This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2022.3171690, IEEE Transactions on Knowledge and Data Engineering

Learning from Ideography and Labels: A Schema-aware Radical-guided Associative Model for Chinese Text Classification

Hanqing Tao, Guanqi Zhu, Enhong Chen, *Senior Member, IEEE,* Shiwei Tong, Kun Zhang, Tong Xu, Qi Liu, *Member, IEEE* and Yew-Soon Ong, *Fellow, IEEE*

Abstract—Reading psychology believes text comprehension to involve a complex psychological construction process, with the reader mind being a dynamic associative system that stores an abundance of schemata. For Chinese text, in particular, the unique ideographic writing system allows its lansign to trigger semantic association and schema recalling without the need of phonetics. In contrast to previous research efforts on text classification problems, in this paper we present an interdisciplinary modeling approach that draws inspirations from the cognitive principles of ideography, schema theory and deep learning to study Chinese text classification. Specifically, we first propose a Radical-guided Associative Model (RAM) for preliminary cognitive imitation, which comprises two coupled spaces, namely the Literal Space and Associative Space. Then, taking consideration of the schemata acquired from the mind of a reader which plays an important role in influencing text-dependent information revision, we extend RAM with a systematic Schema-aware Radical-guided Associative Model (SRAM) that embeds label semantics as essential text-independent human knowledge for real-world abstraction. In SRAM, the Schema Space is introduced and a Schema Attention module is proposed with a novel loss paradigm that includes the linkage and interaction between text-dependent prior concepts and text-independent label schemata. Extensive experiments on three real-world datasets demonstrate the effectiveness and rationality of our proposed method.

Index Terms—Chinese Text Classification, Ideography, Association Mechanism, Schema Theory, Label Embedding.

1 Introduction

Text classification is a fundamental natural language processing (NLP) task, which plays an indispensable role in various data mining scenarios, such as document retrieval, news filtering, public opinion analysis and so on [1], [2]. Traditional text classification methods usually pay attention to theory investigations and feature modeling based on Latin scripts and perform well in English or other phonetic languages, but the performance and cognitive principles behind Chinese language materials are either mediocre or still unexplored. We might hardly imagine the reason for this phenomenon is mainly due to the unique ideographic writing system of Chinese. To bridge this gap and offer an innovative Chinese text classification modeling perspective, we in this paper will deeply delve into the essence of ideographic characteristics, and present a holistic Chinese text modeling framework by combining ideography, schema theory with deep learning together.

Most of the time, when people receive a certain text, they will not only grasp it according to the literal features of the text, but also expand a series of associations in their minds based on those features [3]. Meanwhile, their acquired knowledge will also subconsciously affect the text meaning judgement. This is a fundamental reason why human beings can still maintain a strong generalization ability in complex language environments [4]. In fact, the language symbols in a text that we can directly obtain or perceive are literal features. Especially for Chinese, its writing system derived from pictographs makes its literal features ideographic [5]. Moreover, as the semantic component used to compose *Phono-semantic Compound Characters* [6] which take up over 80% of all Chinese characters, each radical has a pictorial glyph origin which is depicted in Figure 1. This vivid feature has been inherited for thousands of years, often allowing

Glyph Origin	Radical (Chinese Characters)	English Words
	亻 (仆, 伴)	Man (servant, partner)
	目 (看, 瞳)	Eye (look, pupil)
	扌 (打, 挖)	Hand (hit, dig)
	足 (路, 踢)	Foot (road, kick)
	雨 (雾, 霜)	Rain (fog, frost)
	山 (峰, 崖)	Mountain (peak, cliff)
	辶 (道, 追)	Walk (path, chase)

Fig. 1. Oracle bone inscriptions of ideographic radicals, and the ubiquitous semantic connection between Chinese Phono-semantic Compound Characters and radicals. These inherited collective semantics are important schemata for reader's text comprehension subconsciously.

- *Hanqing Tao, Guanqi Zhu, Enhong Chen, Shiwei Tong, Tong Xu and Qi Liu are with the Anhui Province Key Laboratory of Big Data Analysis and Application, School of Computer Science and Technology, University of Science and Technology of China, Hefei, Anhui 230027, China.*
E-mail: {hqtao, zgq, tongsw}@mail.ustc.edu.cn, {cheneh, tongxu, qiliuql}@ustc.edu.cn
- *Kun Zhang is with the School of Computer Science and Information Engineering, Hefei University of Technology (HFUT), Hefei, Anhui 230009, China. E-mail: zhang1028kun@gmail.com*
- *Yew-Soon Ong is with the School of Computer Science and Engineering, Nanyang Technological University (NTU), Singapore. E-mail: asysong@ntu.edu.sg*

Corresponding author: Enhong Chen.

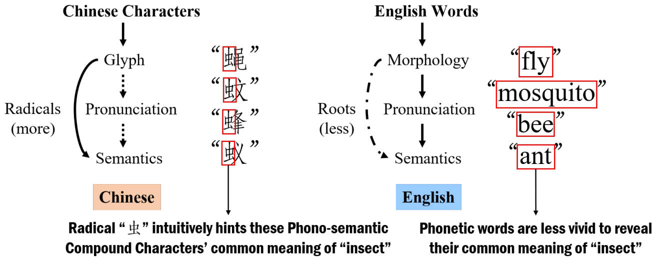


Fig. 2. Comparison of the main semantic conveying mode between Chinese characters and English words. Radical carrying intuitive prior concepts is the essence of ideography, while Latin words are often phonetic substitutes for abstract concepts.

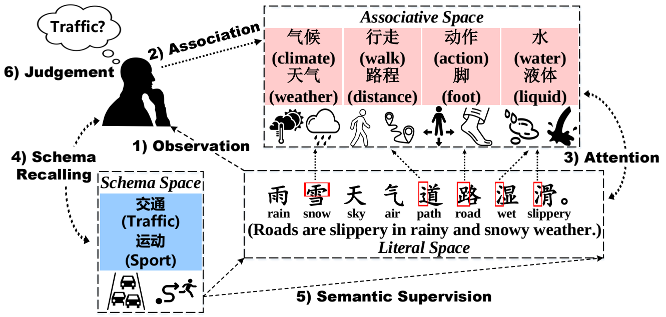


Fig. 3. The cognitive process when understanding a Chinese text and inferring its possible label, where radical of each Phono-semantic Compound Character is circled in red. Many highly relevant concepts and schemata could be derived via association and schema recalling due to the ideographic property of Chinese.

readers to understand the meaning of Chinese characters without knowing their pronunciation, which forms a unique cognitive process in the mode of conveying semantics compared with English and other phonetic languages (Figure 2 gives an intuitive comparision) [7]. Furthermore, as shown by the toy text classification example in Figure 3, ideographic radicals prompt us to associate prior concepts with corresponding Chinese characters: "*climate*" for "snow", "*foot*" for "road", "*water*" for "slippery", etc., which help us grasp relevant attributes of characters. At the same time, our acquired experience and knowledge also enable us to be aware of the candidate semantics in advance: "Traffic": "people or goods transported by road, air, train, or ship"; "Sport": "all types of physical activity that people do to keep healthy or for enjoyment" [8]. Combining the above thinking process, we may comprehensively evaluate and commit the judgement of classification label "Traffic".

Looking into the cognitive principles behind, on the one hand, association in psychology refers to the psychological connection between concepts, events or mental states, usually derived from specific experiences [9]. It allows people to use prior concepts outside a given text to assist comprehension during reading, which is quite ubiquitous in Chinese text due to its ideographic visual irritants. Actually, associative behavior is a fundamental and effective principle in psycho-linguistics for explaining examples of cognition and knowledge learning through accumulated experience [10]. More importantly, since language involves complex human physiological activities, language research is inseparable from cognitive science [11], and more and more researchers have regarded language learning as a cognitive phenomenon [12], [13]. However, based on these interdisciplinary theories of psychology and cognitive science, how to leverage association mechanism to import desired human prior concepts into Natural Language Processing (NLP) is an urgent problem for current deep learning, which faces great challenges.

Unfortunately, traditional text modeling methods often ignore the participation of human associative behavior in the process of text comprehension, just sticking to the analysis of the literal space in isolation to deal with the linguistic symbols [14]. This perspective is very limited now, especially for short texts whose literal features are very sparse [15]. Therefore, introducing some external information reasonably to enrich text representation is more in line with human cognition. Fortunately, as a treasure-house of human knowledge, language dictionaries (e.g., Xinhua Dictionary and Oxford Dictionary) are very efficient for inferring basic information of radicals (roots), characters, words and common concepts to help understand texts in our daily life [16], which leaves a valuable way for us to further improve text representation and enhance classification.

To achieve the goal of imitating readers' cognitive behavior with addressing the association mechanism in our preliminary work [17], we first proposed a novel **R**adical-guided **A**ssociative **M**odel (**RAM**) for Chinese text classification, which can take both literal features and associative prior concepts into consideration with the help of language dictionaries. Specifically, we first introduced a *Literal Space* and devised a serialized structure to model the sequential information of Chinese text. Then, we proposed an *Association* module and a strategy of using radicals as the medium for *Radical-Word Association*, so as to model text-dependent associative contents in *Associative Space*. Afterwards, we designed an *Associative Attention* module by imitating the cognitive process in people's mind to model the matching and decision between *Literal Space* and *Associative Space*.

On the other hand, regarding the cognitive principles behind Figure 3, schema theory is an explanation of how readers use prior knowledge to comprehend and learn from text, and the previously acquired knowledge structures are called schemata [18], [19]. The psychological term schema is a concept similar to common sense, but it has its unique characteristics since schema theory assumes that written text does not carry meaning by itself [20]. Rather, a text only provides directions for readers as to how they should retrieve or construct meaning from their own previously acquired knowledge [21]. For example, "Traffic" is a scheme of affair schema structure, which includes driver, vehicles, road and the knowledge related to "Traffic". When the human sensory system receives a textual message in Figure 3, much related knowledge and definition in the schema net structure will be activated and the schema is used to explain or supervise some particular plots, so the relevant knowledge will be particularized by some specific information. This particularization process is the so-called comprehension process [22]. Hence, we could see that schema recalling is a key factor of human text comprehension, which plays a fundamental role when we try to comprehend unfamiliar things by mobilizing prior experience. However, how to combine "schema" with deep learning strategies and achieve the modeling is also a challenging problem.

Therefore, we may find that the design and architecture of RAM are somewhat inadequate, which does not have the capability of modeling such text-independent schema semantics to help text classification. Also, RAM cannot reflect the important role that human accumulated experience and schema recalling play in the supervision of semantic understanding. In fact, only after we have successfully used the experience to supervise, compare and summarize new things, can we form a holistic cognitive process in our mind [22]. Thus, it is valuable if we could remind RAM about this finding to supervise the overall text modeling. In response to the challenges and schema principles mentioned above, we in this paper extend RAM and propose a more human-like and more systematic **S**chema-aware **R**adical-guided **A**ssociative **M**odel (**SRAM**) by incorporating label semantics as important human prior schemata for real-world abstraction and supervising the whole modeling. The main contributions of our work are summarized as follows:

- We extend the overall perspective of text comprehension by formalizing schema theory as a label learning problem in deep learning and propose a novel *Schema Space*, where each label will be mapped into a formally defined textual description in Oxford Dictionary Chinese version.
- We design a new *Schema Attention* module so that the descriptions of labels will be interacted with contextual information to act as important text-independent prior schemata to improve the literal features learning.
- Given that contents in Associative Space and Schema Space are actually text-dependent and text-independent prior concepts respectively, and the two kinds of prior concepts should be mutually reinforcing, we finely devise a new loss paradigm which can jointly evaluate their similarity and supervise the whole modeling.
- Finally, we conduct extensive experiments on three datasets, where the experimental results not only demonstrate the effectiveness and rationality of SRAM but also provide good cognitive and interdisciplinary insights for future language modeling.

2 Related Work

Text Classification & Deep Learning. Text classification is a fundamental natural language processing (NLP) task, which plays an indispensable role in various scenarios, such as document retrieval, news filtering, public opinion analysis [1], [2]. Recent years have witnessed the success of deep learning in this field, no matter in terms of the construction of deep classification model [23], [24] or word embedding approaches including CBOW, Skipgram, GloVe and so on [25], [26], [27]. Given the sequential property of human language, Recurrent Neural Network (RNN) [28], its improved version Long Short-Term Memory (LSTM) [29] and Bidirectional LSTM (BiLSTM) have been proposed for capturing the long-range information of the context [30], which has a profound effect on the subsequent study of text modeling. Currently, there has been another wave in the field of natural language processing, that is the emerging model of pre-training [31]. Among them, the most successful model might be BERT [32], which combines Transformer's powerful representation ability with some language-related pre-training goals to address many NLP tasks while exhibiting impressive performance [33], [34].

Human Cognitive Modeling. No matter in the early days or now, imitating human cognitive principles has always been the original intention of deep learning [35], [36]. Initially, the fully-connected edges designed in artificial neural networks ideally mimic the numerous dendrites of nerve cells. Then, to improve the nonlinear expression ability of neural networks, the activation functions (e.g., sigmoid and ReLU) were proposed by imitating the activation threshold of biofilm action potential [37], [38]. More importantly, the attention mechanism [39], [40] was proposed to mimic the fact of eye allocation when people are reading a text or observing an image [41], which exhibits superior performance and psychological interpretability at the same time [42].

Label Embedding. As an abstract summary of the commonality of similar things in the real world, labeling plays an important role in supervised learning and can guide the training of models [43]. In truth, the powerful connection between language and cognition in humans begins in infancy, and decades of research has revealed that labeling and naming can facilitate infants' classification ability and help them know the world [44]. Currently, the development of Artificial Intelligence is also in its infancy, and many modeling methods imitating human principles still need to be developed [45]. Rationally incorporating label's semantics to text modeling, so as to let the model know what is the real meaning of training targets, is more in line with our intuition [46], [47]. Inspiringly, label embedding has indeed been shown to be effective and yield performance improvement in various domains (such as natural language processing and computer vision) recently [48], [49].

Chinese-specific Methods. In recent years, the human brain investigations about the differences between Chinese and phonetic languages have prompted researchers to explore the uniqueness of Chinese lansigns [50], [51]. Scholars have also found that Chinese is a highly analytic language with flexible expressions [52], and the low-level features of Chinese characters such as radical [53], pinyin (Wang et al. [54]), stroke [55] and glyph [56] also have certain semantics. By introducing them into word or sentence representation learning, the performance can indeed be improved. At the same time, for the study of Chinese downstream tasks, a proper text modeling method can highlight the characteristics of Chinese, which is an important factor to improve the performance [57], [58]. Lately, Tao et al. [59] have achieved impressive results by directly introducing radicals to participate in Chinese text representation and classification.

3 Problem Overview

Here, we introduce the text classification problem studied in this work and give it a formal definition. Given an arbitrary unlabeled text $T = \{x_1, x_2, ..., x_m\}$ and a pre-defined set of labels V, the goal of our task is to train and obtain a classification function $\mathcal{F}$ with the ability to assign a proper label $l \in V$ for T:

$$\mathcal{F}(T) \to l, \tag{1}$$

where $x_i \in T$ $(0 \le i \le m)$ stands for a feature token in T after text preprocessing.

Generally, existing methods tend to regard word or character as feature token x_i respectively or together. However, the semantics of each label in V and the inherent semantical relations conveyed in a Chinese word or character via radicals are regularly ignored in this way. To achieve this

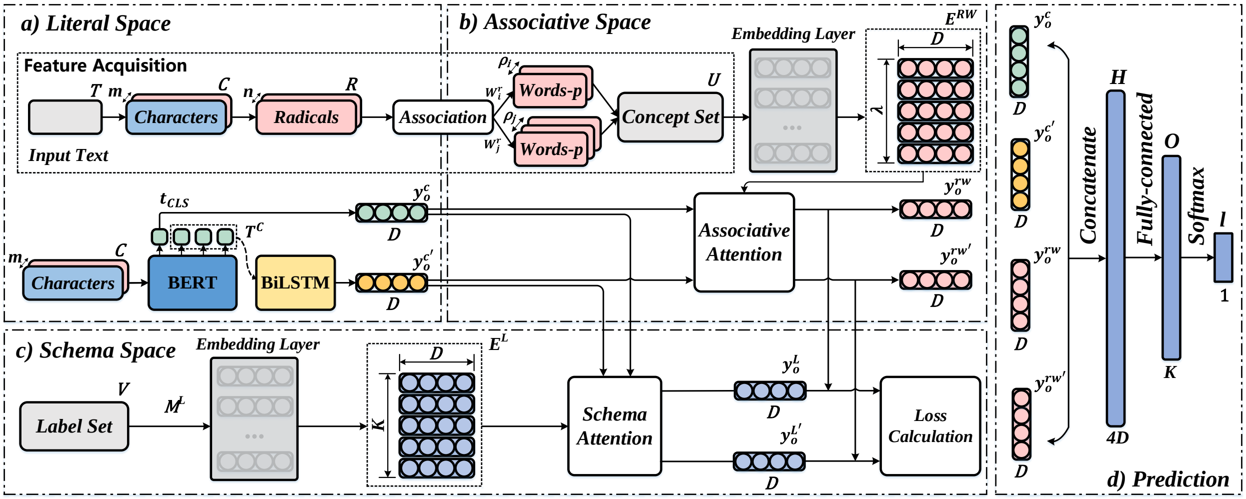


Fig. 4. The overall architecture of the proposed SRAM Model, where RAM (*Literal Space* and *Associative Space*) is part of SRAM.

goal, we propose RAM and SRAM by exploiting the usage of label descriptions and different x_i together in a human-like fashion, including Chinese character, radical and word.

4 Methodology

In this section, we first introduce the implementation of our **R**adical-guided **A**ssociative **M**odel (**RAM**) that could directly achieve the primary goal of association mechanism modeling. Then, we extend RAM to the proposed **S**chema-aware **R**adical-guided **A**ssociative **M**odel (**SRAM**) by elaborating how to incorporate label semantics as important human prior schemata to make the Chinese text classification framework more solid and systematic.

4.1 RAM Model

As shown in Figure 4, RAM is part of our proposed SRAM model. Specifically, RAM mainly comprises two coupled spaces (i.e., *Literal Space* and *Associative Space*, together with one *Feature Acquisition* process and two modules, namely *Association* and *Associative Attention*. The technical details of each part will be presented as follows.

4.1.1 Feature Acquisition

Before delving into the details of each component, it is necessary for us to understand how the features required by our model are obtained. In line with the cognitive behavior of people observing a text and acquiring features shown in Figure 3, we first propose the *Feature Acquisition* process illustrated by Figure 5 to extract the features required for our model, i.e., characters, radicals and words. In fact, there are six kinds of Chinese characters according to "six writings"[1], but only the radicals of *Phono-semantic Compound Characters* contain informative semantics [6]. Therefore, in this paper, we mainly pay attention to *Phono-semantic Compound Characters* and their radicals. Then, we will use these radicals to obtain corresponding associative words with the help of three Chinese language dictionaries.

1) Character Type Masking. As intuitively depicted in Figure 5, given an input Chinese text T containing m characters, we first segment it into a character sequence $C = \{c_1, c_2, ..., c_m\}$ according to the string operation, where C actually stands for the character-level feature of T. Then, by referring to Chinese Character Type Dictionary [60], we are able to label each character with a type tag so as to realize the *Character Type Masking* process:

$$Mask(c_i) = \begin{cases} 1 & c_i = C_p, \\ 0 & c_i = Others, \end{cases} \tag{2}$$

where C_p represents *Phono-semantic Compound Characters*. $Mask(\cdot)$ denotes the masking function, and c_i $(0 \leq i \leq m)$ is the i-th character in C.

2) Radical Distilling. After getting the mask code of each character, we could carry out the following *Radical Distilling* process. That is to say, in order to extract the radicals that have significant ideographic effects from text T (i.e., radicals of *Phono-semantic Compound Characters*) and thus help to convey semantics, we need to remove other useless contents. So, we multiply each character's mask code with itself to determine which characters could retain for querying radicals from Chinese dictionary:

$$R = Radical\_Query(C \odot Mask(C)), \tag{3}$$

where $\odot$ is an element-wise product operation, and $Radical\_Query$ operation allows us to map each Chinese character into a single radical with the help of Xinhua Dictionary [61]. Additionally, we filter out all the repeated radicals in R to avoid redundant processing. As a result, $R = \{r_1, r_2, ..., r_n\}$ is the distilled radicals of character sequence C, where $n \in [0, m]$.

3) Radical-Word Association. Instead of using radicals directly as an additional feature [59], we regard the distilled radicals as the medium for associating highly relevant associative words that indicate attributes and extensional meaning. Formally, we call this strategy *Radical-Word Association*, which corresponds to the *Association* module in Figure 4. As a result, associative words connected with *Phono-semantic Compound Characters* are denoted as *Words-p* (red). By referring to Radical Concept Dictionary [62], each distilled radical $r_i \in R = \{r_1, r_2, ..., r_n\}$ will correspond to a list of associative words:

$$W_i^r = Concept\_Query(r_i) = \{w_1^r, w_2^r, ..., w_{\rho_i}^r\}. \tag{4}$$

[1] https://en.wikipedia.org/wiki/Chinese_characters

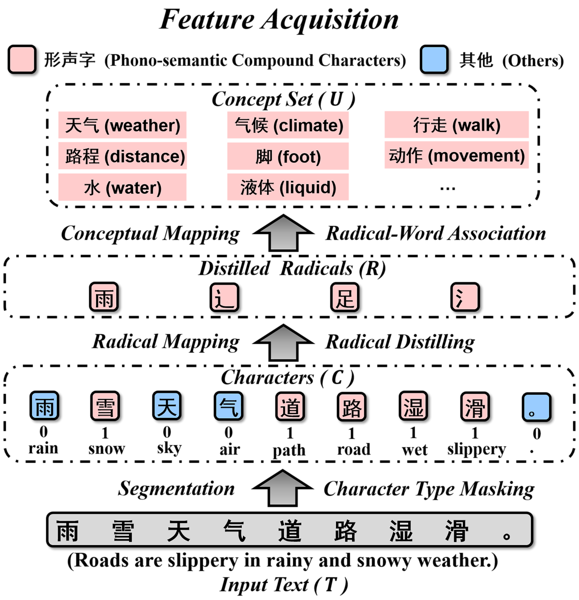


Fig. 5. An intuitive and detailed illustration of the *Feature Acquisition* process for Chinese text.

Here, $\rho_i \geq 1$ denotes the number of associative words for r_i, which will vary from radical to radical. Therefore, all the radicals R extracted from text T could form a set of associative words $U = \{w_1, w_2, ..., w_\lambda\}$:

$$U = \bigcup_{i=1}^{n} W_i^r, \tag{5}$$

where U actually stands for the imported external word-level feature for T and $\lambda = \sum_{i=1}^{n} \rho_i$ denotes the total words number of U. Since different radicals may correspond to the same associative words, the set operation here allows repeated associative words to be merged into one.

4.1.2 Literal Space Modeling.

Given an input Chinese text T containing m characters, RAM will literally project it into a character sequence $C = \{c_1, c_2, ..., c_m\}$ for subsequent processing (each punctuation will also be regarded as a character). Then, we devise a deep modeling structure by harnessing the power of pre-trained BERT (Bidirectional Encoder Representations from Transformers) [32], which has embraced abundant "*accumulated experience*" (statistical language information) based on very large training materials [10], to obtain the sentence representation $t_{CLS} \in \mathcal{R}^{1 \times D}$ and character representation $T^C = \{t_1, t_2, ..., t_m\}$ of T as follows:

$$t_{CLS},\ T^C = BERT([CLS], C), \tag{6}$$

where the first token $[CLS]$ added in front of every sequence C is always a special classification token, and the final hidden state t_{CLS} corresponding to this token is used as the aggregate sequence representation for classification tasks. Because t_{CLS} acts as the output of BERT for later classification, we also use y_o^c to denote it for convenience.

Meanwhile, T^C represents the hidden vectors of corresponding m characters contained in text T. Then, we send the hidden states together as a new sequence into BiLSTM to further learn the context dependencies, which is depicted in Figure 6. Formally, we take the rest output of BERT, i.e., $T^C = \{t_1, t_2, ..., t_m\}$, as the sophisticated representations for each character c_i $(1 \leq i \leq m)$ in C. Afterwards, we apply BiLSTM to further imitate the *conceptual change* [63] under the specific context of T^C, which is consistent with the process of people adjusting to a new text based on their accumulated experience. Thus, given the vector embedding sequence of BERT output T^C, the hidden vectors of BiLSTM are calculated by receiving T^C as input:

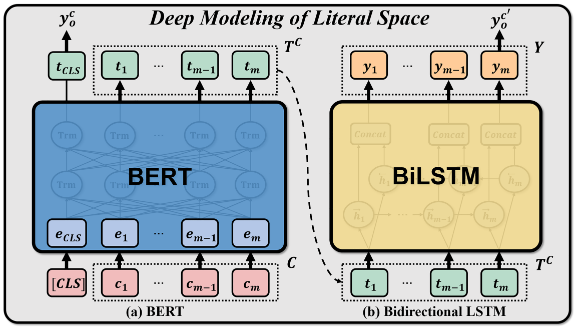


Fig. 6. Diagrammatic sketch of literal space modeling.

$$\begin{aligned}
\overrightarrow{h_i} &= LSTM(\overrightarrow{h}_{i-1}, s_i), \\
\overleftarrow{h_i} &= LSTM(\overleftarrow{h}_{i+1}, s_i), \\
y_i &= concatenate(\overrightarrow{h_i}, \overleftarrow{h_i}),
\end{aligned} \tag{7}$$

where $\overrightarrow{h_i}$ and $\overleftarrow{h_i}$ denote the forward hidden vector and backward hidden vector respectively at the i-th time step s_i $(1 \leq i \leq m)$ in the BiLSTM unit. While y_i is the concatenation of $\overrightarrow{h_i}$ and $\overleftarrow{h_i}$. As a result, the final output of BiLSTM (i.e., y_m) will integrate the forward and backward contextual information. For convenience, we also use $y_o^{c'}$ to denote it for subsequent calculation.

4.1.3 Associative Space Modeling.

As mentioned above, the ideographic characteristics of Chinese characters are deeply rooted and ubiquitous [50], which is a crucial factor for readers to associate relevant concepts with radicals. Now that we have obtained associative words through the *Association* module described in *Feature Acquisition* process, we should further represent those words and highlight the information that we need.

1) Associative Word Embedding. In order to represent the associative words in the concept set $U = \{w_1, w_2, ..., w_\lambda\}$ for subsequent calculation, we need to map each word into a low-dimension real-value vector. Here, we apply an external well pre-trained embedding model based on distributional assumption [25], [28] and an *Embedding Layer* to get the embedding vectors for associative words obtained by *Radical-Word Association*:

$$E^{RW} = Embedding(U) = \{e_1^{rw}, e_2^{rw}, ..., e_\lambda^{rw}\}, \tag{8}$$

where λ denotes the total associative words number of U.

2) Associative Attention Module. The attention mechanism in deep learning is essentially similar to the selective visual attention mechanism of human beings. In fact, as for reading comprehension, people usually tend to first read through the sentence to form a preliminary cognition in their minds, and then back to select and match the proper concepts based on the overall context of the sentence [64]. Inspired by this cognitive process, we design an *Associative Attention* module which can focus our model on relatively important associative words in U back with the consideration of learned contextual representation explained before, i.e., y_o^c and $y_o^{c'}$.

Formally, we regard y_o^c and $y_o^{c'}$ as *query*s, E^{RW} as *key* and *value* at the same time to implement attention mechanism. That is, given the associative word representations obtained in *Associative Space*, i.e., $E^{RW} = \{e_1^{rw}, e_2^{rw}, ..., e_\lambda^{rw}\}$, we use the contextual representations obtained in *Literal Space*, i.e., y_o^c and $y_o^{c'}$, to attend to each associative word $w_i \in U$ and get the attention weight for each $e_\epsilon^{rw} \in E^{RW}$ and $e_\theta^{rw} \in E^{RW}$ $(1 \leq \epsilon \leq \lambda, 1 \leq \theta \leq \lambda)$:

$$\begin{aligned} &\alpha^{'} = [\alpha_1^{'}, ..., \alpha_\epsilon^{'}, ..., \alpha_\lambda^{'}],\ \alpha_\epsilon^{'} = f(y_o^c, e_\epsilon^{rw}), \\ &\beta^{'} = [\beta_1^{'}, ..., \beta_\theta^{'}, ..., \beta_\lambda^{'}],\ \beta_\theta^{'} = f(y_o^{c'}, e_\theta^{rw}), \end{aligned} \tag{9}$$

where $\alpha^{'} \in \mathcal{R}^{1\times\lambda}$ and $\beta^{'} \in \mathcal{R}^{1\times\lambda}$ are two vectors for E^{RW} respectively, representing the attention weight from two contextual aspects of y_o^c and $y_o^{c'}$. Besides, $\alpha_\epsilon^{'}$ and $\beta_\theta^{'}$ denote the ϵ-th or the θ-th weight of an associative word respectively, and $f(\cdot,\cdot)$ denotes the distance function which is stated as an element-wise dot product operation in this paper. Then, we need to normalize $\alpha^{'}$ and $\beta^{'}$ with the $softmax$ function:

$$\begin{aligned} \alpha_i &= \frac{exp\ (\alpha_i^{'})}{\sum_{\epsilon=1}^{\lambda} exp\ (\alpha_\epsilon^{'})},\ where \sum_{i=1}^{\lambda} \alpha_i = 1, \\ \beta_j &= \frac{exp\ (\beta_j^{'})}{\sum_{\theta=1}^{\lambda} exp\ (\beta_\theta^{'})},\ where \sum_{j=1}^{\lambda} \beta_j = 1. \end{aligned} \tag{10}$$

Afterwards, the two-aspect attentive representations y_o^{rw} and $y_o^{rw'}$ for associative words could be obtained through attentive weighted sum as:

$$y_o^{rw} = \sum_{\epsilon=1}^{\lambda} \alpha_\epsilon e_\epsilon^{rw},\ y_o^{rw'} = \sum_{\theta=1}^{\lambda} \beta_\theta e_\theta^{rw}, \tag{11}$$

where α_ϵ is the ϵ-th dimensional value of $\alpha \in \mathcal{R}^{1\times\lambda}$, and β_θ is the θ-th dimensional value of $\beta \in \mathcal{R}^{1\times\lambda}$ $(1 \leq \epsilon \leq \lambda,$ $1 \leq \theta \leq \lambda)$. Consequently, the attentive representations y_o^{rw} and $y_o^{rw'}$ have precisely fused the information of *Literal Space* and *Associative Space* together. Then, for better illustrating the parallel design in the following Schema Space part, we choose to formalize the attention mechanism described above and denote Equation (9) and (10) together as an integrated *Attention* operation. In another words, given the input query (y_o^c and $y_o^{c'}$), key (E^{RW}) and value (E^{RW}), we could rewrite Equation (9) and (10) with its output y_o^{rw} and $y_o^{rw'}$ in Equation (11):

$$\begin{aligned} y_o^{rw} &= Attention(y_o^c, E^{RW}, E^{RW}), \\ y_o^{rw'} &= Attention(y_o^{c'}, E^{RW}, E^{RW}). \end{aligned} \tag{12}$$

4.2 SRAM Model

In our RAM model, we take radicals as a valuable guide to text-dependent prior concepts based on readers' associative behavior when reading Chinese text. Through the *Radical-Word Association* strategy and the *Associative Attention* module, RAM can rationally integrate and balance literal and associative information of Chinese text, while perfectly avoiding the hidden adverse effect of wrong Chinese word segmentation (CWS) [65]. However, due to the inability to clarify the connection and evaluate the interaction between associative concepts and reader's active thinking system, the modeling structure of RAM is not yet sufficient.

Imagine that we readers are now given an unfamiliar text: 1) Have we already acquired some text-independent knowledge in our minds before we see the text? 2) Will we comprehend the text under the supervision of our previously formed knowledge schema? The answers to both questions are of course "Yes".

In fact, as we have noted previously in the principle of "comprehension process", when readers' sensory system receives a textual stimulus, much related knowledge and definition in his/her schema net structure will be activated and the schema is used to explain some particular plot and supervise related judgement [22]. Correspondingly, in terms of supervised learning and text classification tasks, the essence of "labeling" is to guide the model to correctly learn the potential correlation between text and labels [43], where the labels are often defined as a kind of "single word" summary description and reflect some real-world collective semantics. Hence, we may find that the description of labels is just in line with the function of schema in readers' mind. As for the Chinese language shown in Figure 1 and Figure 2, its ideographic literal features actually act as a kind of visual and textual irritant at the same time, the activation of schema recalling is more easily and frequently compared with phonetic lansigns [7].

In view of the facts above, we hold that Chinese text representation and classification will be much more rational and solid if the definition or description of labels could be formally taken into consideration. Thus, it is valuable if we could remind RAM about this text-independent information so that our model could better prepare the target training for text classification. Therefore, inspired by psychological research findings, we extend RAM model and propose a novel *Schema Space* to systematically bring label semantics into Chinese text modeling. To be specific, our newly proposed SRAM model contains three extra modules, i.e., *Schema Space Modeling*, *Schema Attention* and *Loss Calculation*. The details will be elaborated as follows.

4.2.1 Schema Space Modeling

1) Label Embedding. In the literature, most existing models are generally trained on a fixed label set using k-hot vectors, and therefore treat target labels as mere numeric symbols without any particular semantic connection to the space of texts [66]. To cope with this limitation, we exploit the usage of dictionary description and pre-trained BERT encoder in deep learning together, so that the labels could be well embedded with input text into the same space.

Formally, given a label set $V = \{l_1, l_2, ..., l_K\}$ containing K different labels depicted in *Problem Overview* and Figure 4, we manually query the most appropriate one description of each label l_i from Oxford Dictionary in Chinese version:

$$W_i^l = Entry\_Query(l_i) = \{w_1^l, w_2^l, ..., w_{\rho_i}^l\}, \tag{13}$$

where $\rho_i \geq 1$ denotes the number of words for label l_i in dictionary description, and ρ_i will vary from label to label. Thus, all the descriptions of labels in V could form a sentence matrix $\mathcal{M}^L$:

$$\mathcal{M}^L = \{W_1^l, W_2^l, ..., W_K^l\}, \tag{14}$$

where $\mathcal{M}^L$ actually stands for the imported text-independent schemata for input text T.

After the query operation, each label would be assigned a textual description $W_i^l \in \mathcal{M}^L (1 \leq i \leq K)$. Then, in order to rationally embed labels V with input text T in the same space, an effective way is to apply a pre-trained BERT model to represent the textual descriptions $\mathcal{M}^L$ so as to achieve the label embedding:

$$E^L = BERT(\mathcal{M}^L) = \{e_1^l, e_2^l, ..., e_K^l\}, \tag{15}$$

where $e_i^l \in \mathcal{R}^D$ is the description embedding of label l_i.

2) Schema Attention Module. According to schema theory, comprehending a text is an interactive process between reader's background knowledge and the text [22]. In other words, people hearing the speech, seeing the text is not equal to accepting the language, only when the textual symbols and the acquired knowledge schemata produce a meaningful match is the real principle of language understanding. Inspired by these facts, we design a *Schema Attention* module which is symmetrical to *Associative Attention* for imitating the interaction process between literal and schema spaces.

Before the implementation of schema attention mechanism, we need to be aware of the purpose for imitation and function of *query*, *key* and *value*. Here, we take y_o^c and $y_o^{c'}$ as *query*s to function as external text stimuli, E^L as *key* and *value* at the same time to function as internal schema correspondence. That is, given the label representations, i.e., $E^L = \{e_1^l, e_2^l, ..., e_K^l\}$, we use the contextual representations obtained in *Literal Space*, i.e., y_o^c and $y_o^{c'}$, to attend to each label embedding $e_i^l \in E^L$:

$$\begin{aligned} y_o^L &= Attention(y_o^c, E^L, E^L), \\ y_o^{L'} &= Attention(y_o^{c'}, E^L, E^L). \end{aligned} \tag{16}$$

Through the calculation above, the attentive representations y_o^L and $y_o^{L'}$ have incorporated the information of *Literal Space* and *Schema Space* together. Like the information processing in human brain, the attention mechanism herein is actually a bridge between the two spaces.

3) Loss Calculation Module. In *Associative Attention* and *Schema Attention* module, we have detailed the necessity of the attention mechanism for its function of information screening and weighting. But three non-negligible questions we may ask are: 1) "Will the schemata and label embeddings directly help enrich the representation of input text?" 2) "How do the schemata and label embedding manifest in the final classification task?" 3) "What is the relation between associative prior concepts and label schemata?"

For solving these doubts, we think deeply about the human thinking and inference process when comprehending the text. As depicted in Figure 3, we may conclude that the schemata activated by labels are independent of input text, which is a kind of invariant abstraction for certain collective semantics to some degree (e.g., label "Traffic" actually embraces a quite stable knowledge structure relevant to it). Thus, we hold that y_o^L and $y_o^{L'}$ are playing a role of global supervision and discriminator for semantical judgement, and the semantics of labels are affecting the whole text comprehension indirectly. While for y_o^{rw} and $y_o^{rw'}$ obtained in *Associative Space*, it is obvious that these associative schemata are dependent of input text, which will help enrich input text representation directly. Furthermore, due to the parallel design and same goal of enhancing text comprehension, the text-dependent prior concepts and text-independent schemata should have similar semantical direction and enjoy mutual enhancement respectively (i.e., y_o^{rw} should be similar to y_o^L, $y_o^{rw'}$ should be similar to $y_o^{L'}$).

Therefore, driven by the analysis above, and as a correspondence to deep learning strategy, we devise a new loss paradigm to allow different loss functions to evaluate and constrain the needful similarity between the aforementioned associative prior concepts and label schemata:

$$\mathcal{L}^S = loss_x(y_o^{rw}, y_o^L) + loss_x(y_o^{rw'}, y_o^{L'}), \tag{17}$$

where $loss_x$ denotes a certain loss function (we choose cosine similarity loss [67] as the default setting for SRAM model), and the performance rendered by different $loss_x$s will be discussed in subsequent Experiments section. Through this way, the two groups of embeddings in *Associative Space* and *Schema Space* will be viewed as a joint modeling process throughout the training of our SRAM model.

4.2.2 Prediction

As in the literal, associative and schema spaces described earlier, we have already obtained six different representations for Chinese text T: two-aspect contextual representations learned through literal features, i.e., y_o^c and $y_o^{c'}$; attentive representations derived from jointly modeling of literal features and associative concepts, i.e., y_o^{rw} and $y_o^{rw'}$; and attentive representations derived from jointly modeling of literal features and label schemata, i.e., y_o^L and $y_o^{L'}$. Since the semantics of labels tend to assist us in making classification decisions indirectly, we choose to directly combine the first four semantic representations to conduct prediction, i.e., y_o^c, $y_o^{c'}$, y_o^{rw} and $y_o^{rw'}$, while leveraging y_o^L and $y_o^{L'}$ to supervise the prediction via *Schema Attention* and *Cosine-loss Calculation* mechanism. In order to systematically integrate and fully learn the information of the four representations, we first conduct a concatenation operation:

$$H = [y_o^c;\ y_o^{c'};\ y_o^{rw};\ y_o^{rw'}], \tag{18}$$

where $H \in \mathcal{R}^{1 \times 4D}$ is the vector concatenated through dimension with an advantage of retaining all the information [42], [68]. Afterwards, we leverage the superior fitting ability of the fully connected neural network to learn the hidden interactions and enhancements among these four representations:

$$\begin{aligned} O &= \sigma(\boldsymbol{W} \times H + \boldsymbol{b}), \\ \sigma(x) &= \frac{1}{1 + e^{-x}}, \end{aligned} \tag{19}$$

where $\boldsymbol{W}$ and $\boldsymbol{b}$ respectively denotes the weight matrix and bias vector fitted by the linear neural network, and $O \in \mathcal{R}^{1 \times K}$ is its output. Note that K represents the size or scale of label set V which has been stated in *Problem Overview*. Finally, the classification could be conducted through the $softmax$ function and $argmax$ operation as:

$$l = argmax(softmax(O)). \tag{20}$$

As a result, the predicted label l, which corresponds to the dimension index with the maximum value, would be assigned to input text T.

4.3 Training Strategy

Currently, we may have a holistic understanding of the architecture of our SRAM model. Next, we will discuss its learning and training strategies, including *Loss Function* and *Model Initialization*.

Loss Function. As the multi-class classification task exhibits an output of distribution with different probabilities on various classes, we need to judge the most significant class and distinguish it from others. According to previous work [57], [59], cross-entropy is a good way to scale the significance of probability distribution for classification:

$$\mathcal{L}^E = -\sum_{T \in Corpus} \sum_{i=1}^{K} p_i(T) \log p_i(T), \tag{21}$$

Additionally, we take into account the analysis and Cosine-loss described in *Schema Space*, and devise a new loss function as follows:

$$Loss = \mu \mathcal{L}^E + (1-\mu)\mathcal{L}^S, \tag{22}$$

where μ is an important hyperparameter to trade-off the two kinds of losses, which will be detailedly discussed in the following Experiments section. In this way, we could find that the cross-entropy loss is the main role responsible for the final classification, while the proposed cosine similarity loss is an effective means to balance the interaction between text-dependent concepts and text-independent schemata.

Model Initialization. Before training, a proper initialization is beneficial for optimizing our model. In conjunction with the architecture of RAM and SRAM, we apply the pre-trained Chinese BERT model with 24-layer, 1024-hidden, 16-heads and 330M parameters[2] to support the literal space modeling and label embedding, so that the labels could be embedded with input text into the same space as we need [66]. In addition, for ensuring the embedding heterogeneity between associative prior concepts and label schemata, hence better verifying the effectiveness of our proposed new loss paradigm, we select a well pre-trained Chinese word embedding model[3] based on distributed assumptions [25] with a dimension of 256 to represent associative words in U (Because homogeneous pre-trained embedding will naturally reduce the impact of $\mathcal{L}^S$). The embedding dimension D is also set as 256, with fully connected neural network for dimension transformation if necessary according to Figure 4, while the hidden size of BiLSTM is set as 1,024. To prevent our model from overfitting, we add a dropout mechanism in front of the embedding layer and fully-connected layer with a drop rate of 0.5. As for the scale size λ of concept set U, it is a variable which will be adapted to Radical Concept Dictionary dynamically. Besides, we apply orthogonal initialization for the parameters of BiLSTM and *Xavier* initialization for the fully connected neural network, and the training epoch of SRAM and RAM is set as 20. Finally, we apply *Adagrad* optimizer with a learning rate of 0.01, and use *Pytorch* to build our model and train it with two 2.30GHz Intel(R) Xeon(R) Gold 5218 CPUs and a Tesla V100-SXM2-32GB GPU.

TABLE 1
The statistics of three real-world datasets.

Datasets		Count	Length (Avg. / Max)	Number of Class
CNT	Train	47,693	17.8/56	32
	Test	15,901	17.8/56	
FCT	Train	8,220	16.3/75	20
	Test	8,115	16.2/64	
TNT	Train	287,007	22.7/150	15
	Test	95,664	22.7/146	

5 Experiments

5.1 Dataset Description

To fit the problems studied in this paper, we carefully select three real-world datasets to evaluate our model with different emphases: the Chinese News Title Dataset (CNT), Fudan Chinese Text Dataset (FCT) and Toutiao News Text Dataset (TNT). The original statistics of them are shown in Table 1. To avoid randomness as much as possible, we merge and shuffle all data for each dataset and conduct 5-fold cross-validation for all comparison methods.

- **CNT** [57] is a public dataset which covers a wide range of 32 different categories of Chinese news. After preprocessing the useless text whose length is lower than 2, it contains 47,693 texts for training and 15,901 for testing, which is suitable for validating the generalization ability of different methods.
- **FCT**[4] is an official dataset provided by Fudan University with 20 categories covering abundant academic texts for validation. To guarantee the quality of implementation, we carefully preprocessed this dataset by correcting and removing unreadable samples. As a result, it contains 8,220 texts for training and 8,115 for testing. Due to the inherent imbalanced samples between different classes, this dataset is a good choice to evaluate the stability and robustness of different methods.
- **TNT**[5] is a public Chinese dataset which covers 15 different categories of Chinese news. It contains 287,007 texts for training and 95,664 for testing, whose scale and volume are quite larger than the former two datasets. Therefore, this dataset is beneficial for validating the comprehensiveness of different methods.

5.2 Dictionary Preparation

In order to guarantee the reliability of our model, we apply three formal Chinese Dictionary datasets to support the process of *Character Type Masking*, *Radical Mapping* and *Conceptual Mapping* process. In addition, we manually annotate the labels in each dataset and query the corresponding label descriptions and definitions from the online Oxford Dictionary in Chinese version[6], which could consequently form a Schema Dictionary. Accordingly, Chinese Character Type Dictionary[7] contains all the information about *Phono-semantic Compound Characters*; Xinhua Dictionary[8] contains the necessary radical information for mapping each character to a radical; Radical Concepts Dictionary[9] includes detailed conceptual information for all Chinese radicals,

[2] https://github.com/ymcui/Chinese-BERT-wwm
[3] https://spaces.ac.cn/archives/4304
[4] https://www.kesci.com/home/dataset/5d3a9c86cf76a600360edd04
[5] https://www.kesci.com/mw/dataset/5dd645fca0cb22002c94e65d/file
[6] https://dictionary.cambridge.org/zhs/
[7] http://zidian.kxue.com/
[8] http://zidian.aies.cn/
[9] http://xh.5156edu.com/page/z2443m7618j19616.html

with over 1,000 concept words in total and 6 concept words for each radical on average; Since the Schema Dictionary is relevant to specific datasets, according to Table 1, there would be 32 label descriptions for CNT dataset, 20 for FCT dataset and 15 for TNT dataset.

5.3 Benchmark Methods

To comprehensively evaluate the performance of our model, we finely select 13 benchmark text classification methods from three perspectives, and the number of them are also noted in Table 2: a) basic deep learning models (1-5); b) recently proposed non-Chinese-specific outstanding models (6-10); c) models considering the effect of Chinese-specific feature granularity (11-13).

- **TextRNN (word/char)** [28] refers to the plain recurrent neural network which processes tokens sequentially. To compare the functionality of Chinese feature granularity in different scenarios, we set character-level and word-level feature as the input respectively.
- **TextCNN (word/char)** [24] is a convolutional neural network-based model for text classification. With the same aim of comparison like TextRNN, the input is also set as two kinds, i.e., character-level and word-level. We apply *jieba*[10] as the segmentation tool to obtain the word-level feature.
- **FastText (word)** [69] is a quite simple but effective model which applies the average of word/n-grams embeddings to achieve text representation and classification.
- **TextGCN (word)** [70] is a widely used graph convolutional network for text classification. Its core idea is to build a graph based on the co-occurrence of words and texts, and then utilize the convolutional operation to capture high-order neighborhoods information.
- **HyperGAT (word)** [71] is a state-of-the-art graph-based network for text classification, which models texts with text-level hypergraphs. Hence, the high-order interaction between words could be captured.
- **LEAM (word)** [46] is a representative label embedding model, which introduces label descriptions and learns embeddings of words and labels in the same space to enhance text representations. To seek common ground while being different from LEAM, our SRAM model employs a new strategy of utilizing label information, which is more suitable and rational for Chinese text classification.
- **BERT (char)** [32] stands for the current state-of-the-art pre-training model for natural language processing, which is usually applied in English materials and performs well. We here take it as an important baseline and fine-tune it to validate the rationality and effectiveness of the design of our SRAM model.
- **LCM (char)** [72] is a recently proposed model enhancement framework for general text classification. We take it to enhance BERT as a contrast benchmark.
- **C-LSTMs/C-BLSTMs (char+word)** [57] are two Chinese-specific text classification models applying two independent LSTMs to concatenate word and character features together. Since both characters and words are important features for Chinese text, they make up for the disadvantages of using one kind of feature unilaterally. And C-BLSTMs is the bidirectional version.
- **RAFG (char+word+radical)** [59] is another Chinese-specific text classification method. This state-of-the-art baseline is a four-granularity model, which integrates two extra kinds of radicals (character-level and word-level) together with corresponding Chinese characters and words to directly help Chinese text classification.

5.4 Experimental Results

The comparison results on three datasets are shown in Table 2, from which we can see that our SRAM model is able to substantially achieve the best results on all datasets, no matter in terms of Accuracy, Recall or F1-score. However, there are still some thought-provoking findings.

5.4.1 Main Results

Firstly, by comparing Chinese-specific methods (11-15) with those non-Chinese-specific ones (1-10), we could notice that the feature granularity of Chinese text is a crucial factor for classification performance. Models using character-level features consistently perform better than word-level ones, which is due to the high recall brought by countable Chinese characters. And utilizing character or word features unilaterally is worse than combing them together, which proves that Chinese characters and words can make up for each other and Chinese word segmentation may cause loss of information unavoidably. In the meantime, we could infer from Table 2 that although BERT-based models (9-10) only take Chinese characters as main input, they can maintain a robust performance on all datasets (more stable Recall and F1 on FCT dataset whose samples are quite inbalanced), which confirms that after large-scale corpus pre-training, single character-level features can effectively adapt to the high recall property of Chinese characters and obtain better robustness when faced with different corpora. All these granularity-relevant findings are quite consistent with the study in [65]. Secondly, the significant performance of label embedding model LEAM (8) compared with recently proposed outstanding models (6-10, graph-based and pre-trained benchmarks) also inspires us, where the utilization of label semantics can yield considerable performance improvement and rationality at the same time. Thirdly, looking back on our modeling of the three features in Chinese (character, radical and word), we can find that RAM and SRAM only use the character features of Chinese text literally, and meanwhile, the word features are associated via the medium of radical instead. This process perfectly avoids the adverse effects of Chinese word segmentation errors, which plays a non-negligible role in promoting the performance of RAM and SRAM. Fourthly, the results are clear that RAM and SRAM have a comprehensive improvement in performance compared with the most advanced pre-trained BERT model, which shows that our modeling strategy based on cognitive principles can better grasp the purport of Chinese text. Last but not least, we might know that radical is a special low-level Chinese feature which does not possess the property of "context", so the way of RAFG directly integrating radicals with Chinese characters and words context is imperfect and taking every radical into consideration is a little improper. Therefore, through the comparison between RAFG, RAM and SRAM, we could learn that a more rational method of utilizing radicals is very beneficial for better understanding hence harnessing the messages conveyed by radicals. Overall, the results in Table 2 demonstrate that our proposed RAM and SRAM can substantially outperform benchmark methods on diverse dataset distributions.

[10] https://github.com/fxsjy/jieba

TABLE 2
Experimental results (average Accuracy, Recall and F1-score with standard deviation under 5-fold cross-validation) of different methods on CNT, FCT and TNT dataset ("c" represents character-level feature, "w" represents word-level feature, "r" represents radical-level feature).

Model (granularity)	CNT			FCT			TNT		
	Accuracy	Recall	F1-score	Accuracy	Recall	F1-score	Accuracy	Recall	F1-score
(1) TextCNN (w)	0.7560±0.003	0.7562±0.004	0.7569±0.003	0.8849±0.008	0.5149±0.016	0.5340±0.022	0.8496±0.002	0.7837±0.003	0.7848±0.002
(2) TextCNN (c)	0.7574±0.005	0.7575±0.004	0.7588±0.005	0.9091±0.003	0.6480±0.012	0.6926±0.017	0.8479±0.002	0.7830±0.002	0.7860±0.001
(3) TextRNN (w)	0.6951±0.005	0.6952±0.005	0.6967±0.006	0.8449±0.005	0.4714±0.019	0.4727±0.023	0.8447±0.001	0.7806±0.002	0.7799±0.001
(4) TextRNN (c)	0.7051±0.008	0.7050±0.007	0.7071±0.007	0.8685±0.012	0.4940±0.011	0.4775±0.011	0.8471±0.003	0.7842±0.003	0.7835±0.003
(5) FastText (w)	0.7349±0.003	0.7383±0.003	0.7393±0.003	0.8653±0.006	0.5034±0.021	0.5030±0.024	0.8410±0.001	0.7771±0.001	0.7791±0.002
(6) TextGCN (w)	0.7696±0.002	0.7693±0.002	0.7689±0.001	0.8602±0.006	0.6060±0.019	0.6520±0.026	0.8707±0.001	0.8036±0.002	0.8070±0.005
(7) HyperGAT (w)	0.7692±0.003	0.7691±0.003	0.7672±0.003	0.8980±0.004	0.6866±0.014	0.6795±0.012	0.8582±0.001	0.8141±0.004	0.8055±0.003
(8) LEAM (w)	0.7991±0.002	0.7991±0.002	0.7984±0.002	0.9320±0.003	0.7217±0.016	0.7646±0.017	0.8641±0.001	0.7989±0.002	0.7992±0.002
(9) BERT (c)	0.8082±0.003	0.8077±0.002	0.8081±0.002	0.9073±0.003	0.7390±0.015	0.7778±0.017	0.8608±0.001	0.7968±0.002	0.7975±0.002
(10) LCM (c)	0.8203±0.011	0.8203±0.009	0.8206±0.010	0.9374±0.003	0.7911±0.034	0.8104±0.023	0.8835±0.001	0.8197±0.002	0.8204±0.003
(11) C-LSTM (c+w)	0.8128±0.003	0.8135±0.003	0.8123±0.002	0.9139±0.003	0.6897±0.012	0.7007±0.009	0.8557±0.001	0.7996±0.001	0.8021±0.001
(12) C-BLSTM (c+w)	0.8178±0.004	0.8145±0.002	0.8160±0.004	0.9172±0.002	0.6893±0.011	0.7046±0.012	0.8566±0.001	0.8024±0.001	0.8046±0.002
(13) RAFG (c+w+r)	0.8253±0.007	0.8256±0.007	0.8253±0.007	0.9139±0.004	0.6981±0.011	0.7164±0.013	0.8630±0.001	0.8041±0.002	0.8066±0.002
(14) RAM (c+w+r)	0.8497±0.002	0.8500±0.002	0.8499±0.002	0.9428±0.002	0.8018±0.025	0.8272±0.027	0.8866±0.003	0.8256±0.003	0.8252±0.003
(15) SRAM (c+w+r)	**0.8550±0.003**	**0.8549±0.002**	**0.8549±0.002**	**0.9448±0.002**	**0.8176±0.020**	**0.8386±0.025**	**0.8888±0.003**	**0.8304±0.002**	**0.8332±0.002**

5.4.2 *Ablation Results*

As mentioned earlier, the designed components of SRAM are solidly based on the cognitive principles between ideography and reading psychology [13]. In order to validate the design of SRAM and determine how each part affects the final results, we conduct an ablation study by removing each module respectively or compositionally, which is summarized in Table 3.

According to the results of six ablation variants, we can observe a consistent performance decline compared with SRAM on all three datasets no matter which module is removed. That is to say, all modules designed in our model contribute a certain degree to performance improvement. Meanwhile, the changes in ablation performance reflected by the three datasets are quite different, namely, different dataset distributions rely on and focus on different modules. Consistent with the purpose of our initially selecting the dataset, for the CNT dataset, the performance of SRAM drops obviously when schema space or associative space is removed. It verifies that the concept words in associative space and label semantics in schema space are indispensable aids to enhance the generalization ability for multi-class Chinese text classification. In the meantime, we can notice that the performance of SRAM on the FCT dataset decreases severely when the associative attention and schema attention module are removed, which shows that simply introducing prior information without filtering and weighting will be harmful to robustness in turn. Furthermore, as for the TNT dataset, there is a significant drop when removing schema space, which means that a proper trade-off between the two spaces interaction (i.e., associative and schema spaces) and holistic classification is necessary for ensuring comprehensiveness. Although some variants on TNT dataset are somewhat more accurate, they cause a significant recall decline after the removal of particular modules. Therefore, the F1-score that embodies the comprehensive performance better illustrates the stability of SRAM. Altogether, all the ablation results could lead us to a conclusion that each module of SRAM is indispensable to achieving excellent performance for Chinese text classification.

5.5 Loss Paradigm Study

In Section Introduction and Section 4.2, we have discussed the inadequacy of the previous RAM model, and left an open choice of loss functions in our proposed Loss Calculation module. In a word, our core idea of entending RAM to SRAM is to let our model obtain as much human prior knowledge as possible about the semantic understanding of input text within a reasonable range, while guaranteeing a balance between text-dependent information (i.e., y_o^{rw}, $y_o^{rw'}$) and text-independent information (i.e., y_o^{L}, $y_o^{L'}$). Therefore, we need to determine some suitable similarity/loss functions to achieve this goal. To comprehensively evaluate the design of our model, we further conduct a loss paradigm study in this section. Specifically, we choose three loss functions, i.e., cosine similarity [67] (default for SRAM model), KL-divergence (Kullback-Leibler divergence) [73] and Mean Square Error (MSE). The experimental results are shown in Table 4, from which we can find that the default setting for SRAM (SRAM-Cos) serves as a moderate benchmark. At the same time, SRAM-KL fits better on the CNT dataset, while SRAM-MSE performs the best on FCT and TNT datasets.

5.6 Hyperparameter Study

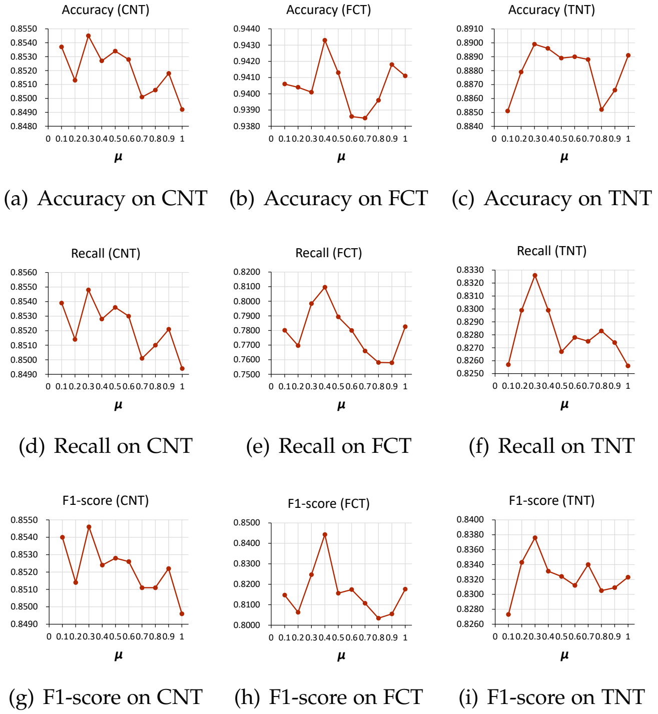
To validate the design of our proposed loss paradigm and explore the hidden correlation between text-dependent prior concepts and text-independent schemata, we make μ increase from 0 to 1 to see how it affects the final performance of SRAM (Since the classification effect of cross-entropy loss will totally disappear when $\mu = 0$, so we have $\mu \in (0, 1]$). The experimental results of adjusting parameter μ are shown in Figure 7. Firstly, from this figure, we could notice an obvious changing trend of rising first and then falling rather than invariant when μ increases from 0 to 1, which indicates that combing the interactions between two spaces and global prediction properly is crucial for achieving better classification performance. Secondly, we can find that when the value of μ reaches 1 ($Loss = \mathcal{L}^E$), the comprehensive performance of SRAM (F1-score) becomes either mediocre or the worst on three datasets. This phenomenon is due to the cancellation of our proposed similarity constraints for text-dependent and text-independent information, which effectively proves the necessity of the proposed loss paradigm $\mathcal{L}^S$. Besides, more informatively, we can also see the peak values of nine curves in Figure 7 correspond to a consistent value range of μ from 0.2 to 0.4 in all three datasets, which means that the weights of $\mathcal{L}^S$ occupy 60% to 80%. That is to say, our proposed loss

TABLE 3
Average ablation performance of SRAM, where "as." denotes "associative" and "sc." denotes "schema" for simplification.

Model	CNT			FCT			TNT		
	Accuracy	Recall	F1-score	Accuracy	Recall	F1-score	Accuracy	Recall	F1-score
(1) w/o as. & sc. attention	0.8516±0.002	0.8516±0.002	0.8517±0.002	0.9418±0.004	0.7841±0.023	0.8077±0.025	0.8903±0.002	0.8265±0.002	0.8279±0.001
(2) w/o associative attention	0.8520±0.002	0.8521±0.001	0.8522±0.001	0.9415±0.003	0.7991±0.029	0.8191±0.021	0.8889±0.001	0.8276±0.001	0.8266±0.002
(3) w/o schema attention	0.8527±0.003	0.8530±0.002	0.8527±0.002	0.9425±0.003	0.7894±0.020	0.8139±0.022	0.8890±0.002	0.8266±0.002	0.8269±0.001
(4) w/o as. & sc. space	0.8499±0.002	0.8500±0.002	0.8498±0.002	0.9423±0.001	0.7969±0.021	0.8239±0.022	**0.8906±0.002**	0.8265±0.002	0.8263±0.002
(5) w/o schema space	0.8497±0.002	0.8500±0.002	0.8499±0.002	0.9428±0.002	0.8018±0.025	0.8272±0.027	0.8866±0.003	0.8256±0.003	0.8252±0.003
(6) w/o label description	0.8524±0.002	0.8525±0.002	0.8525±0.002	0.9420±0.002	0.8038±0.021	0.8250±0.019	0.8900±0.002	0.8279±0.002	0.8281±0.003
(7) SRAM	**0.8550±0.003**	**0.8549±0.002**	**0.8549±0.002**	**0.9448±0.002**	**0.8176±0.020**	**0.8386±0.025**	0.8888±0.003	**0.8304±0.002**	**0.8332±0.002**

TABLE 4
Average performance of SRAM applying different loss functions in the Loss Calculation module.

Model	CNT			FCT			TNT		
	Accuracy	Recall	F1-score	Accuracy	Recall	F1-score	Accuracy	Recall	F1-score
(1) SRAM-Cos	0.8550±0.003	0.8549±0.002	0.8549±0.002	**0.9448±0.002**	0.8176±0.020	0.8386±0.025	0.8888±0.003	0.8304±0.002	0.8332±0.002
(2) SRAM-KL	**0.8557±0.004**	**0.8559±0.003**	**0.8557±0.003**	0.9447±0.002	0.8080±0.029	0.8356±0.027	**0.8890±0.002**	0.8294±0.002	0.8315±0.002
(3) SRAM-MSE	0.8549±0.003	0.8551±0.003	0.8551±0.003	0.9426±0.004	**0.8208±0.028**	**0.8395±0.027**	0.8883±0.003	**0.8312±0.004**	**0.8341±0.004**



(a) Accuracy on CNT (b) Accuracy on FCT (c) Accuracy on TNT

(d) Recall on CNT (e) Recall on FCT (f) Recall on TNT

(g) F1-score on CNT (h) F1-score on FCT (i) F1-score on TNT

Fig. 7. Performance of our SRAM model on three real-world datasets when hyperparameter μ is ranging from 0 to 1 ($\mu \in (0, 1]$).

paradigm plays a salient and effective role in boosting the holistic classification performance.

5.7 Efficient Analysis

To show SRAM is effective not only in learning knowledge but also in operating efficiency, we further conduct an efficient analysis in this section. To be specific, since RAM and SRAM are both BERT-based models, we select BERT as the best benchmark pre-trained model for the ease of control variates method. The corresponding results are shown in Table 5. Clearly from the results, we can find that the average inference time of the three models is very fast, with a minor difference. Although the total training time of RAM and SRAM is longer, the convergence is much faster than BERT. At the same time, SRAM is able to achieve a quite lower validation loss on all datasets, which is considerably due to our proposed loss paradigm. Thus, we can get the conclusion that our SRAM model has superior ability and efficiency to tackle Chinese-specific features across different corpora, which demonstrates the effectiveness of the proposed SRAM framework again. Because all models reach their lowest loss before 20 epochs, we keep this setting for each model's training on three datasets.

TABLE 5
Average runtime of BERT, RAM and SRAM, where the "Train_time" means total training time (20 epochs). "Conv_time" and "Conv_loss" means the average convergence time and convergence validation loss.

Model		BERT	RAM	SRAM
CNT	Train_time	26m41s	38m59s	43m0s
	Conv_time	23m7s	18m58s	24m46s
	Conv_loss	0.5977	0.5040	**0.1881**
FCT	Train_time	6m27s	11m28s	11m49s
	Conv_time	5m37s	4m58s	6m5s
	Conv_loss	0.3440	0.2257	**0.1618**
TNT	Train_time	256m49s	306m12s	346m5s
	Conv_time	231m46s	194m18s	232m48s
	Conv_loss	0.4772	0.3829	**0.1316**
Inference time (ms/text)		1.33	2.43	2.47

5.8 Case Study

To provide some intuitionistic examples for explaining why our model gains a better performance than any other baseline methods, we conduct a case study similar with [74] to see what is happening in the working flow of SRAM, where the specific cases could be found in Table 6. Taking the first example to say, we notice that the associative words and literal features can enhance each other, i.e., associative words "*plant*" and "*agriculture*" associated by SRAM are important clues for inferring the concept of "Eggplant", while other associative words (e.g., "*action*" suggests the attribute of "salvation", and "*liquid*" indicates the property of "sauce") could be regarded as complementary contents for source text thus helping us grasp less prominent but global semantics. Then, for the second example, we could find that the label description in dictionary enables SRAM to be aware of the semantics of "Dress" in advance, and associative words "*vegetation*", "*material*" and "*hair*" globally reflect the trait of "fur", while "*condition*", "*property*" and "*time*" together help us recognize the semantics of "fashion" hence lead us to the idea of "Dress". As for the BERT model, due to its inability to understand the label semantics of "Dress", hence the prediction is disturbed by certain characters in the input text.

TABLE 6
A case study for some Chinese source texts, where the Phono-semantic Compound Characters are all painted in red.

Source Text	Distilled Radicals	Ground Truth, Label Description in Oxford Dictionary Chinese version & English	Top 2 Associative Words for each Distilled Radical (sorted by attention weights)	Model Prediction	
				BERT	SRAM
Chinese: 拯救夏日胃口的肉菜：茄汁里脊 **English**: Salvation of Summer Appetite: Tenderloin in Eggplant Sauce.	扌、攵、艹、氵	**Label**: Food **Description**: 人和动物吃的东西，或植物吸收的东西来维持它们的生存 **English**: something that people and animals eat, or plants absorb, to keep them alive	**Distilled Radicals**：扌、攵、艹、氵 **Chinese**: (动作、行为)、(做、行为)、(植物、农业)、(液体、水) **English**: (action, behavior)、(do, behavior)、(plant, agriculture)、(liquid, water)	Food (✓)	Food (✓)
Chinese: 皮草早春混搭新时髦 **English**: The new fashion of mixing and matching fur in early spring.	艹、氵、扌、斤、日、髟	**Label**: Dress **Description**: (尤指适用于某种场合的)服装 **English**: used, especially in combination, to refer to clothes of a particular type, especially those worn in particular situations	**Distilled Radicals**：艹、氵、扌、斤、日、髟 **Chinese**: (草木、材料)、(状态、液体)、(动作、行为)、(性质、工具)、(时间、太阳)、(毛发、胡须) **English**: (vegetation, material)、(condition, liquid)、(action, behavior)、(property, tool)、(time, sun)、(hair, beard)	Star (×)	Dress (✓)
Chinese: 利用城市污泥防治水土流失 **English**: Applying urban sludge to prevent and control soil erosion.	刂、土、氵、阝	**Label**: Environment **Description**: 人、动物和植物居住的空气、水和陆地 **English**: the air, water, and land in or on which people, animals, and plants live	**Distilled Radicals**：刂、土、氵、阝 **Chinese**: (工具、动作)、(土地、建筑)、(流体、水)、(山地、地形) **English**: (tool, action)、(soil, building)、(fluid, water)、(mountainous region, topography)	Economy (×)	Environment (✓)
Chinese: 短跑运动员专项力量练习的设计与选择 **English**: Design and Selection of Special Strength Practice for Sprinters.	矢、足、辶、力、页、纟、讠、扌	**Label**: Sports **Description**: 根据规则需要体力劳动和技能的游戏、比赛或活动 **English**: a game, competition, or activity needing physical effort and skill that is played or done according to rules	**Distilled Radicals**：矢、足、辶、力、页、纟、讠、扌 **Chinese**: (长度、度量)、(脚、动作)、(行走、路程)、(力量、行为)、(颈部、数量)、(纺织、行为)、(交流、语言)、(手、动作) **English**: (length, measurement)、(foot, action)、(walk, distance)、(strength, behavior)、(neck, number)、(weave, behavior)、(communication, language)、(hand, action)	Sports (✓)	Sports (✓)

Moreover, as for the remaining two examples, we could also observe that the prior information provided by associative words and label semantics are mostly informative indicators for determining the ground truths. Although there might be some associative words which are not directly related to the semantics of ground truths (e.g., "*liquid*" for "Dress" and "*weave*" for "Sports"), those words actually reflect the original meaning of corresponding radicals, which will be balanced under the *Associative Attention* module and may be helpful in another context. In fact, when we humans are associating related concepts and recalling relevant schema to help text comprehension in our minds, we tend to think of all possible meanings. This is similar to the unconscious iceberg effect [75], [76], i.e., although some associative contents seem to be irrelevant to the classification ground truth of the current text, the sufficient associative information and priorly learned label semantics is actually a key hidden factor to grasp original meanings of characters and ensure the understanding robustness. In summary, all the above findings could finally enable us to confirm the rationality and effectiveness of our model.

6 Conclusion

In this paper, we conducted an explorative but focused study on Chinese text classification from an interdisciplinary viewpoint of human beings, and proposed a novel **S**chema-aware **R**adical-guided **A**ssociative **M**odel (**SRAM**) for this task. Unlike previous methods which neglect the cognitive principles of language comprehension such as association and schema recalling, SRAM comprises three coupled spaces called *Literal Space*, *Associative Space* and *Schema Space*, which ideally imitates the real process in readers' mind when comprehending a Chinese text. While combining computer science and language-related interdisciplinary theories, our model can balance and correspond to key technologies in the field of deep learning, so that the performance and interpretability of our SRAM model can coexist. Through extensive experiments, our study has gone some way towards enhancing our understanding between ideography, schema theory and human cognition.

7 Acknowledgments

This research was partially supported by grants from the National Natural Science Foundation of China (Grants No. 61727809, 61922073 and U20A20229).

References

[1] A. Hotho, A. Nürnberger, and G. Paaß, "A brief survey of text mining." in *Ldv Forum*, vol. 20. Citeseer, 2005, pp. 19–62.
[2] E. D'Andrea, P. Ducange, A. Bechini, A. Renda, and F. Marcelloni, "Monitoring the public opinion about the vaccination topic from tweets analysis," *Expert Systems with Applications*, vol. 116, pp. 209–226, 2019.
[3] M. Kuvska, R. Trnka, A. A. Kubena, and J. G. Ruvzivcka, "Free associations mirroring self- and world-related concepts: Implications for personal construct theory, psycholinguistics and philosophical psychology," *Frontiers in Psychology*, vol. 7, 2016.
[4] J. Bresnan, A. Cueni, T. Nikitina, and R. H. Baayen, "Predicting the dative alternation," in *Cognitive foundations of interpretation*. KNAW, 2007, pp. 69–94.
[5] J. M. Unger, *Ideogram: Chinese characters and the myth of disembodied meaning*. University of Hawaii Press, 2004.
[6] T. Tung, *The Six Scripts Or the Principles of Chinese Writing by Tai Tung: A Translation by LC Hopkins, with a Memoir of the Translator by W. Perceval Yetts*. Cambridge University Press, 2012.
[7] O. J. Tzeng, D. L. Hung, B. Cotton, and W. S. Wang, "Visual lateralisation effect in reading chinese characters," *Nature*, vol. 282, no. 5738, p. 499, 1979.
[8] O. E. Dictionary, "Oxford english dictionary," *Simpson, Ja & Weiner, Esc*, 1989.
[9] S. B. Klein, *Learning: Principles and applications*. Sage Publications, 2018.
[10] A. Dickinson, "Associative learning and animal cognition," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 367, no. 1603, pp. 2733–2742, 2012.
[11] W. Marslen-Wilson and L. K. Tyler, "The temporal structure of spoken language understanding," *Cognition*, vol. 8, no. 1, pp. 1–71, 1980.
[12] A. G. Huth, W. A. De Heer, T. L. Griffiths, F. E. Theunissen, and J. L. Gallant, "Natural speech reveals the semantic maps that tile human cerebral cortex," *Nature*, vol. 532, no. 7600, p. 453, 2016.
[13] N. C. Ellis, "Essentials of a theory of language cognition," *The Modern Language Journal*, vol. 103, pp. 39–60, 2019.
[14] C. Wang, Y. Fan, X. He, and A. Zhou, "Decoding chinese user generated categories for fine-grained knowledge harvesting," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 8, pp. 1491–1505, 2019.
[15] X. Zhou, X. Wan, and J. Xiao, "Cminer: Opinion extraction and summarization for chinese microblogs," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 7, pp. 1650–1663, 2016.
[16] S. Nielsen, "The effect of lexicographical information costs on dictionary making," *Lexikos*, vol. 18, 2008.
[17] H. Tao, S. Tong, K. Zhang, T. Xu, Q. Liu, E. Chen, and M. Hou, "Ideography leads us to the field of cognition: A radical-guided associative model for chinese text classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 15, 2021, pp. 13 898–13 906.

[18] F. C. Bartlett and F. C. Bartlett, *Remembering: A study in experimental and social psychology*. Cambridge University Press, 1995.

[19] D. E. Rumelhart, "On evaluating story grammars." 1980.

[20] S. A. Widmayer, "Schema theory: An introduction," *Retrieved December*, vol. 26, p. 2004, 2004.

[21] J. Bai, L. Li, D. Zeng, and Q. Li, "Associated activation-driven enrichment: Understanding implicit information from a cognitive perspective," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 12, pp. 2655–2668, 2017.

[22] S. An, "Schema theory in reading." *Theory & Practice in Language Studies*, vol. 3, no. 1, 2013.

[23] C. C. Aggarwal and C. Zhai, "A survey of text classification algorithms," in *Mining text data*. Springer, 2012, pp. 163–222.

[24] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1746–1751.

[25] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[26] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014.

[27] B. Li, A. Drozd, Y. Guo, T. Liu, S. Matsuoka, and X. Du, "Scaling word2vec on big corpus," *Data Science and Engineering*, pp. 1–19, 2019.

[28] T. Mikolov, S. Kombrink, L. Burget, J. Černockỳ, and S. Khudanpur, "Extensions of recurrent neural network language model," in *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2011, pp. 5528–5531.

[29] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[30] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*. Ieee, 2013, pp. 6645–6649.

[31] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365*, 2018.

[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[33] J. Yu and J. Jiang, "Adapting bert for target-oriented multimodal sentiment classification," in *IJCAI*, 2019.

[34] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, "Ernie: Enhanced language representation with informative entities," *arXiv preprint arXiv:1905.07129*, 2019.

[35] J. C. Bezdek, "On the relationship between neural networks, pattern recognition and intelligence," *Int. J. Approx. Reason.*, vol. 6, pp. 85–107, 1992.

[36] S. Sardi, R. Vardi, Y. Meir, Y. Tugendhaft, S. Hodassman, A. Goldental, and I. Kanter, "Brain experiments imply adaptation mechanisms which outperform common ai learning algorithms," *Scientific Reports*, vol. 10, no. 1, pp. 1–10, 2020.

[37] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *The Journal of physiology*, vol. 117, no. 4, p. 500, 1952.

[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[40] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Attention branch network: Learning of attention mechanism for visual explanation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 705–10 714.

[41] Y. Yin, Z. Huang, E. Chen, Q. Liu, F. Zhang, X. Xie, and G. Hu, "Transcribing content from structural images with spotlight mechanism," *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.

[42] K. Zhang, G. Lv, L. Wang, L. Wu, E. Chen, F. Wu, and X. Xie, "Drr-net: Dynamic re-read network for sentence semantic matching," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 7442–7449.

[43] M. Xu, Y.-F. Li, and Z.-H. Zhou, "Robust multi-label learning with pro loss," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 8, pp. 1610–1624, 2020.

[44] A. LaTourrette and S. R. Waxman, "A little labeling goes a long way: Semi-supervised learning in infancy," *Developmental science*, vol. 22, no. 1, p. e12736, 2019.

[45] L. Feigenson and J. Halberda, "Conceptual knowledge increases infants' memory capacity," *Proceedings of the National Academy of Sciences*, vol. 105, no. 29, pp. 9926–9930, 2008.

[46] G. Wang, C. Li, W. Wang, Y. Zhang, D. Shen, X. Zhang, R. Henao, and L. Carin, "Joint embedding of words and labels for text classification," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2321–2331.

[47] H. Peng, J. Li, S. Wang, L. Wang, Q. Gong, R. Yang, B. Li, P. S. Yu, and L. He, "Hierarchical taxonomy-aware and attentional graph capsule rcnns for large-scale multi-label text classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 6, pp. 2505–2519, 2021.

[48] H. Zhang, L. Xiao, W. Chen, Y. Wang, and Y. Jin, "Multi-task label embedding for text classification," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 4545–4553.

[49] C. Du, Z. Chen, F. Feng, L. Zhu, T. Gan, and L. Nie, "Explicit interaction model towards text classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6359–6366.

[50] L. H. Tan, J. A. Spinks, J.-H. Gao, H.-L. Liu, C. A. Perfetti, J. Xiong, K. A. Stofer, Y. Pu, Y. Liu, and P. T. Fox, "Brain activation in the processing of chinese characters and words: a functional mri study," *Human brain mapping*, vol. 10, no. 1, pp. 16–27, 2000.

[51] Y.-h. Hung, D. L. Hung, O. J.-L. Tzeng, and D. H. Wu, "Tracking the temporal dynamics of the processing of phonetic and semantic radicals in chinese character recognition by meg," *Journal of Neurolinguistics*, vol. 29, pp. 42–65, 2014.

[52] C. Wang, X. He, and A. Zhou, "Open relation extraction for chinese noun phrases," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 6, pp. 2693–2708, 2021.

[53] Y. Sun, L. Lin, N. Yang, Z. Ji, and X. Wang, "Radical-enhanced chinese character embedding," in *International Conference on Neural Information Processing*. Springer, 2014, pp. 279–286.

[54] Y. Wang, S. Ananiadou, and J. Tsujii, "Improving clinical named entity recognition in chinese using the graphical and phonetic feature," *BMC Medical Informatics and Decision Making*, vol. 19, 2019.

[55] S. Cao, W. Lu, J. Zhou, and X. Li, "cw2vec: Learning chinese word embeddings with stroke n-gram information," in *AAAI*, 2018.

[56] W. Wu, Y. Meng, Q. Han, M. Li, X. Li, J. Mei, P. Nie, X. Sun, and J. Li, "Glyce: Glyph-vectors for chinese character representations," *arXiv preprint arXiv:1901.10125*, 2019.

[57] Y. Zhou, B. Xu, J. Xu, L. Yang, and C. Li, "Compositional recurrent neural networks for chinese short text classification," in *Web Intelligence (WI), 2016 IEEE/WIC/ACM International Conference on*. IEEE, 2016, pp. 137–144.

[58] H. Peng, E. Cambria, and X. Zou, "Radical-based hierarchical embeddings for chinese sentiment analysis at sentence level," in *The 30th International FLAIRS conference. Marco Island*, 2017.

[59] H. Tao, S. Tong, H. Zhao, T. Xu, B. Jin, and Q. Liu, "A radical-aware attention-based model for chinese text classification," in *AAAI*, 2019.

[60] S. Liang, X. Tang, R. Hu, J. Wu, and Z. Liu, "Measurement and application of chinese component semantic ability based on distributed representation," in *CCL*, 2019.

[61] Z. Han, "Xinhua zidian (xinhua dictionary)," 2009.

[62] J.-F. Hong and C.-R. Huang, "A hanzi radical ontology based approach towards teaching chinese characters," in *Workshop on Chinese Lexical Semantics*. Springer, 2012, pp. 745–755.

[63] N. R. Council *et al.*, *How people learn: Brain, mind, experience, and school: Expanded edition*. National Academies Press, 2000.

[64] N. A. Taatgen, H. Van Rijn, and J. Anderson, "An integrated theory of prospective time interval estimation: The role of cognition, attention, and learning." *Psychological Review*, vol. 114, no. 3, p. 577, 2007.

[65] X. Li, Y. Meng, X. Sun, Q. Han, A. Yuan, and J. Li, "Is word segmentation necessary for deep learning of chinese representations?" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3242–3252.

[66] N. Pappas and J. Henderson, "Gile: A generalized input-label embedding for text classification," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 139–155, 2019.

[67] Y.-S. Lin, J.-Y. Jiang, and S.-J. Lee, "A similarity measure for text classification and clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 7, pp. 1575–1590, 2014.

[68] K. Zhang, H. Zhang, Q. Liu, H. Zhao, H. Zhu, and E. Chen, "Interactive attention transfer network for cross-domain sentiment classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 5773–5780.

[69] A. Joulin, É. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 2017, pp. 427–431.

[70] L. Yao, C. Mao, and Y. Luo, "Graph convolutional networks for text classification," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 7370–7377.

[71] K. Ding, J. Wang, J. Li, D. Li, and H. Liu, "Be more with less: Hypergraph attention networks for inductive text classification," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 4927–4936.

[72] B. Guo, S. Han, X. Han, H. Huang, and T. Lu, "Label confusion learning to enhance text classification models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 14, 2021, pp. 12 929–12 936.

[73] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[74] C. Qin, H. Zhu, T. Xu, C. Zhu, C. Ma, E. Chen, and H. Xiong, "An enhanced neural network approach to person-job fit in talent recruitment," *ACM Transactions on Information Systems (TOIS)*, vol. 38, no. 2, pp. 1–33, 2020.

[75] S. Freud, *The unconscious*. Penguin UK, 2005, vol. 8.

[76] A. Rogers, *The base of the iceberg: Informal learning and its impact on formal and non-formal learning*. Verlag Barbara Budrich, 2014.

Hanqing Tao received the B.S. degree in electrical engineering and automation from China University of Mining and Technology, Xuzhou, China, in 2017. He is currently working toward the Ph.D. degree in the Department of Computer Science and Technology from University of Science and Technology of China (USTC). His research interests include data mining, deep learning, natural language processing, Chinese language analysis and interpretable artificial intelligence. He has published several papers in referred conference proceedings, such as AAAI, ICDM, ICME, CCL, etc.

Guanqi Zhu received the B.E. degree in Computer Science and Technology from the University of Science and Technology of China (USTC), Hefei, China, in 2020. He is currently working toward the MS degree in the Anhui Province Key Laboratory of Big Data Analysis and Application (BDAA), School of Computer Science and Technology, University of Science and Technology of China (USTC), Hefei, China. His research interests include data mining, natural language processing, and machine learning.

Enhong Chen (Senior Member, IEEE) received the Ph.D. degree from the University of Science and Technology of China (USTC), Hefei, China, in 1996. He is currently a director (professor) of the Anhui Province Key Laboratory of Big Data Analysis and Application (BDAA), University of Science and Technology of China (USTC), Hefei, China. He is an executive dean of the School of Data Science, USTC, and vice dean of the School of Computer Science and Technology, USTC. He is a CCF fellow. His research interests include data mining, machine learning, and recommender systems. He has published more than 200 refereed international conference and journal papers. He was the recipient of KDD' 08 Best Application Paper Award and ICDM' 11 Best Research Paper Award.

Shiwei Tong received the B.S. degree in computer science and technology from the Department of Computer Science and Technology from University of Science and Technology of China (USTC), in 2017. He is currently working toward the Ph.D. degree in the Department of Computer Science and Technology from University of Science and Technology of China (USTC). His research interests include educational data mining, deep learning, natural language processing and interpretable artificial intelligence. He has published several papers in referred conference proceedings, such as KDD, IJCAI, ICDM, etc.

Kun Zhang received the PhD degree in computer science and technology from University of Science and Technology of China, Hefei, China, in 2019. He is is currently a faculty member with the Hefei University of Technology (HFUT), China. His research interests include Natural Language Understanding, Recommendation System. He has published several papers in refereed journals and conferences, such as the IEEE Transactions on Systems, Man, and Cybernetics: Systems, the ACM Transactions on Knowledge Discovery from Data, AAAI, KDD, ACL, ICDM. He received the KDD 2018 Best Student Paper Award.

Tong Xu received the Ph.D. degree in University of Science and Technology of China (USTC), Hefei, China, in 2016. He is currently working as an Associate Professor of the Anhui Province Key Laboratory of Big Data Analysis and Application, USTC. He has authored 50+ journal and conference papers in the fields of social network and social media analysis, including IEEE TKDE, IEEE TMC, IEEE TMM, KDD, AAAI, ICDM, etc.

Qi Liu (Member, IEEE) received the Ph.D. degree from the University of Science and Technology of China (USTC), Hefei, China, in 2013. He is currently a professor with the Anhui Province Key Laboratory of Big Data Analysis and Application (BDAA), School of Computer Science and Technology, University of Science and Technology of China (USTC), Hefei, China. His research interests include data mining, machine learning, and recommender systems. He was the recipient of KDD' 18 Best Student Paper Award and ICDM' 11 Best Research Paper Award. He was also the recipient of China Outstanding Youth Science Foundation, in 2019.

Yew-Soon Ong (Fellow, IEEE) received the Ph.D. degree on artificial intelligence in complex design from the Computational Engineering and Design Center, University of Southampton, Southampton, U.K., in 2003. He is a President's Chair Professor of the School of Computer Science and Engineering, Nanyang Technological University (NTU), Singapore, where he is also the Director of the Data Science and Artificial Intelligence Research Center and the Principal Investigator of the Data Analytics and Complex Systems Programme with Rolls-Royce at NTU Corporate Lab. His current research interests include computational intelligence which spans across memetic computing, complex design optimization, and big data analytics.