Associated Activation-Driven Enrichment: Understanding Implicit Information from a Cognitive Perspective

Jie Bai, Student Member, IEEE, Linjing Li, Member, IEEE, Daniel Zeng, Fellow, IEEE, and Qiudan Li, Member, IEEE

Abstract—In this paper, we propose a novel text representation paradigm and a set of follow-up text representation models based on cognitive psychology theories. The intuition of our study is that the knowledge implied in a large collection of documents may improve the understanding of single documents. Based on cognitive psychology theories, we propose a general text enrichment framework, study the key factors to enable activation of implicit information, and develop new text representation methods to enrich text with the implicit information. Our study aims to mimic some aspects of human cognitive procedure in which given stimulant words serve to activate understanding implicit concepts. By incorporating human cognition into text representation, the proposed models advance existing studies by mining implicit information from given text and coordinating with most existing text representation approaches at the same time, which essentially bridges the gap between explicit and implicit information. Experiments on multiple tasks show that the implicit information activated by our proposed models matches human intuition and significantly improves the performance of the text mining tasks as well.

Index Terms—Text analysis, knowledge representation, cognitive simulation, association rules

1 INTRODUCTION

Text mining, aimed at acquiring high-quality information, patterns or semantics from text data, is a necessary pathway through which we understand the outside world. Various text mining approaches are used in a wide range of applications to analyze online information. However, in practice, getting reliable information or semantics from online text is extremely difficult. Generally speaking, understanding human language is a huge challenge that is still the focus of many studies, including in the Internet environment, which is full of noise and content loss. For example, the word “LOL” (the abbreviation for “Laugh Out Loud”) in tweets usually signals surprised or amused sentiment, but it cannot be identified in text mining works without a specific knowledge base. The following movie review provides another example: “It obviously got this big secret to hide but it seems to want to hide it completely until its final five minutes.” We can judge it as a negative comment for a suspense film from the attitude implied in the sentence, but it is hard for a machine to understand the implication based on the literal meaning of the words. The lack of linkage between explicit text and implicit concepts leads to a critical bottleneck hindering text mining.

Most text mining approaches suffer from limitations in text representation. On the one hand, most commonly used text representation approaches focus on representing text literally while ignoring the contextual background information [1], [2], [3]. On the other hand, the newly proposed text representation approaches usually learn high-order vector representations of text, where the semantic insights may be contained in the vector but the original information expressed by the literal words is lost [4], [5], [6], [7]. Thus, integrating explicit and implicit information in text is an important issue. Research based on human cognition, which excels in information integration, knowledge contextualization and intuitive understanding, is urgently needed. In this paper, we apply ideas from cognitive theories to bridge the gap between explicit and implicit information in text representation and enhance text mining performance.

Some cognitive psychological theories maintain that human reading is a process of interaction between the text seen by the reader (stimulants) and the internal knowledge kept in the reader’s brain (long-term memory) [8], which results in understanding the current text (working memory). These interactions involve how and which part of internal knowledge is activated by the text read, i.e., how
different pieces of information in long-term memory are retrieved based on external stimulants [9]. Various theories have been proposed to describe the process, among which Adaptive Control of Thought (ACT) [10] and its extensions [11], [12], [13] are widely-accepted. ACT claims that a specific concept can be jointly activated by a few stimulants based on associations between the stimulants and the concept. For instance, the word “Noah” can be jointly activated by the words “Bible,” “animals” and “flood,” although the association is much weaker when the three words are presented individually.

Based on the basic cognitive process of human reading, an activation-inspired text mining process can be derived through analogy. If the existing knowledge base serves as the long-term memory that stores the knowledge accumulated from previous experiences and the documents serve as the stimulants that need to be understood, then the text mining work is similar to the activation process from the perspective of understanding the text and acquiring information needed. As such, the cognitive psychology theories about memory activation, like ACT, can be applied to supplement existing text mining approaches with the help of well-designed knowledge bases and activation mechanisms.

Inspired by the ACT theory, we propose a new text representation paradigm called Associated Activation-Driven Enrichment (AADE), as well as a set of novel AADE models to extend existing text representation methods by learning implicit information from the given text. Intuitively, a single document may have some implicit information, but a large collection of documents usually contains a massive amount of explicit information that may aid understanding the implicit information in the document. Fig. 1 illustrates how the AADE models work for a specific document. The dashed circle outside represents the information that one can get from the document with adequate knowledge background, while the little gray circles represent the information that the words in the document have (stimulants). The purpose of AADE is to activate the implicit information (working memory) given the literal words, i.e., to find the words that can reveal the central idea of the document. By taking advantage of the word embeddings [14], which is a kind of vector representation for words learned from neural network language models, we build the knowledge base and store the distributed semantics of words from a large collection of documents. Then the activation process can be transformed into (1) getting the “central idea” of the document and (2) searching for the words that are highly correlated with the “central idea,” which is the same as the “joint activation” elaborated in the ACT theory.

In this paper, we first discuss formal text enrichment problems and propose the AADE framework as a special issue. With this framework, we adopt perplexity as the semantics gauge and infer key factors to enable activation of implicit information, which can also be treated as a demonstration of the effectiveness of ACT from the information theory point of view. Based on the inferences under different assumptions, we propose a set of AADE models accordingly. Extensive experiments were conducted to evaluate the proposed AADE models. The experimental results show that the AADE models significantly improve multiple text mining tasks on various text with linear time consumption.

The rest of this paper is organized as follows. Section 2 reviews the related works. In Section 3 we propose the problem definition and a common framework for AADE. The details of building AADE models are elaborated in Section 4. Experiments and findings are presented in Section 5, and Section 6 is the conclusion.

2 RELATED WORK

This section investigates literatures from two research fields: (1) text representation approaches and (2) cognitive activation theories and their applications in text mining.

2.1 Text Representation Methods

Text representation is the fundamental process of conventional text mining. It aims to numerically represent unstructured text documents to make them mathematically computable [15]. We divide the text representation approaches into three parts, i.e., (1) the initial Vector Space Model (VSM) presentations, (2) text enrichment by incorporating external knowledge and (3) text representation by exploring internal semantic relationships in the corpus.

Among text representation approaches, the VSM [16] paradigm is commonly applied due to its flexibility and effectiveness. In VSM, documents are mapped into a vector space and each dimension, corresponding to a particular term, reflects the weight of the term in the document. Bag-of-Words (BoW) is the most basic and the most adopted VSM representation, where the term weight is counted according to the frequency that the term appears in the document. Improvement in TF-IDF [1] considers the term specificity in the whole documents collection. Other VSM approaches include information gain [2] and pointwise mutual information [3]. These initial VSM representations
have a wide range of applications in text mining and have been extended from text to other fields such as image and video [17], [18]. However, they usually assume that the terms are independent; thus, the word correlations and the contextual information are ignored in this scheme.

To overcome the limitations of VSM, various approaches have been proposed to incorporate word correlations and contextual information into text representation models. Text enrichment, which usually leverages external knowledge for advanced VSM representation, is one method. For instance, some studies tried to introduce semantic information into text representation by incorporating existing knowledge resources, like the hierarchical structure in WordNet [19], word-word and word-document semantic relations extracted from Wikipedia [20] or conceptual features captured from domain ontology [21]. Phan et al. [22] used an external corpus for enriching short text with the help of topic models, which can be seen as an advanced application of using external resources for text enrichment. Other studies tried to extend text representation by means of other techniques. For instance, Tang et al. [23] proposed to enrich text representation by multi-language knowledge integration through machine translation; while Guo et al. [24] and Zhu et al. [25] solved the same problem by employing search engines. All these approaches need external resources or techniques to enrich the text, which is not flexible enough and is difficult to update in real-time due to rapidly changing information availability on the Internet [26].

Another approach attempts to explore information embodied in a large corpus and learn vector representation for the raw text directly. Zhu et al. [27] proposed mining word-word and text-text relationships, then integrating them into a word-text matrix to present more semantic information. Inspired by word embeddings [14], Le and Mikolov [6] proposed Paragraph Vectors to learn fixed-length vectors of sentences or documents. They trained paragraph vectors by concatenating them with the vector presentations of a few words in the paragraph and then predicting the next few words. Yu et al. [7] proposed a three-layer neural network for text representation, adding co-occurring information to enrich short text during the model training. Probabilistic topic modeling [4], [5], which learn “topics” for the word collections to obtain high-order document presentations, is another approach. Similar studies include microtext normalization. These works usually focus on the informal tokens in tweets or SMSs, replacing them with formal words [28] or mapping them in a uniform vector space [29]. These studies learned high-level abstractions for text, but the origin information represented by word sequences was lost, making it difficult to integrate into other text mining methods.

Our work is similar to these studies, since we leverage semantics embedded in the corpus to recognize implicit words and enrich document text. However, we apply cognitive psychological theories to guide the method of mining internal information from the corpus and propose a general text representation framework, which is also consistent with other text representation approaches like VSM and topic models. More generally, we further consider the nature of human cognition in text representation, which leads to advanced models in both validity and universality.

2.2 Cognitive Activation Theories and Their Applications in Text Analysis

The idea of memory activation, or recall mechanism, appeared in the early 1970s when the theory of human cognition was developed. A representative model is Human Associative Memory (HAM) theory [30], which assumes that one searches the items in long-term memory and evaluates the contextual associations of the items to determine what should be recalled.

The ACT theory [10] was proposed to include a production rule system of procedural memory to complement HAM’s declarative memory. It provides an activation function for the cognitive process. For each potential target item $i$, the activation level $A_i$ can be calculated in a concept network according to (1), where $B_i$ reflects the activation baseline of $i$, $W_j$ reflects the weight given to stimulant $j$, and $S_{ij}$ reflects the strength of associations between stimulant $j$ and potential target $i$.

$$A_i = B_i + \sum_j W_j S_{ij}$$ (1)

The ACT theory insisted that two major components are needed to build the cognition process: (1) an associative network of long-term memory that encodes the inherent knowledge about the outside world and (2) an activation mechanism that operates on the associative network to perform the cognitive task. ACT was further extended to ACT* [11], ACT-R [12] and others [13]. In ACT* [11], the inherent knowledge is conceived as a more complex structure, the cognitive units. Each cognitive unit has its own knowledge types, properties and hierarchical relationships, which give rise to more sophisticated storage and activation mechanisms. The ACT-R (R is for rational) [12] theory was introduced to solve the rational analysis problem, which seeks to maximize the achievement of information-processing goals while minimizing computational costs. An alternative to the ACT series is SAM (Search of Associative Memory) [31], which claims that memory traces vary with their familiarity rather than their activation.

Applying memory activation theories to understand natural languages has not been well-studied. Li and Larid [32] argued that the current deliberately cued retrieval mechanisms require prior knowledge about when and what to retrieve, which is often missing or incorrect. They applied the ACT-R theory and proposed to collect the necessary prior knowledge from the problem-solving context as a heuristic for relevance, then presented a functional mechanism of spontaneous, uncued retrieval. Cambria et al. [33] employed the dual-process (unconscious and conscious) models in sentiment analysis and proposed a human-like two-level affective reasoning framework, where they exploited multidimensionality reduction and graph mining techniques to mimic the integration of conscious and unconscious reasoning. Pickett and Aha [34] believed that higher-level cognition processes are built on the same fundamental building blocks as low-level perception and proposed to present relational structures as feature bags for cortically inspired algorithms in analogical learning and reasoning. Wei et al [35] claimed that existing text representation approaches usually consider the semantics in text itself and ignore the background knowledge. They built a text representation framework TRMBK...
3 ASSOCIATED ACTIVATION-DRIVEN ENRICHMENT FRAMEWORK

We shall first define some notations, which are summarized in Table 1. Let \( w \) (or \( y \)) be a word, a text fragment \( w \) is a sequence of words, i.e., \( w = \{ w_1, w_2, \ldots, w_j \} \), where \( |w| \) is the length of fragment \( w \), which is equal to the total number of words included in \( w \). The text fragment provides us a unified concept to deal with phrase, sentence, paragraph and document. We shall denote a document as \( d \); since document is an important concept in text mining, it deserves a specific symbol. The set of all documents under consideration forms the corpus \( D = \{ d_1, d_2, \ldots, d_j \} \), where \( |D| \) is the total number of documents. Further, all distinctive words included in the corpus \( D \) constitute the vocabulary \( V \); its size is denoted as \( |V| \). The Kleene closure \( V^* \) of \( V \) contains all text fragments defined in \( V \) [36]. In this sense, a corpus \( D \) is just a subset of \( V^* \), or an element of the power set, \( 2^V \), of \( V^* \). However, it is obvious that not all elements of \( 2^V \) could be treated as a meaningful corpus.

### 3.1 Text Enrichment

With the help of these notations, Text Enrichment can be viewed as a transformation \( \delta \) defined on \( V^* \). Intuitively, for an input fragment \( w \), enrichment \( \delta : V^* \rightarrow V^* \) returns a fragment \( \delta(w) \) that contains more information to help comprehend the text fragment. To make this formal, we first need a semantics information gauge function.

**Definition 1 (Semantics Gauge).** A function \( \mathcal{L} : 2^V \rightarrow \mathbb{R}_+ \) is called a semantics gauge, if \( \forall \xi_1, \xi_2 \in 2^V, \mathcal{L}(\xi_1) < \mathcal{L}(\xi_2) \) whenever \( \xi_1 \) is semantically clearer than \( \xi_2 \).

For a text fragment \( w \), we have \( \{ w \} \in 2^V \) which is a subset of \( V^* \), thus \( \mathcal{L}(\{ w \}) \) is well-defined; for the sake of simplicity, we denote it as \( \mathcal{L}(w) \). In general, gauging semantic clarity is highly application dependent. A widely employed semantics gauge is the so-called perplexity, which derives from entropy on an ergodic assumption [37] and is commonly used to evaluate probability language models.

**Definition 2 (Text Enrichment).** A transformation \( \delta : V^* \rightarrow V^* \) is called a text enrichment, if \( \forall w \in V^* \), \( \delta(w) \in V^* \) and \( \mathcal{L}(\delta(w)) < \mathcal{L}(w) \).

### 3.2 Associated Activation-Driven Enrichment

An input text fragment \( w \) can be treated as a set of stimulants; from the viewpoint of cognitive activation theories, \( w \) can be used to activate an implicit word set \( y \) with the help of the corpus \( D \) applied as the knowledge base.

**Definition 3 (Text Fragment Addition).** Text fragment addition is a binary operation \( + : V^* \times V^* \rightarrow V^* \), such that \( \forall w_1, w_2 \in V^*, w_1 + w_2 \) is the concatenation of fragment \( w_2 \) after fragment \( w_1 \).

**Definition 4 (Associated Activation-Driven Enrichment).**

Given a corpus \( D \), an enrichment \( \delta_D : V^* \rightarrow V^* \) is called an
Associated Activation-Driven Enrichment, if \( \forall \mathbf{w} \in V^* \), there exists a unique fragment \( \mathbf{y} \in V^* \), such that \( A(y) = \mathbf{w} + \mathbf{y} \).

AADE extracts a set of implicit words \( \mathbf{y} \) for the given text fragment \( \mathbf{w} \). How corpus \( D \) is used to capture semantic information will be demonstrated in the next section. Here we propose a general framework to apply AADE and prove some necessary conditions to guide its design.

As described in the ACT theory, an activation function could be built to calculate the activation level for the potential target words in \( V \), and the high-level words could be activated as the implicit information. This process is illustrated in Fig. 2. The key problems consist of (1) generating stimulants from the corpus, (2) constructing a knowledge base from the corpus, and (3) building an activation function.

### 3.3 AADE Under Perplexity

Hereafter, we shall adopt perplexity as the semantics gauge function to infer the activation functions. For any text fragment \( \mathbf{w} \), the per-word perplexity is formulated as

\[
P(\mathbf{w}) = \exp \left\{ -\sum_{j=1}^{\left|\mathbf{w}\right|} \log p(w_j) \right\},
\]

(2)

For a specific stimulant set \( \mathbf{w} \), our goal is to activate appropriate implicit word set \( \mathbf{y} = \{y_1, y_2, \ldots, y_{|\mathbf{y}|}\} \), so as to reduce the perplexity, i.e., \( P(\mathbf{w} + \mathbf{y}) < P(\mathbf{w}) \). This can be formulated with (2) as follows:

\[
P(\mathbf{w} + \mathbf{y}) = \exp \left\{ -\sum_{j=1}^{\left|\mathbf{w}\right|} \log p(w_j) + \sum_{i=1}^{\left|\mathbf{y}\right|} \log p(y_i|\mathbf{w}) \right\} < P(\mathbf{w}) \\
= \exp \left\{ -\sum_{j=1}^{\left|\mathbf{w}\right|} \log p(w_j) \right\}.
\]

(3)

It can be proved that (3) is equivalent to

\[
\sum_{i=1}^{\left|\mathbf{y}\right|} \log p(y_i|\mathbf{w}) > \sum_{j=1}^{\left|\mathbf{w}\right|} \log p(w_j).
\]

(4)

Then we analyze (4) under different assumptions.

**Assumption 1.** \( \mathbf{w} \) and \( \mathbf{y} \) are independent identically distributed (IID), respectively, i.e.,

\[
p(w_1) = p(w_2) = \cdots = p(w_{|\mathbf{w}|}) = p(y_1) = p(y_2) = \cdots = p(y_{|\mathbf{y}|})
\]

\[
p(y_1|\mathbf{w}) = p(y_2|\mathbf{w}) = \cdots = p(y_{|\mathbf{y}|}|\mathbf{w}).
\]

Then (4) can be transformed as follows:

\[
\log \left( \frac{p(y_i|\mathbf{w})}{p(\mathbf{w})p(y_i)} \right) > 0, \forall y_i \in \mathbf{y}.
\]

(5)

The left side of (5) correlates to the definition of point-wise mutual information (PMI) [3], [38], which reflects that the association between \( y_i \) and \( \mathbf{w} \) (corresponding to \( S_{y_i, \mathbf{w}} \)) should be strong enough.

**Assumption 2.** \( \mathbf{w} \) and \( \mathbf{y} \) are independent, but not identically distributed, i.e., the marginal probability of each word is unequal.

By including \( \alpha = |\mathbf{y}|/|\mathbf{w}| \) in (4), we can get

\[
\sum_{i=1}^{\left|\mathbf{y}\right|} \log p(y_i|\mathbf{w}) > \alpha \sum_{j=1}^{\left|\mathbf{w}\right|} \log p(w_j) \\
\sum_{j=1}^{\left|\mathbf{w}\right|} \log \left( \frac{p(y_j|\mathbf{w})}{p(\mathbf{w})p(y_j)} \right) > 0 \\
\sum_{i=1}^{\left|\mathbf{y}\right|} \log p(y_i) > 0
\]

(6)

Eq. (6) reflects that we should search for the implicit words \( \mathbf{y} = \{y_1, y_2, \ldots, y_{|\mathbf{y}|}\} \) such that (1) the associations of any \( y_i \) and \( \mathbf{w} \) should be strong enough and (2) the marginal probability of \( y_i \) should be high enough, given \( \mathbf{w} = [w_1, w_2, \ldots, w_{|\mathbf{w}|}] \) and \( |\mathbf{y}| = \alpha |\mathbf{w}| \). Here we would like to point out that the above inference is highly coherent with the ACT theory, if we explain the baseline activation index \( B \), in (1) as the marginal probability term \( \log \left( p(y_j) \right) \), and the word association index \( \sum_{j} W_{ij}S_{ij} \) as the PMI term \( \log \left( \frac{p(y_j|\mathbf{w})}{p(\mathbf{w})p(y_j)} \right) \).

Generally speaking, our inference for getting the activation functions can serve as an explanation of the ACT theory from the information theory point of view, which mathematically explains how ACT can help (by reducing perplexity) with natural language understanding as well as text mining tasks.

### 4 Activation Computing

In this section, we first discuss how to learn knowledge base from the corpus. Then we describe three AADE models.

#### 4.1 Association Computing and Knowledge Building

Before building the AADE models, the association term requires further definition. It is obvious that the PMI factor \( \log \left( \frac{p(y_j|\mathbf{w})}{p(\mathbf{w})p(y_j)} \right) \) is actually unreachable, since \( p(y_j, \mathbf{w}) \) is usually 0 in practical document collections where \( \mathbf{w} \) may not appear repeatedly. So it is necessary to replace \( \log \left( \frac{p(y_j|\mathbf{w})}{p(\mathbf{w})p(y_j)} \right) \) with another appropriate metric.

PMI, along with Jaccard coefficient and cosine distance, was proposed as a measure of word association [3], [38]. As a matter of fact, internal correlations exist among different measures. Look at \( \log \left( \frac{p(y_j|\mathbf{w})}{p(\mathbf{w})p(y_j)} \right) \) from the perspective of cosine distance. If we present a specific word as a vector to show if it appears in each document respectively, the cosine distance between any specific \( w_i \) and \( w_j \) is equal to the ratio that the co-occurrence frequency of \( w_i \) and \( w_j \) versus the square root of frequency \( w_i \) and \( w_j \) appears independently, which is highly correlated with PMI.
When using the cosine distance to measure the similarity between word embeddings, it is reasonable to replace the "center" of the word embedding \( w \) with the cosine distances of some finely designed word vectors. We choose word embeddings trained by the skip-gram model. Word embeddings are vector-space word presentations derived from neural language models that are good at capturing syntactic and semantic regularities in language [14].

As described in Table 1, a word \( w \) can be denoted by its word embedding \( \mathbf{v}_w \). Then the association of the specific \( w_i \) and \( w_j \) is equal to the cosine distance of \( \mathbf{v}_{w_i} \) and \( \mathbf{v}_{w_j} \):

\[
S_{i,j} = \cos(\mathbf{v}_{w_i}, \mathbf{v}_{w_j}) = \frac{\mathbf{v}_{w_i} \cdot \mathbf{v}_{w_j}}{||\mathbf{v}_{w_i}|| \cdot ||\mathbf{v}_{w_j}||}.
\]

For any word \( y \) and a specific stimulant set \( w \), their associations can be approximated by the cosine distance of \( y \) and the "center" of \( w \):

\[
S_{y,w} = \sum_{w \in w} \frac{W_{w,y,w}}{||w||} \approx \frac{\mathbf{v}_y \cdot \mathbf{v}_w}{||\mathbf{v}_y|| \cdot ||\mathbf{v}_w||} \quad \text{(9)}
\]

\[
\mathbf{v}_w = \sum_{w \in w} \mathbf{v}_w/||w||. \quad \text{(10)}
\]

The skip-gram model [39] is superior in training high quality word embeddings with low computational cost. It has been proved that the skip-gram model actually factorizes a matrix that consists of word-context PMI values [3], and this will result in more correlations between PMI and the cosine distance of word embeddings. The operation replacing PMI with cosine distance is also consistent with what we described in Section 1, i.e., the activation process is actually searching for the words that are highly correlated with the "center" of the document.

From the inferences and demonstrations above, the knowledge base needed includes two parts: (1) the marginal probability of any words in the vocabulary \( V \) and (2) the word embeddings used in association computing. The marginal probability is easy to acquire since we can count the frequency that any word appears in the corpus \( D \) and calculate the further probability. The word embeddings will be learned through skip-gram models. Then we can build some specific AADE models.

### 4.2 Three AADE Models

Three AADE models are built based on the progressive assumptions in Section 3.3. Specifically, we define three activation functions in accordance with the above inferences to evaluate the activation level for every word in \( V \), given the stimulant set \( w \). For each document \( d \) in \( D \), the number of activated words \( |y| \) is decided by \( w|w_i \), then the implicit word set \( y \) is constructed by the activated words selected according to their activation level. The activated words in \( y \) reveal the implicit information of document \( d \), and they can be added to the original BoW presentation of document \( d \) as an augmented representation. Fig. 3 illustrates the overall workflow of the proposed AADE models, which are described below.

#### 4.2.1 Global Equal Base

The Global Equal Base model is built under Assumption 1, i.e., all of the words are IID. Taking the words in \( d \) as a stimulant set, the activation level of a specific word \( y \) is decided by its association with the words in the given document \( d \). This can be formulated as

\[
A_y^d = S_{y,d} = \frac{\mathbf{v}_y \cdot \mathbf{v}_d}{||\mathbf{v}_y|| \cdot ||\mathbf{v}_d||} \quad \text{(11)}
\]

\[
\mathbf{v}_d = \sum_{w \in w} \mathbf{v}_w/||w||. \quad \text{(12)}
\]

This model is built based on the widely accepted IID assumption, which is a simplification of the actual scenario that we need to deal with.

#### 4.2.2 Global Activate

The Global Activate model is built under Assumption 2, i.e., the activation level of a specific word \( y \) is decided by its occurrence probability as well as its association with the words in the given document \( d \). This can be formulated as

\[
A_y^d = \beta p(y) + (1 - \beta) S_{y,d} \quad \text{(13)}
\]

In (13), the hyper parameter \( \beta \in (0, 1) \) denotes the trade-off between the occurrence probability and the word associations. This model is a generalization of Global Equal Base. It incorporates word occurrence probability (corresponding to the baseline activation in the ACT theory) to make the model more consistent with the actual scenario.
TABLE 2
Local Activate Algorithm

Input: Vocabulary \( V \), the document \( d \), the activation rate \( \alpha \)
Output: The implicit word set \( y \)

Local Activate:

Initialize \( y = \emptyset \), \( |y| = \alpha|w| \)
while \# of words in \( y < |y| \), do
  Extract a subsequence \( w_t \), randomly from \( d \), \( |w_t| \sim U[0, |d|] \)
  \( N_t = \alpha|w_t| \)
  \( y_t \leftarrow \) activate \( N_t \) words according to (14) and (15)
while \# of words in \( y < |y| \), do
  Pop a word form \( y \), and append it to \( y \)
return \( y \)

4.2.3 Local Activate

Different from the above models which treat all the words in the document as a stimulant set, the Local Activate model is based on an additional assumption that the implicit words are activated by some subsequences instead of the whole document. We build the model to extract the subsequences from a specific document \( d \) and activate words based on these subsequences. Specifically, denote a subsequence as \( w_t \), and the activation level of the word \( y \) given \( w_t \) can be formulated as

\[
A_y = \beta p(y) + (1 - \beta)S_p(y, w_t)
\]

\[
= \frac{N_t + \beta}{N_t} \frac{v_y v_{w_t}}{||v_y|| ||v_{w_t}||}
\]

\[
v_{w_t} = \sum_{w \in w_t} v_w
\]

The workflow of the Local Activate model for the specific document \( d \) is described in Table 2. It assumes that we understand text by word sequences (like sentences or phrases) as units, instead of understanding the document as a whole. Compared with Global Equal Base and Global Activate, this model fits reality better.

4.3 Computational Complexity Analysis

The computing time includes two parts: (1) training word embeddings and (2) activating implicit words with AADE models.

To train word embeddings, using Hierarchical Softmax for the skip-gram training process, the time cost per document is \( O(L \times \log_2(V)) \) [39]. Here \( L \) denotes the dimensionality of the word embeddings, according to Table 1. The overall computational complexity is \( O([D] \times L \times \log_2(V)) \).

The main time cost of activating implicit words with the models is in computing the cosine distance, which is determined by the dimensionality of the word embeddings \( L \), the vocabulary size \( V \) and the AADE model we choose. For Global Equal Base and Global Activate, the activation process, i.e., the cosine distance computing for the word associations, performs \( V \) times per document and the time cost is \( O(L \times V) \). Thus, the overall computational complexity is \( O([D] \times L \times V) \). For Local Activate, the expected times of the activation process is the natural logarithm \( e \)

\[ \text{Output: The implicit word set } y \]

5 EXPERIMENT

Extensive experiments were conducted on various text datasets. According to the source, style and functionality, we divided the datasets into two parts, i.e., “Reviews and Microblogs” and “News.” The Reviews and Microblogs datasets are texts generated by web users, usually short, part of which has been labeled as positive or negative. The News datasets consist of traditional news texts that have been labeled according to their subjects, which are longer and more formally written. Considering their different conditions, we conducted different experiments for these two kinds of datasets separately, so as to evaluate the quality of the proposed AADE models from various perspectives.

Comparison Methods

- **Origin**: Use the original documents without enrichment.
- **Global EB**: Enriching documents with the Global Equal Base model.
- **Global Actv**: Enriching documents with the Global Activate model.
- **Local Actv**: Enriching documents with the Local Activate Base model.
- **Similarity**: Enriching documents according to the pair-wise similarities among word embeddings [40]. This method is adopted as a baseline method since word-embedding similarity-based methods have provided the state-of-the-art performance for some text analysis tasks.
- **Random**: Enriching documents with words randomly selected from the vocabulary. This method is adopted as a control approach to evaluate the influence of the documents’ number of words on the analysis performance.

For Global EB, Global Actv, Local Actv, Similarity and Random, the enriched words are added to the BoW representation of each document (Origin). Thus, the BoW representation of the documents can be augmented while their original sequence information remains.

Model training. In the pre-training step, each dataset was used as its own training corpus for learning word embeddings. The open-source toolkit Gensim\(^3\) was adopted for training the word embeddings with skip-gram models. The hidden layer size of the skip-gram model was set as 50, i.e., \( L = 50 \).

During the experiments, the hyper parameters \( \alpha \) and \( \beta \) were learned through cross validation if there was no additional explanation.

2. The source code of the AADE models can be found at https://github.com/IBASIC-github/AADE
5.1 Experiments on Reviews and Microblogs

To verify the effectiveness of our proposed models in different scenarios and different languages, three datasets were selected for this part of the experiments.

**Movie Review:** This is a movie review dataset, “sentence polarity dataset v1.0” collected by Pang and Lee [41], which has 5,331 positive and 5,331 negative sentences collected from the Rotten Tomatoes website.

**Twitter:** This is a sentiment analysis dataset collected by Go et al. [42]. It is a massive dataset containing 1,595,509 training samples, all of which are labeled as positive or negative semantic polarity. Considering the experiment running time and the conformity with practical applications, we randomly selected 10 percent of the training set as our experiment dataset.

**Weibo:** This dataset was provided by COAE 2014. It has 40,000 samples; about 7,000 samples have been labeled as positive or negative by humans.

There are three parts of the Reviews and Microblogs experiments. The first part qualitatively evaluates if the implicit words activated by the AADE models correspond to human intuition; the second part quantitatively evaluates if AADE can help with the semantic analysis task; the third part evaluates the effectiveness of AADE for improving topic analysis. In the experiments we split the labeled samples into 80 percent training set and 20 percent test set. The details of the experiments are described below.

### 5.1.1 Words Activated

We listed the words activated by our proposed models to see to what extent these words match human intuition. The sentences were selected from both English and Chinese datasets, which may often confound regular text mining approaches.

#### Table 3: Activation Results of the AADE Models

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Sentence</th>
<th>Global EB</th>
<th>Global Actv</th>
<th>Local Actv</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Movie Review</td>
<td>offers that rare combination of entertainment and education</td>
<td>subtly; slight; riveting; fable; heartfelt; achingly; bio; unforced</td>
<td>subtly; slight; riveting; fable; heartfelt; achingly; bio; unforced</td>
<td>slight; subtly; bleak; heartwarming; thoughtful; versus; concept; honest</td>
<td>middle; films; life; piece; impact; funny; age; audience</td>
</tr>
<tr>
<td>Twitter</td>
<td>peak's subject and lead actor elevate his film above its mediocre production</td>
<td>quality; terrific; average; classics; offset; psychological; direct; direct</td>
<td>quality; terrific; average; classics; offset; psychological; direct; direct</td>
<td>quality; average; darn; directing; material; delivers; excitement; careers</td>
<td>doesn’t; t; love; people; fun; final; style; material; hollywood</td>
</tr>
<tr>
<td>Weibo</td>
<td>obviously got this big secret to hide but it seems to want to hide it completely until its final five minutes</td>
<td>hell; catch; worrying; nobody; joking; concluded; wrong; happening</td>
<td>hell; catch; worrying; nobody; joking; concluded; wrong; happening</td>
<td>catch; hell; alive; risk; him; impossible; joking; audience</td>
<td>passion; formula; predictable; age; genre; honest; effects; adaptation</td>
</tr>
<tr>
<td>I was serious! LOL</td>
<td>clever; sounded; vegetable; disrespectful; sarcastic; nifty; aha; naughty</td>
<td>clever; sounded; vegetable; disrespectful; sarcastic; nifty; aha; naughty</td>
<td>clever; sounded; vegetable; disrespectful; sarcastic; nifty; aha; naughty</td>
<td>lmao; jk; umm; yea; umm; naw; nah; lolz</td>
<td>excess; increased; recycling; clutter; inflation; corners; increase</td>
</tr>
<tr>
<td>I’m really cold. I don’t want to go to sleep yet but there’s nothing to do</td>
<td>wanna; desperately; sleepy; motivation; anywhere; motivate; fade; arsed</td>
<td>wanna; desperately; sleepy; motivation; anywhere; motivate; fade; arsed</td>
<td>wanna; anywhere; desperately; sleepy; arsed; walls; ssap; motivate</td>
<td>heat; temperature; warm; freezing; winter; sweating; wintery; windy</td>
<td></td>
</tr>
<tr>
<td>早就喝不了了，珍爱生命，远离美牛！(Already do not drink for a long time. Cherish life, stay away from Mengniu!)</td>
<td>超市, 不敢, 坚决, 反正, 自杀, 成, 说谎, 一, 晚上, 答案, 总, 赞同, 决定, 段, 不敢, 解决, 没有, 总, 没有, 表, 直接, 好, 才, 不敢, (supermarket; daren’t; resolved; anyway; suicide; bottle; collapse; ungrateful)</td>
<td>超市, 坚决, 吃, 反正, 自杀, 告, 一, 晚上, 答案, 总, 赞同, 决定, 段, 不敢, (supermarket; resolved; eat; anyway; suicide; collapse; bottle; daren’t)</td>
<td>超市; 飞机; 双; 奶粉; 健康; 理; 很; 长; 产品 (supermarket; airplane; Shuanghui; milk powder; shame; Guangming; for long; product)</td>
<td>坚决; 讨; 吃; 打; 同; 度; 理; 刚; 缺; 毫; 理 (resolved; for long time; kill; problem; BY-HEALTH; Yih; diarrhea)</td>
<td></td>
</tr>
<tr>
<td>三星, 手机中的巨无霸 (Samsung, the jumbo in mobile phones)</td>
<td>平板, 蓝色, 世, 寸, 电视, 优点, 机器, 电池, 砖头 (tablet, Galaxy; cun, computer, advantage, machine, battery, bricks)</td>
<td>平板, 蓝色, 世, 寸, 电视, 优点, 机器 (tablet, Galaxy; cun, computer, compare, cun, Galaxy; big, machine)</td>
<td>平板, 蓝色, 世, 寸, 电视, 优点, 机器 (use; apple; tablet, Galaxy; cun, computer, advantage, machine)</td>
<td>用, 苹果; 好; 袋; 联; 好; 运; 一份; 饰品; 上 (use; apple; broken; luxury; luck; piece; accessories)</td>
<td></td>
</tr>
<tr>
<td>奥迪的前灯设计确实不错 (The headlights design of Audi is really good)</td>
<td>婚车; 婚车; 内饰; 远处; 灯光, 全; 展厅, 试驾 (wedding car; pass by; upholstery; distance, light, omnidistance; gallery, test drive)</td>
<td>婚车; 婚车; 看到; 内饰; 灯光, 展厅, 试驾 (wedding car; pass by; sec; upholstery; light; gallery, test drive)</td>
<td>婚车; 婚车; 展厅, 照, 进; 一路, 摄, 华丽 (bodywork; wedding car; gallery; shocked; aggressive; all the way; alert; gorgeous)</td>
<td>别克; 感, 上; 也许; 退; 内饰; 华丽, 尤其 (Buick; load; up; maybe; rear-end; upholstery; gorgeous, especially)</td>
<td></td>
</tr>
</tbody>
</table>

5.1.1 Words Activated

We listed the words activated by our proposed models to see to what extent these words match human intuition. The sentences were selected from both English and Chinese datasets, which may often confound regular text mining approaches.

For each sentence, we listed the top 8 activated words. Since some proper names may appear in the activated words list, we removed them for a better understanding of the implicit information. The activation results are shown in Table 3, from which we can see that the activated words conform with human intuition to some extent, and some implicit information that reveals the sentence meaning usually
The average accuracy was used to evaluate the models. The experiments were conducted based on HowNet Chinese. and HowNet English. For the Chinese dataset Weibo, the experiments were conducted based on both SentiWordNet and HowNet. Unlike SentiWordNet, which has a specific score, HowNet, we combined them to judge the polarity of a word. Since there is a sentiment score and sums up the scores as the overall polarity of the document.

We chose three lexicons from two knowledge bases for the experiment. SentiWordNet\(^4\): This is a lexical resource based on WordNet 3.0, which assigns three sentiment scores to each item: positivity, negativity and objectivity. Since a word in SentiWordNet may have more than one meaning, we extracted the first sense of each word (labeled as #1) [43].

HowNet\(^5\): This is a common-sense knowledge base developed based on the WordNet framework, which contains both English and Chinese versions. Since there is a sentimentality polarity lexicon and an evaluative polarity lexicon in HowNet, we combined them to judge the polarity of a word. Unlike SentiWordNet, which has a specific score, HowNet only lists the polarity of words. Thus we simply assigned 1.0 and -1.0 to positive words and negative words, respectively.

For the English Movie Review and Twitter datasets, the experiments were conducted based on both SentiWordNet and HowNet English. For the Chinese dataset Weibo, the experiments were conducted based on HowNet Chinese. The average accuracy was used to evaluate the models. The experimental results are listed in Table 4.

Table 4 shows that our proposed AADE models improve the accuracy of the lexicon-based reviews and microblogs semantic polarity classification. The improvement is especially remarkable in the Chinese Weibo dataset. Given the fact that lexicon-based semantic polarity classification for short text is a challenging task, the AADE models effectively promote its performance with the limited corpus and the naive classification approach. For parameter learning, the value of the learned parameter \(\alpha\) usually ranges from 20 to 50, and \(\beta\) usually floats around 0.8. The exception is Random, whose performance grows worse with increasing \(\alpha\), thus the best \(\alpha\) for Random is around 0.5. Here the value of \(\alpha\) is high, which indicates that a large number of implicit words should be activated for each piece of reviews or microblogs to achieve the best performance. One reason for this is that there are usually only a few words in each review or microblog, thus compared with the formal long text, \(\alpha\) needs to be larger to activate enough number of words to aid the understanding of the text; Another reason is that the lexicon-based classification approach determines the semantic polarity using the critical polarity words, so higher \(\alpha\) is needed to make sure that enough polarity words are activated.

### 5.1.2 Semantic Polarity Analysis

In the semantic polarity analysis experiment, we chose a naive lexicon-based approach, which matches the words in the document with a specific lexicon to get the semantic polarity scores and sums up the scores as the overall polarity of the document.

We chose three lexicons from two knowledge bases for the experiment.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Lexicon</th>
<th>Origin</th>
<th>Global EB</th>
<th>Global Actv</th>
<th>Local Similarity Actv</th>
<th>Random Actv</th>
</tr>
</thead>
<tbody>
<tr>
<td>Movie Review</td>
<td>SWN</td>
<td>0.581</td>
<td>0.593</td>
<td>0.596</td>
<td>0.576</td>
<td>0.582</td>
</tr>
<tr>
<td></td>
<td>HNE</td>
<td>0.573</td>
<td>0.607</td>
<td>0.604</td>
<td>0.606</td>
<td>0.575</td>
</tr>
<tr>
<td>Twitter</td>
<td>SWN</td>
<td>0.545</td>
<td>0.590</td>
<td>0.591</td>
<td>0.584</td>
<td>0.569</td>
</tr>
<tr>
<td></td>
<td>HNE</td>
<td>0.557</td>
<td>0.613</td>
<td>0.605</td>
<td>0.596</td>
<td>0.607</td>
</tr>
<tr>
<td>Weibo</td>
<td>HNC</td>
<td>0.597</td>
<td>0.711</td>
<td>0.714</td>
<td>0.686</td>
<td>0.645</td>
</tr>
</tbody>
</table>


### 5.1.3 Perplexity Test through LDA

LDA, a widely used topic model, was used in this section to evaluate whether the AADE models can help with topic analysis for reviews and microblogs. Besides training word embeddings, Gensim was also used for training the LDA models. The topic number \(K\) was set as 50, which means that the topic models had 50 “topics” constructed by distributions over words.

Here we used test perplexity as the metric to evaluate LDA models built based on different enrichment methods. Following literature [4], the perplexity of a test set \(D_t\) under a probabilistic topic modeling framework is a generative process derived from (2), which is formulated as

\[
P(D_t) = \exp \left\{ -\frac{1}{|D_t|} \sum_{j=1}^{|D_t|} \log p(w_j | z_k)p(z_k | d_j) \right\}
\]

Based on the experience from the above experiments, the activation rate \(\alpha\) ranged from 2 to 40 to measure the test perplexity of each model. The hyper parameter \(\beta\) was decided by 10-fold cross validation for the models separately beforehand. The experimental results are shown in Fig. 4.

Fig. 4 shows that the enrichment methods are effective for most cases. The perplexity of the models decreases sharply and then stabilizes as activation rate \(\alpha\) grows. However, Random has the worst performance, and our proposed AADE
TABLE 5  
Basic Information of the News Datasets  

<table>
<thead>
<tr>
<th>Dataset</th>
<th># training samples</th>
<th># test samples</th>
<th># topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>20-Newsgroups</td>
<td>11,314</td>
<td>7,532</td>
<td>20</td>
</tr>
<tr>
<td>Reuters-21578</td>
<td>15,390</td>
<td>6,188</td>
<td>120</td>
</tr>
</tbody>
</table>

models are slightly better than Similarity. Although in the semantic polarity classification test the learned parameter $\alpha$ usually ranges from 20 to 50, the perplexity curves become stable when $\alpha$ is around 2. This means that adding more implicit words will not help the perplexity when $\alpha$ is large enough, although it may be useful for some specific tasks.

5.2 Experiments on News Text 

We chose the well-known datasets 20-Newsgroups and Reuters-21578 as the news text datasets. The basic information of these the datasets is listed in Table 5. Note that Reuters-21578 is a biased dataset, where only 57 topics contain more than 20 documents. 

In this section, we combined the AADE models with LDA to perform the implicit semantic analysis work. For all of the LDA models trained, the topic number $K$ was set as 50. 

The experiment includes three parts that describe the model quality, time consumption and performance on the downstream tasks. Each part may contains several subtasks, so as to evaluate the models comprehensively.

5.2.1 Model Quality 

We evaluated the model quality quantitatively and qualitatively. Specifically, the LDA models built based on the enrichment methods are assessed through test perplexity and their topic distribution results. 

In the perplexity tests for the News datasets, the basic settings were the same as for Reviews and Microblogs. However, the activation rate $\alpha$ ranged from 0.2 to 5.0. 

The curves in Fig. 5 show remarkable improvements in the effectiveness of topic modeling using our proposed models. The curves change sharply when $\alpha$ ranges from 0.0 to 1.0, and more or less stabilize after $\alpha$ reached 2.0. Different from the results in Fig. 4, the perplexity of Random is usually lower than that of Origin. This shows the difference between review or microblog text and news text, where the former is shorter and sparser, and the latter is longer and more normal. The curves corresponding to Random also show that for the news text, the performance improvement from using the AADE models has no relation to the number of words, which is also consistent with the analysis in Section 3.3. The curves corresponding to Similarity have the same trend as that of the AADE models, although performance is not as good, which means that Similarity is also an effective method. Actually, Similarity can be seen as a special case of Local Actv. When comparing our proposed models, we note that the performance of Global Actv is slightly better than Global EB, while Local Actv is slightly better than both.

Intuitively, the model performance is influenced by data size, since the quality of word embeddings partly relies on the training data size. To verify whether and how data size affects the AADE model performance, we tested the perplexity of different AADE models under different proportions of the datasets and then visualized the results through thermodynamic charts. The charts on the three models are similar; thus, we only show Global Actv tested under 20-Newsgroups as an example, see Fig. 6 (Appendix B, available in the online supplemental material, provides all of the thermodynamic charts). The thermodynamic chart shows that, on the one hand, larger data size indeed reduces the perplexity with fixed $\alpha$; on the other hand, fixing the data size, the perplexity is also reduced with $\alpha$ increasing. This effect is more significant when the data size is small, which means that Global Actv is still effective in improving topic modeling for small data size cases.

To examine the topic distribution details qualitatively, we compared the topic distributions of the LDA models combined with different enrichment methods. The results are shown in Tables 6 and 7. For each dataset, 6 representative topics among 50 are listed and each topic is represented by its top 8 words.

It can be seen that the topics generated from our proposed models are much more reasonable than those from Origin, Similarity or Random, according to human intuition. This may be because our proposed AADE models add more related background information for each single document, and this leads to a better performance of the topic modeling, which relies heavily on word co-occurrence in model training.
5.2.2 Time Consumption

We evaluated time consumption by comparing LDA (learned by EM algorithm) training and predicting time with those of the AADE models and Similarity. The results are shown in Fig. 7. The time of activating training data and activating test data were measured separately for more convenient comparisons.

Fig. 7 shows that the activation time varies according to the different AADE models. The activation time of Similarity increases sharply as the activation rate $\alpha$ grows. In contrast, for our proposed AADE models the activation time stays almost steady as $\alpha$ grows, while the LDA training and test time grows linearly with the number of words in the documents.

5.2.3 Downstream Tasks

To test the models in practical application scenarios, we conducted multiple experiments including document retrieval and document classification.

- Document retrieval. For document retrieval, we trained topic models with the training set and generated document vector presentations with the trained topic models for the training set and the test set...
IEEE Proof

Retrieval was set up with the test set acting as the queries and the training set acting as the database. The curves in Fig. 8 show that the retrieval precision varies with the recall percentage.

For document clustering, the training and test set were merged to perform the topic modeling and K-means clustering. The clustering number K was set according to the dataset topics, i.e., K = 20 for the 20-Newsgroups dataset, K = 57 for the Reuters-21578 dataset. The results shown in Table 8 compare different models through homogeneity (denoted as H), completeness (denoted as C) and V_measure (denoted as V) [44].

The experimental results shown in Fig. 8, Tables 8 and 9 reveal the effectiveness of our proposed AADE models in practical applications. The AADE models outperformed Origin, Similarity and Random in most cases and achieved the best results in every experiment task. Although Local Actv has lower test perplexity compared with the other two AADE models, it does not perform better in practical applications, which also suggests that the simple models sometimes work better. As for the parameters learned, the hyper parameter $\beta$ usually achieves best performance at around 0.8, which is the same as in the short text experiments. However, $\alpha$ usually ranges from 0.5 to 1.0, largely different from the experiments in Section 5.1.2. We have discussed the reasons for a large $\alpha$ in that section; for news text, it usually has more explicit information to state its idea, as such less implicit information is needed, which also results in a small $\alpha$.

Generally speaking, the experiments conducted on News datasets prove the effectiveness of our proposed AADE models for topic modeling. The experimental results for Local Actv and the learned hyper parameter value suggest that test perplexity cannot evaluate the quality of a topic model.


<table>
<thead>
<tr>
<th>Dataset</th>
<th>Metric</th>
<th>Origin</th>
<th>Global EB</th>
<th>Global Actv</th>
<th>Local Actv</th>
</tr>
</thead>
<tbody>
<tr>
<td>20-News</td>
<td>H</td>
<td>0.339</td>
<td>0.481</td>
<td>0.486</td>
<td>0.448</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.460</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.298</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0.373</td>
<td>0.505</td>
<td>0.508</td>
<td>0.472</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.352</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.362</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>0.355</td>
<td>0.493</td>
<td>0.497</td>
<td>0.459</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.399</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.327</td>
</tr>
<tr>
<td>Reuters-21578</td>
<td>H</td>
<td>0.446</td>
<td>0.466</td>
<td>0.464</td>
<td>0.454</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.441</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.416</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0.169</td>
<td>0.178</td>
<td>0.178</td>
<td>0.171</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.167</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.160</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>0.245</td>
<td>0.258</td>
<td>0.257</td>
<td>0.249</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.242</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.231</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Metric</th>
<th>Origin</th>
<th>Global EB</th>
<th>Global Actv</th>
<th>Local Actv</th>
</tr>
</thead>
<tbody>
<tr>
<td>20-News</td>
<td>P</td>
<td>0.561</td>
<td>0.627</td>
<td>0.626</td>
<td>0.611</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.580</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.503</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>0.561</td>
<td>0.617</td>
<td>0.613</td>
<td>0.601</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.588</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.509</td>
</tr>
<tr>
<td></td>
<td>F1</td>
<td>0.555</td>
<td>0.613</td>
<td>0.612</td>
<td>0.596</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.580</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.503</td>
</tr>
<tr>
<td>Reuters-21578</td>
<td>P</td>
<td>0.720</td>
<td>0.750</td>
<td>0.749</td>
<td>0.748</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.738</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.705</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>0.755</td>
<td>0.782</td>
<td>0.783</td>
<td>0.779</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.769</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.736</td>
</tr>
<tr>
<td></td>
<td>F1</td>
<td>0.703</td>
<td>0.746</td>
<td>0.749</td>
<td>0.743</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.743</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.668</td>
</tr>
</tbody>
</table>
model precisely. It has also been proved that the perplexity of topic models cannot match human intuition [45]. However, the proposed AADE models outperform Origin, Similarity and Random in test perplexity, the reasonability of topic distributions and various downstream document analysis tasks, which proves that the AADE models provide significant performance improvements for topic modeling.

6 Conclusion

There are three main contributions of this paper. First, we incorporated cognitive activation theories into text representation and proposed a new paradigm called Associated Activation-Driven Enrichment. Second, we conducted a formal demonstration to prove the value of incorporating memory activation theories to aid automatic understanding of human language. Third, we proposed a set of AADE models inspired by the activation theory. The experiments on various text datasets show that the proposed AADE models can provide significant performance improvements. The ACADE models can also be used online without training, with the word embeddings and the word counts provided in advance and updated dynamically.

Nevertheless, some problems exist in ACADE. First, since the input of the ACADE models is word sequences and the output is the additional implicit word list, the output will not follow the sequence information of the original documents, thus ACADE cannot coordinate with sequential-based analysis methods like CRF or RNN. Second, the activation results depend heavily on the quality of the word embeddings. Our further work will focus on solving these problems by embedding ACADE into deep neural networks and exploring new semantic gauges for the ACADE framework.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant No. 2016QY02D0205, the National Natural Science Foundation of China under Grant Nos. 71202169, 71602184, 71621002, and Q.-T. Ha, “A hidden topic-based framework toward building applications with short web documents,” in Proc. Int. Conf. Advances Neural Inf. Process. Syst., vol. 22, no. 1, pp. 2279–2305, 2013.

References


Jie Bai received the BE degree in communication engineering from the North China University of Technology, Beijing, China. She is currently working toward the PhD degree in the Institute of Automation, Chinese Academy of Sciences, China. Her research interests include cognitive based text analysis, machine learning, and social computing. She is a student member of the IEEE.

Linjing Li received the BE degree in electrical engineering and automation and the ME degree in control theory and control engineering from the Harbin Institute of Technology, Harbin, China, and the PhD degree in computer applied technology from the Graduate University of the Chinese Academy of Sciences, Beijing, China. He is currently an associate professor with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include game theory, mechanism design, auction theory, and machine learning. He is a member of the IEEE.

Daniel Zeng BS degree in economics and operations research from the University of Science and Technology of China, Hefei, China, and the MS and PhD degrees in industrial administration from Carnegie Mellon University. He is Gentile Family Professor in the Department of Management Information Systems at the University of Arizona. He also holds a Research Fellow position at the Institute of Automation, Chinese Academy of Sciences. His research interests include intelligence and security informatics, infectious disease informatics, social computing, recommender systems, software agents, and applied operations research and game theory. He has published more than 300 peer-reviewed articles. He currently serves as the editor in chief of ACM Transactions on MIS, and President of the IEEE Intelligent Transportation Systems Society. He is a Fellow of IEEE.

Qiudan Li received the PhD degree in computer science from the Da Lian University of Technology, Da Lian China. She is currently an associate professor in the State Key Laboratory of Management and Control for Complex Systems, the Institute of Automation, Chinese Academy of Sciences. Her research interests include web mining and machine learning. She is a member of the IEEE.

For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.