器 LGBM，使用累加 13、12 個特徵，即能達到準確率 48%、46%，鄰近準確率 76%、81%（詳見表 4）。

從過去的研究，可以知道語言特徵再結合語意空間，例如，Word2vec（Maddela & Xu, 2018）、LSTM（Liu et al., 2017）、BERT（Tseng et al., 2019）等，可以提升可讀性模型的效能。因此，本研究在未來可以基於現在的成果知道對於國文科 1-12 年級可讀性模型有效果的特徵（詳見表 7）。在未來，可以再結合其他的特徵，如語意空間等，來提升模型的準確率。

Acknowledgments

本研究承國科會研究計畫 111-2410-H-011-039-MY2 與教育部高教深耕計畫補助國立臺灣科技大學技職賦能研究中心謹此致謝。

參考文獻

Aghdam, M. H., Ghasem-Aghaee, N., & Basiri, M. E. (2009). Text feature selection using ant colony optimization. *Expert systems with applications*, *36*(3), 6843-6853.

Chen, Y. H., & Lin, T. C. (2014). Dimension reduction techniques for accessing Chinese readability. In *2014 International Conference on Machine Learning and Cybernetics* (Vol. 1, pp. 434-438). IEEE.

Dale, E., & Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational research bulletin*, 37-54.

Dale, E., & Chall, J. S. (1949). The concept of readability. *Elementary English*, *26*(1), 19-26.

De Clercq, O., & Hoste, V. (2016). All mixed up? Finding the optimal feature set for general readability prediction and its application to English and Dutch. *Computational Linguistics*, *42*(3), 457-490.

Dell'Orletta, F., Wieling, M., Venturi, G., Cimino, A., & Montemagni, S. (2014). Assessing the readability of sentences: which corpora and features?. In *Proceedings of the ninth workshop on innovative use of NLP for building educational applications* (pp. 163-173).

DuBay, W. H. (2007). *Smart Language: Readers, Readability, and the Grading of Text*.

Eesa, A. S., Orman, Z., & Brifcani, A. M. A. (2015). A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems. *Expert systems with applications*, *42*(5), 2670-2679.

Feng, L., Elhadad, N., & Huenerfauth, M. (2009). Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)* (pp. 229-237).

Feng, L., Jansche, M., Huenerfauth, M., & Elhadad, N. (2010). A comparison of features for automatic readability assessment.

François, T., & Fairon, C. (2012). An "AI readability" formula for French as a foreign language. In *Proceedings of the 2012 joint conference on empirical methods in Sciences language processing and computational Sciences language learning* (pp. 466-477).

Ibrahim, A. A., Ridwan, R. L., Muhammed, M. M., Abdulaziz, R. O., & Saheed, G. A. (2020). Comparison of the CatBoost classifier with other machine learning methods. *International Journal of Advanced Computer Science and Applications*, *11*(11).

Imperial, J. M., & Ong, E. (2020). Exploring hybrid linguistic feature sets to measure filipino text readability. In *2020 International Conference on Asian Language Processing (IALP)* (pp. 175-180). IEEE.

J. Cambria (2010). "Motivating and engaging students in reading," New England Reading Association Journal, vol. 46, Jan.

Karabulut, E. M., Özel, S. A., & Ibrikci, T. (2012). A comparative study on the effect of feature selection on classification accuracy. *Procedia Technology*, *1*, 323-327.

Kuo, B. C., Liao, C. H., & Chang, C. J. (2018). Using LSA-Based Tools to Enhance Students' Chinese Reading Ability. *數位學習科技期刊*, *10*(1), 31-55.

Liao, G. (2011). On the development of reading ability. *Theory and Practice in Language Studies*, *1*(3), 302-305.

Liu, H., Li, S., Zhao, J., Bao, Z., & Bai, X. (2017). Chinese teaching material readability assessment with contextual information. In *2017 International Conference on Asian Language Processing (IALP)* (pp. 66-69). IEEE.

Liu, Y. N., Chen, K. Y., Tseng, H. C., & Chen, B. (2015). 可讀性預測於中小學國語文教科書及優良課外讀物之研究 (A Study of Readability Prediction on Elementary and Secondary Chinese Textbooks and Excellent Extracurricular Reading Materials)[In Chinese]. In *Proceedings of the 27th Conference on Computational Linguistics and Speech Processing (ROCLING 2015)* pp. 71-86).

Maddela, M., & Xu, W. (2018). A word-complexity lexicon and a neural readability ranking model for lexical simplification. *arXiv preprint arXiv:1810.05754*.

Mc Laughlin, G. H. (1969). SMOG grading-a new readability formula. *Journal of reading*, *12*(8), 639-646.

Pedregosa *et al.*, Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825–2830, 2011.

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage*, *45*(1), S199-S209.

Scikit-Learn. (2023). https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html

Si, L., & Callan, J. (2001). A statistical model for scientific readability. In *Proceedings of the tenth international conference on Information and knowledge management* (pp. 574-576).

Sima, C., & Dougherty, E. R. (2006). What should be expected from feature selection in small-sample settings. *Bioinformatics*, *22*(19), 2430-2436.

Sung, Y. T., Chang, T. H., Lin, W. C., Hsieh, K. S., & Chang, K. E. (2016). CRIE: An automated analyzer for Chinese texts. *Behavior research methods*, *48*, 1238-1251.

Sung, Y. T., Chen, J. L., Cha, J. H., Tseng, H. C., Chang, T. H., & Chang, K. E. (2015). Constructing and validating readability models: the method of integrating multilevel linguistic features with machine learning. *Behavior research methods*, *47*(2), 340-354. (SSCI)

Tseng, H. C., Chen, H. C., Chang, K. E., Sung, Y. T., & Chen, B. (2019). An innovative bert-based readability model. In *Innovative Technologies and Learning: Second International Conference, ICITL 2019, Tromsø, Norway, December 2–5, 2019, Proceedings 2* (pp. 301-308). Springer International Publishing.

Wang, B., & Zou, B. (2018). Exploring language specificity as a variable in Chinese-English interpreting. A corpus-based investigation. *Making way in corpus-based interpreting studies*, 65-82.

Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., & Saeed, J. (2020). A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. *Journal of Applied Science and Technology Trends*, *1*(2), 56-70.

台灣 PISA 國家研究中心. (2023). https://pisa.irels.ntnu.edu.tw/project.html

南一教師網. (2009). https://trans.nani.com.tw/NaniTeacher/ (2023 年 8 月訪問)

教育部全球資訊網. (2023). https://www.edu.tw/News_Content.aspx?n=D33B55D537402BAA&s=37E2FF8B7ACFC28B#

康軒. (2009). https://www.knsh.com.tw/ (2023 年 8 月訪問)

翰林出版. (2009). https://www.hle.com.tw/ (2023 年 8 月訪問)