The 2018 Conference on Computational Linguistics and Speech Processing
ROCLING 2018, pp. 116-125

探索結合快速文本及卷積神經網路於可讀性模型之建立

Exploring Combination of FastText and Convolutional Neural Networks for Building Readability Models

曾厚強　Hou-Chiang Tseng
國立臺灣師範大學資訊工程學系 / 心理與教育測驗研究發展中心
Department of Computer Science and Information Engineering
National Taiwan Normal University
Research Center for Psychological and Educational Testing
ouartz99@gmail.com

陳柏琳　Berlin Chen
國立臺灣師範大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan Normal University
berlin@ntnu.edu.tw

宋曜廷　Yao-Ting Sung
國立臺灣師範大學教育心理與輔導學系所
Department of Educational Psychology and Counseling
National Taiwan Normal University
sungtc@ntnu.edu.tw

摘要

可讀性模型可以自動估測文本的可讀性，幫助讀者去挑選符合自己閱讀能力的文件，以達到更好的學習效果。長久以來，研究人員致力於可讀性模型或特徵的研發，尤其近年來隨著表示學習法的蓬勃發展，使得訓練可讀性模型所需要的特徵可以不再需要仰賴專家，這也使得可讀性模型的發展有了一個嶄新的研究方向。然而，不同表示學習法所抽取出來的特徵各有所長，但過去的可讀性研究大多只用單一方法來訓練可讀性模型。因此，本論文嘗試利用類神經網路來融合卷積神經網路及快速文本兩種表示學習法，以訓練出一個能夠分析跨領域文件的可讀性模型，並可以因應文件內容多元主題的特性。從實驗結果可以發現本論文所提出的可讀性模型，其效能可略微勝出單一表示學習法所訓練的可讀性模型。

關鍵詞：可讀性，詞向量，卷積神經網路，表示學習法，快速文本

一、緒論

可讀性(Readability)是指閱讀材料能夠被讀者所理解的程度[1],[2],[3],[4]，當讀者閱讀高可讀性的文件時，會產生較好的理解及學後保留效果[2],[3]。由於文件的可讀性在知識傳遞扮演極為重要的角色，因此西方的可讀性公式發展的非常早[5],[6]，據 Chall 與 Dale[7]在 1995 年的統計，到 1980 年為止相關的可讀性公式就已經超過 200 多個。這些傳統的可讀性研究大多使用較淺層的語言特徵來發展線性的可讀性公式。然而，傳統可讀性公式所採用的淺層語言特徵，並不足以反映文件難度。Graesser、Singer 和 Trabasso[8]便指出，傳統語言特徵公式無法反映閱讀的真實歷程，文件的語意和語法只是文件的淺層語言特徵，並沒有考量文件的凝聚特性。Collins-Thompson[9]亦指出傳統可讀性公式僅著重在文件的表淺資訊，而忽略文件重要的深層特徵。這也讓傳統可讀性公式在預測文本可讀性的結果常遭受到質疑，甚至因為可讀性公式所採用的可讀性特徵過少，導致容易受到有心人士為了達到特定的可讀性數值，而刻意針對可讀性特徵的特性來修改文章，使得文章呈現許多簡短而破碎的句子，降低了文本的流暢度與連貫性，增加閱讀難度[10]。直至今日，可讀性的研究仍持續不斷。研究人員為了克服傳統可讀性公式的缺點，嘗試利用更細緻的機器學習演算法來發展出非線性的可讀性模型，並納入更多元的可讀性指標來共同評量文本的可讀性，除了可以提升可讀性模型的效能，亦可防止有心人士去操弄文本的可讀性[11],[12],[13]。

然而可惜的是，研究人員發現採用一般語言特徵的可讀性模型在應用到特定領域文時，一般語言特徵並無法判斷相同詞彙在不同領域文本時背後所代表的意義。其原因在於特定領域文本的內容著重在闡述領域的「知識概念」，而這樣子的描述方式有別於一般語文的敘述文或故事體的結構。Yan 等人[14]就明確指出美國大型醫學資料庫(Medical Subject Headings, MeSH)中的醫學專業術語的難度與語言特徵公式的音節數、字長無關。針對一般語言特徵無法表徵特定領域知識結構的問題，開始有學者針對這個議題進行研究。例如，Yan 等人[14]利用本體論的技術將美國國家醫學資料庫(Medical Subject Headings, MeSH)的醫學符號階層資料庫作為概念資料庫，從中找出每一個醫學類文件中的概念，並計算概念到此樹狀結構最底部的距離，得出每篇文件概念深度的指標(Document Scop)。Borst 等人[15]則是利用詞表的方式將每個詞彙的「類別複雜度」與「詞頻」兩個分數加總來計算詞彙複雜度，作為評估醫學線上文件的詞彙、句子及文件難度的依據。Tseng 等人[16]則是利用表徵學習(Representation Learning)方法(如：卷積神

經網路、快速文本)自動從原始資料中抽取可讀性特徵，分別訓練出一個能夠分析跨領域文件的可讀性模型，以求可以更加貼近現實應用的需求。

根據過去的研究可以發現，多數的可讀性研究利用單一機械學習的模型來整合不同類型的可讀性特徵去訓練出可讀性模型，讓可讀性模型可以從更多面向來考量文本的可讀性。但若能夠進一步在訓練可讀性模型過程中就可以讓不同的演算法相互分享資訊，則在訓練可讀性模型的過程中便可以享有更多元的資訊，以提升可讀性模型的效能。基於這樣子的想法，本研究將基於類神經網路來融合兩種不同的表示學習法，訓練一個能夠分析跨領域文件的可讀性模型，其效能可優於單一表示學習法所訓練而成的可讀性模型。本論文的內容安排如下：第二節將描述兩種應用於中文可讀性模型的表示學習法。第三節將基於類神經網路來融合兩種不同的表示學習法，以訓練出一個能夠同時分析不同領域文件的可讀性模型。第四節將呈現本論文所提出可讀性模型的效能。最後第五節是總結及未來研究的方向。

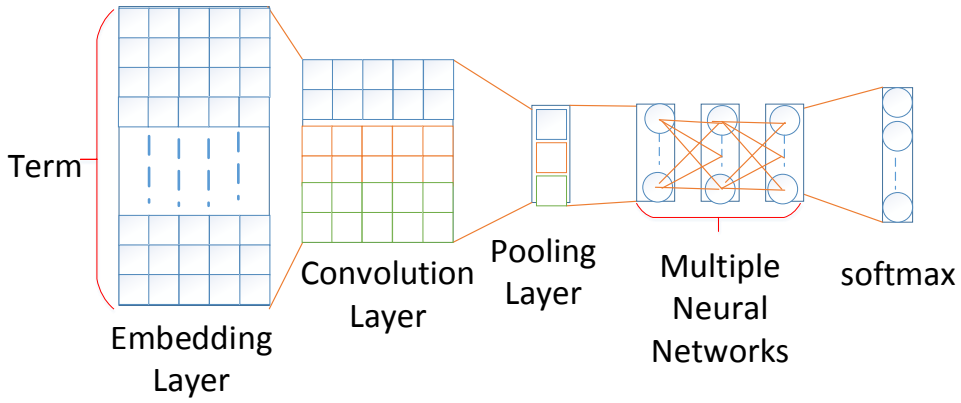
二、基於表示學習技術之可讀性模型相關研究

表示學習法是指學習如何表徵資料的演算法，可以用來抽取有用的資訊來幫助建立預測模型[17]。對於可讀性的研究而言，透過表示學習技術可以自動抽取出訓練可讀性模型所需之可讀性特徵，是一個富有潛力的研究方向。以下將簡述兩個應用於中文可讀性模型的表示學習法。

(一)、卷積神經網路

卷積神經網路(Convolutional Neural Network, CNN)是一種分層式的結構，每個模組都是由卷積層(Convolutional Layer)和池化層(Pooling Layer)來組成[18]，通過模組不斷的疊加或是加上多層的類神經網路後形成深度學習的模型。整個卷積神經網路透過三個重要的思想來幫助改進訓練的效率及效能：稀疏交互(Sparse Interactions)、參數共享(Parameter Sharing)及等變表示(Equi-variant Representations)[19]。稀疏交互又稱為稀疏連接(Sparse Connectivity)，主要是利用數個核(Kernel)基於自定核的大小(Kernel Size)來局部連結兩層網路，使得整個模型所要儲存的參數變少，可以有效減少計算量和提升計算效率。參數共享指的是在同一個核中，每一個元素在不同位置所作用的權重都是相同的。

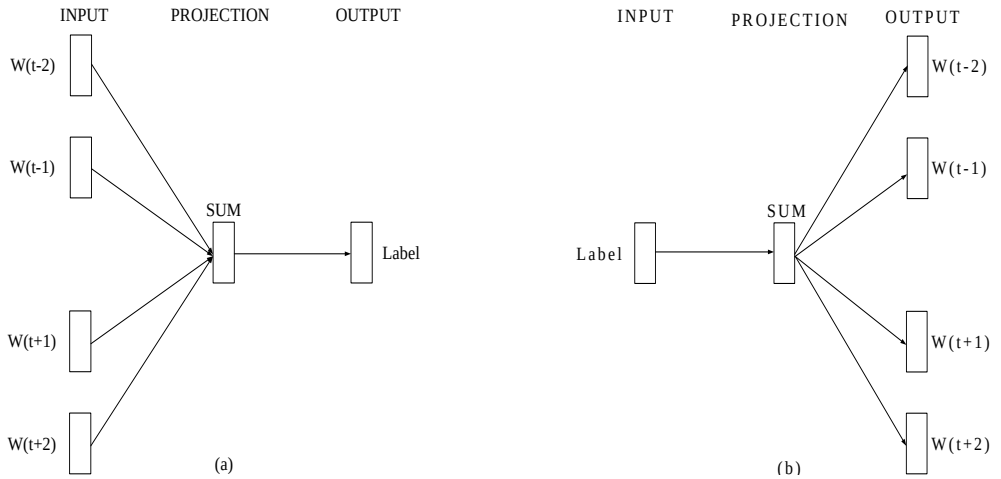
這意味著在卷積運算的過程中，模型只需要學習一組固定的參數即可。因此這也將大幅度提高模型的訓練效能。而參數共享的機制再加上適當的池化策略，也讓卷積神經網路對於局部平移有一些不變的特性可以應用於圖像的處理或語音辨識，尤其是在關心某個特徵是不是有出現，而不是關心它出現的具體位置時[19]。這樣的特性也適用於文本可讀性的分析：在意的是對於文本中所提及的主題或知識概念的難度，而不是關心它在文本中具體的位置。因此，過去已有研究將卷積神經網路用來自動抽取可讀性模型所需要的特徵，並利用深層類神經網路訓練出可讀性模型[16]，其架構如圖一所示，在訓練的過程中，該研究使用到 Dropout[20]的技巧來避免模型過度適配(overfitting)外，並利用 rectified linear units (ReLU) [21]作為的激發函數(active function)，以避免典型的梯度消失(gradient vanish)問題。



圖一、基於卷積類神經網路及深層類神經網路之可讀性模型架構

(二)、快速文本 (fastText)方法

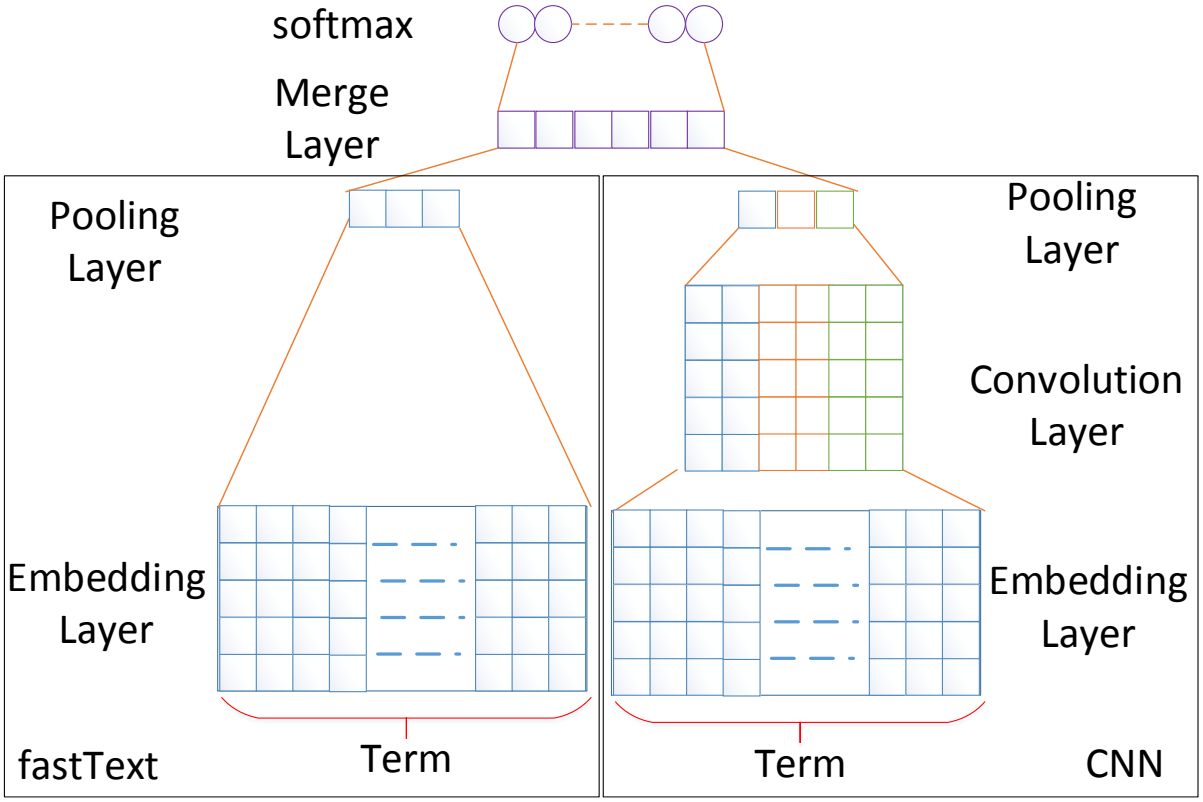
快速文本[22]是繼 Word2vec[23]之後，Joulin 等人持續發展出的架構，其架構如圖二(a)及圖二(b)所示，快速文本與 Word2vec 一樣有連續詞袋模型和略詞模型兩種訓練的架構。但不一樣的地方在於輸入層改為 n-gram 的方式來考慮句子或文章的詞序關係，並且將目標詞彙改換成訓練資料的類別。而這樣子的作法可以讓語意空間特別針對類別去進行優化，這將有助於區別語意或主題相近但實際上卻屬於不同類別的句子或文章。在過去，Tseng 等人[16]也嘗試將快速文本採用連續詞袋演算法來訓練出中文的可讀性模型。



圖二、(a)快速文本之連續詞袋模型。(b)快速文本之略詞模型。

三、基於類神經網路融合不同表示學習法的可讀性模型

不同的表示學習法各有所長，例如卷積神經網路可以透過多個核來觀察文本中多個局部範圍的文本資訊來提取出特徵。有別於卷積神經網路，快速文本則是透過 n-gram 的方式來看同一個滑動視窗中詞彙與類別之間的關係，使其所訓練的詞向量所形成的語意空間可以特別針對類別去進行優化。假若在訓練可讀性模型的過程中可以同時融合這兩種演算法，其訓練可讀性模型的過程中就可以相互分享資訊，為可讀性模型帶來更豐富的資訊去評估文本的可讀性。基於這個想法，本研究利用類神經網路來融合卷積神經網路及快速文本兩種不同的表示學習演算法，其示意圖如圖三所示，在整個可讀性模型的訓練過程中，卷積神經網路和快速文本所產生的特徵會以向量的形式在融合層進行相加、相乘、平均或串聯等不同的運算方式進行融合，而融合後的特徵便可以讓可讀性模型在訓練的過程中享有不同表徵學習法所帶來資訊。



圖三、融合卷積神網路及快速文本的可讀性模型架構

四、實驗及結果

(一)、實驗材料

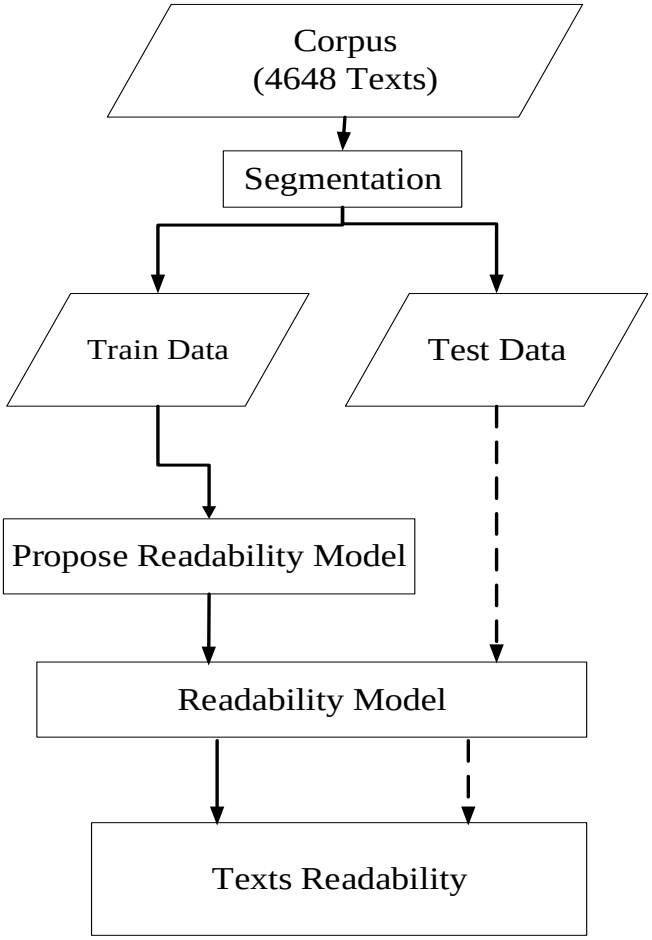
本研究材料選自98年度臺灣H、K、N三大出版社所出版的1-12年級審定版的國語科、社會科和自然科等三個領域的教科書全部共計4,648篇，各版本教科書均經由專家根據課程綱要編制而成，其實驗材料的年級分佈如表一所示。

表一、實驗材料在各年級的數量分佈

年級	1	2	3	4	5	6	7	8	9	10	11	12
社會科	0	0	80	74	85	81	389	407	325	340	331	270
自然科	0	0	72	67	67	62	172	175	157	211	355	295
國文科	24	67	61	71	69	70	37	34	28	84	41	47

(二)、訓練可讀性模型

首先將實驗材料利用 WECAn[24]來進行中文斷詞的前處理程序，接著再以四比一的比例亂數從不同的年級去挑選訓練資料和測試資料。最後，將訓練資料輸入本研究所提出的架構來訓練可讀性模型，而可讀性模型所需的類別就是課文所屬的年級。待可讀性模型訓練完成後，將測試資料輸入至可讀性模型，可讀性模型便會自動預測該文本的可讀性。整體實驗利用 Keras[25]進行實作，流程如圖四所示。



圖四、可讀性模型訓練及測試流程圖

(三)、實驗結果

實驗結果如表二所呈，利用相乘方式所融合兩種表徵學習法的可讀性模型準確率為 79.42%，皆優於單一表徵學習法的可讀性模型，對於卷積神經網路而言準確率多了 3.77%，快速文本的準確率則微幅提升了 0.97%。若進一步將放寬上、下一個年級的標

準來統計出相鄰準確率，以觀察可讀性模型錯誤預測的程度是否嚴重。從表二也可以發現以相乘方式融合的可讀性模型的相鄰準確率為 91.59%，仍是所有可讀性模型中相鄰準確率最高的。

此外，本研究亦將過去研究[13]成功應用於國語科文本的一般語言特徵利用深層類神經網路來訓練可讀性模型，以觀察其模型分析跨領域文件的能力為何。從表二的結果可以發現性其準確率與本研究所提出的可讀性模型相比差了 38.09%。而這樣子的結果也符合國外可讀性研究所得出的結論：一般語言特徵的確無法有效適用於評估特定領域文本的可讀性。

表二、基於卷積神經網路及快速文本之可讀性模型效能比較

適用年級	適用領域	可讀性模型	融合方式	準確率	相鄰準確率
1-12 年級	國語科、社會科、自然科共計 4,648 篇	卷積神經網路		75.65%	89.33%
		快速文本		78.45%	90.95%
		卷積神經網路 + 快速文本	相加	74.89%	87.93%
		卷積神經網路 + 快速文本	相乘	**79.42%**	**91.59%**
		卷積神經網路 + 快速文本	串聯	76.40%	90.30%
		卷積神經網路 + 快速文本	平均	74.78%	89.55%
		一般語言特徵 + 深層神經網路		41.33%	69.11%

五、結論

有鑑於過去的可讀性研究大多只用單一演算法來訓練可讀性模型，本研究利用類神經網路來融合卷積神經網路及快速文本兩種表示學習法，以訓練出一個能夠分析跨領域文件的可讀性模型。實驗結果顯示，雖然利用快速文本就已經可以達到不錯的模型效能，但若可以融合卷積神經網路，則不論是準確率或相鄰準確率都可以往上再提升。這個現象初步顯示在訓練可讀性模型的過程中，若可以找到適當的融合方式，是可以交織出更優質的特徵來提升可讀性模型的效能。在未來我們將以此為基礎，除了嘗試更多複雜的表徵學習演算法來進行橫向的融合外，也將進一步探討與縱向融合之間差異為何。

參考文獻

[1]E. Dale and J. S. Chall, "The concept of readability," *Elementary English*, vol. 26, pp. 19–26, 1949.

[2]G. R. Klare, "Measurement of Readability," 1963.

[3]G. R. Klare, "The measurement of readability: useful information for communicators," *ACM Journal of Computer Documentation (JCD)*, vol. 24, pp. 107-121, 2000.

[4]G. H. McLaughlin, "SMOG grading: A new readability formula," *Journal of reading*, vol. 12, pp. 639–646, 1969.

[5]B. A. Lively and S. L. Pressey, "A method for measuring the vocabulary burden of textbooks," *Educational administration and supervision*, vol. 9, pp. 389–398, 1923.

[6]M. Vogel and C. Washburne, "An objective method of determining grade placement of children's reading material," *The Elementary School Journal*, pp. 373–381, 1928.

[7]J. S. Chall and E. Dale, *Readability Revisited: The new Dale-Chall Readability Formula*, Brookline Books, 1995.

[8]A. C. Graesser, M. Singer, and T. Trabasso, "Constructing inferences during narrative text comprehension," *Psychological Review*, vol. 101, pp. 371, 1994.

[9]K. Collins-Thompson, "Computational assessment of text readability: A survey of current and future research," *International Journal of Applied Linguistics*, vol. 165, pp. 97–135, 2014.

[10] B. Bruce, A. Rubin, and K. Starr, "Why readability formulas fail," *IEEE Transactions on Professional Communication*, pp. 50-52, 1981.

[11] S. E. Petersen and M. Ostendorf, "A machine learning approach to reading level assessment," *Computer Speech & Language*, vol. 23, pp. 89–106, 2009.

[12] L. Feng, M. Jansche, M. Huenerfauth, and N. Elhadad, "A comparison of features for automatic readability assessment," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, 2010, pp. 276–284.

[13] Y. T. Sung, J. L. Chen, J. H. Cha, H. C. Tseng, T. H. Chang, and K. E. Chang, "Constructing and validating readability models: the method of integrating multilevel linguistic features with machine learning," *Behavior research methods*, vol. 47, pp. 340–354, 2014.

[14] X. Yan, D. Song, and X. Li, "Concept-based document readability in domain specific information retrieval," in *Proceedings of the 15th ACM international conference on Information and knowledge management*, pp. 540–549, 2006.

[15] A. Borst, A. Gaudinat, C. Boyer, and N. Grabar, "Lexically based distinction of readability levels of health documents," *Acta Informatica Medica*, vol. 16, pp. 72–75, 2008.

[16] H. C. Tseng, B. Chen, and Y. T. Sung, "Exploring the Use of Neural Network based Features for Text Readability Classification," *International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP)*, vol. 22, pp. 31–46, 2017.

[17] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” *IEEE transactions on pattern analysis and machine intelligence,* vol. 35, pp. 1798–1828, 2013.

[18] Y. LeCun, “Generalization and network design strategies,” *Connectionism in perspective,* pp. 143–155, 1989.

[19] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning (adaptive computation and machine learning series)*. MIT Press, 2016.

[20] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *Journal of machine learning research,* vol. 15, pp. 1929–1958. 2014.

[21] Nair, Vinod, and Geoffrey E. Hinton. “Rectified linear units improve restricted boltzmann machines,” in *Proceedings of the International Conference on Machine Learning*, pp. 807–814. 2010.

[22] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, “Bag of tricks for efficient text classification,” *arXiv preprint arXiv:1607.01759*. 2016.

[23] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781,* 2013.

[24] T. H. Chang, Y. T. Sung, and Y. T. Lee, “A Chinese word segmentation and POS tagging system for readability research,” in *Proceedings of the Annual Meeting of the Society for Computers in Psychology*, 2012.

[25] F. Chollet, “Keras: Deep learning library for theano and tensorflow. *URL: https://keras.io*.” 2015.