information

Article

A Hybrid Semantic Representation Method Based on Fusion Conceptual Knowledge and Weighted Word Embeddings for English Texts

Zan Qiu [1,2], Guimin Huang [1,*], Xingguo Qin [1], Yabing Wang [1], Jiahao Wang [1] and Ya Zhou [1]

[1] Guangxi Key Laboratory of Image and Graphic Intelligent Processing, School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, China; 22031102007@mails.guet.edu.cn (Z.Q.)
[2] School of Computer Science and Engineering, Guilin University of Aerospace Technology, Guilin 541004, China
* Correspondence: sendhuang@126.com

Abstract: The accuracy of traditional topic models may be compromised due to the sparsity of co-occurring vocabulary in the corpus, whereas conventional word embedding models tend to excessively prioritize contextual semantic information and inadequately capture domain-specific features in the text. This paper proposes a hybrid semantic representation method that combines a topic model that integrates conceptual knowledge with a weighted word embedding model. Specifically, we construct a topic model incorporating the Probase concept knowledge base to perform topic clustering and obtain topic semantic representation. Additionally, we design a weighted word embedding model to enhance the contextual semantic information representation of the text. The feature-based information fusion model is employed to integrate the two textual representations and generate a hybrid semantic representation. The hybrid semantic representation model proposed in this study was evaluated based on various English composition test sets. The findings demonstrate that the model presented in this paper exhibits superior accuracy and practical value compared to existing text representation methods.

Keywords: text representation; conceptual knowledge; word embeddings; information fusion

check for updates

Citation: Qiu, Z.; Huang, G.; Qin, X.; Wang, Y.; Wang, J.; Zhou, Y. A Hybrid Semantic Representation Method Based on Fusion Conceptual Knowledge and Weighted Word Embeddings for English Texts. *Information* **2024**, *15*, 708. https://doi.org/10.3390/info15110708

Academic Editor: David Martins de Matos

Received: 9 September 2024
Revised: 31 October 2024
Accepted: 1 November 2024
Published: 5 November 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The advent of the information age has precipitated an extensive reliance on text as a simplistic and convenient medium for disseminating information across diverse fields. In contrast to structured data, text is inherently unstructured, presenting significant challenges for automated analysis by machines. As widely acknowledged, a fundamental prerequisite for natural language processing tasks is to enable computers to recognize and process unstructured text in a manner analogous to numerical data. This necessitates transforming the textual information into an appropriate numerical representation that facilitates computer storage and analysis while preserving its inherent semantic content [1]. The efficient and accurate vectorization of text has thus emerged as one of the fundamental and pivotal tasks in the field of natural language processing [2]. A robust text representation method should not only comprehensively capture the semantic information embedded in natural language expressions but also effectively incorporate additional textual features, such as domain-specific knowledge, to provide a more comprehensive feature set for subsequent text processing tasks [3].

In various text analysis tasks such as topic models, machine translation, and information retrieval, enhancing the semantic vectorization of text has emerged as a pivotal aspect in advancing natural language processing for improved text understanding [4]. Early rule-based or model-based text representation methods necessitated extensive manual annotation of corpora and the development of vocabularies or dictionaries [5]. In this research

context, a semantic space was proposed to enhance the accuracy of text representation methods, addressing two fundamental challenges in natural language: lexical mismatch and inherent ambiguity [6]. Due to the inherent richness of word semantics, multiple lexical choices can be employed to convey identical meanings, while a single word may also encompass diverse semantic interpretations. By establishing associative relationships among words, semantic space facilitates more effective clustering of lexemes that express similar conceptual nuances [7]. The latent Dirichlet allocation topic model (LDA) aims to enhance the extraction of textual topics by constructing a document-topic matrix and a topic-lexicon matrix to capture the global semantics of the text. Subsequently, a hierarchical Latent Dirichlet Allocation (hLDA) was introduced, which incorporates a hierarchical structure into the potential Dirichlet topic model. It uncovers latent semantic topics in documents and represents them through a distribution with a number of topics that adapt to the number of documents [8]. With the continuous advancement of deep language learning models, the utilization of deep semantic representation has become increasingly prevalent. The Word2Vec model, a distributed word embedding approach based on neural networks proposed by Mikolov et al., enhances the precision of word clustering in the semantic space through an analysis of contextual semantic information within documents [9]. The Glove model, proposed by Pennington et al., posits that there should be a strong coherence between word vectors and co-occurrence matrices. In other words, the word vector should encapsulate the information embedded in the co-occurrence matrix, thereby capturing global semantic knowledge from documents and enhancing the representational capacity of the semantic space [10]. Knowledge bases typically encompass a wealth of entity and relationship information, which can be further enriched through semantic augmentation to enhance the model's performance. The Probase conceptual knowledge base boasts an extensive concept space, spanning conceptual knowledge across diverse domains [11]. Xu et al. proposed incorporating context-related concepts into text representations using the Probase concept knowledge base. They extracted concept and context features through concept and context embeddings, respectively. By assigning varying weights to concepts using attention layers, they calculated a weighted sum embedding matrix to represent context-related concepts, aiming to enhance model performance [12]. Although the approximate addition and subtraction features within the semantic space of word embeddings can be utilized for text representation at various levels of granularity, direct coarse-grained text representation through cumulative word embeddings inevitably leads to a loss of certain semantic features and amplifies the noise introduced by word embeddings, thereby diminishing the inherent advantages of word embeddings in semantic representation. Furthermore, while word embeddings can be directly utilized for word-level text representation, it is worth noting that these models are typically trained solely on the local contextual semantic information of words, disregarding the statistical characteristics of words within relevant subject domains or across the entire dataset. This oversight results in a deficiency of feature information related to broader fields, thereby impacting the comprehensiveness and accuracy of text representation to some extent [13]. In recent years, numerous studies have demonstrated that incorporating global information into deep semantic representation can effectively capture the semantic nuances of text. Consequently, by combining topic information and word embedding data for text representation, the comprehensiveness and accuracy of textual depiction can be enhanced. By building upon these notions, this paper proposes a hybrid semantic representation approach that integrates conceptual knowledge from a topic model with a weighted word embedding model.

The primary focus of the paper encompasses the following facets:

- It proposed a novel topic model that integrates conceptual knowledge by combining the Probase concept knowledge base with the LDA model, thereby enhancing the semantic information of topic vectors;
- It constructed a deep semantic representation model based on weighted word embeddings, leveraging the TF-IWF [14] value of each word to assess the significance of word embeddings and enhance the precision of text representation;

- It designed a hybrid semantic representation model that combines a topic model integrating conceptual knowledge with a weighted word embedding model and employs a feature information fusion strategy to enhance the accuracy and comprehensiveness of text representation;
- The hybrid semantic representation method, which combines a topic model integrating conceptual knowledge with a weighted word embedding model, was comprehensively evaluated on the English composition dataset to assess its performance.

2. Related Work

The topic model, being the earliest text analysis tool, currently stands as one of the most widely adopted language models. It effectively identifies multiple sets of keywords from a given document collection to represent each topic within the corpus. And it plays a crucial role in uncovering latent semantic structures embedded within textual data, thereby significantly contributing to refining text topics and enhancing our understanding of text semantics [15]. As an unsupervised learning method, a topic model can unveil the underlying probability distribution of topics in a text by leveraging word co-occurrence information within a given corpus. Currently, topic models have found applications across various domains including topic extraction, text classification, and analysis of social network relationships [16]. The most commonly employed approach among them is the LDA probabilistic topic model proposed by Blei et al., which represents each document in the corpus as a multinomial distribution of topics, with each topic being represented as a multinomial distribution of words [17]. The optimal parameters of the model are obtained via two inference methods, namely the Gibbs sampling algorithm and variational Bayesian inference, to identify the potential distribution of topic semantics and vocabulary in documents, thereby achieving semantic representation of text in a low-dimensional topic space [18]. Despite the remarkable success of conventional topic models across various tasks, obtaining precise textual topics solely based on statistical word information in a corpus remains challenging due to the limited co-occurrence data available. Consequently, the conventional probabilistic topic model may encounter challenges such as diminished precision in identifying topics and limited coherence in capturing topic information from English texts [19]. The sparsity of co-occurring words in the corpus poses a hindrance to the formation of text topic probability distribution, thereby facilitating the generation of topics with incoherent semantics and inaccurate vocabulary distribution by the topic model.

External knowledge can effectively complement and guide the process of topic reasoning in the text, leading to more meaningful clustering results for text topics. Furthermore, it enables the incorporation of prior domain knowledge into traditional probabilistic topic models through various approaches [20]. The DF-LDA model proposed by Andrzejewski et al. enables the incorporation of Must-link and Cannot-link knowledge of documents into the Dirichlet prior to the LDA topic model, thereby introducing semantic relationships between words to further constrain the word relationships among documents. The model incorporates prior knowledge into a Dirichlet prior distribution, replacing the previous Dirichlet prior on the topic-word distribution, and employs this informative prior for parameter inference. It leverages additional co-occurrence information of words to enhance the interpretability of the topic model. Experimental results demonstrate that incorporating word semantic similarity knowledge improves the accuracy of topic distributions in probabilistic topic models [21]. Chen et al. proposed the MWCK-LDA model, which uses vocabulary knowledge in the dictionary as prior knowledge of the LDA topic model to further constrain the training process of the topic model. The model helps the model generate more coherent topic information by fusing the semantic similarity relationship between words and using types of semantic knowledge such as synonyms, antonyms, and similar words between words [22]. With the emergence of knowledge graphs, topic models based on knowledge graphs have been proposed one after another. Zhu et al. proposed a topic model that utilizes a common sense knowledge base. This model uses common sense knowledge in ConceptNet to process each document in a given corpus into

a document graph, and then uses a neural variational inference framework to infer the model parameters, using a variety of Regularizations to constrain the topic distribution of documents [23]. Experiments show that using a common sense knowledge base in topic models can increase the semantic information and topic interpretability of topic models. Yao et al. proposed the KGE-LDA topic model, which processes knowledge in the form of triples of facts. This model uses the entities and their entity relationships in the knowledge graph as prior knowledge of the LDA topic model; TransE interprets the relationship in the knowledge graph as a conversion between subject and object and retains the inherent structure of the original knowledge graph [24]. The model incorporates external entity knowledge into the traditional probabilistic topic model, further mining deep semantic information and improving topic coherence. In addition, Shi et al. used the correlation of word contextual semantics and proposed using the non-negative matrix factorization model (NMF) to mine text topic information [25]. To solve the problem of sparse short text data and lack of co-occurrence, Ozyurt et al. proposed a sentence segment LDA (SS-LDA) to extract aspect-level emotional features in product reviews [26]. Panichella et al. aimed at the problem that most LDA applications use default parameters and can only obtain suboptimal topic distributions. By comparing two types of methods for optimizing LDA parameters, including seven state-of-the-art meta-heuristic algorithms and three alternative indicators, and by providing suggestions for combining the two types of parameters for optimization [27], Chen et al. compared LDA and NMF on multiple public short text datasets and proposed knowledge-guided non-negative matrix factorization for topic discovery of short texts to address the problem of high time complexity of NMF [28].

The utilization of word embedding stands as one of the most efficacious techniques for representing textual information in contemporary times. Currently, a majority of text representation methodologies are founded upon either word embedding or deep semantic models. Within the realm of natural language processing, word embedding entails the mapping of encoded-words onto vectors within a specific vector space through the implementation of intricate language models. The primary concept behind word vector representation is to transform the simplistic encoding of each word into a low-dimensional and dense vector space through extensive training on vast corpora, thereby addressing the issue of word relationships and mitigating the dimensionality catastrophe caused by simple coding. Bengio et al. proposed a language model composed of a feedforward neural network, which was generated using the N-Gram method. At that time, the word vectors in the model were just incidental to training the language model. With the rise of deep neural networks, a large number of deep language models suitable for natural language have been proposed. Mikolov et al. proposed the famous Word2Vec model based on the feedforward neural network language model. Word2vec includes two models, CBOW and Skip-Gram. The idea of CBOW is to predict the central word according to the context of each central word and then adjust the distribution representation of the context words by gradient descent according to the prediction results. In the process of model training, each word is taken as the central word in turn. The Skip-Gram model is the opposite of CBOW in that it predicts context words based on a given central word; the idea of the Skip-Gram model is opposite to CBOW, which predicts the context words based on the given center word. This kind of vectorized representation of words trained through context can not only obtain low-dimensional dense vectors, but these low-dimensional dense vectors also imply the semantic relationships between words. Based on global word sharing, Pennington et al. proposed the Glove model, which incorporates local contextual semantic features into the co-occurrence matrix by combining global matrix decomposition with local context windows. In recent years, with the rapid development of computer hardware, a large number of language models based on deep learning have been proposed, such as ELMo and BERT, which can dynamically adjust the semantic meaning of words according to context information to achieve more accurate representation. However, such models are deep semantic feature learning models based on the context association of words. Insufficient consideration of the global information in the text results in incomplete

semantic expression [29]. Therefore, in recent years, researchers have attempted to study hybrid semantic representation methods of text. A common approach is to combine word embeddings with shallow statistical features. Peinelt et al. studied the standard method of combining LDA with the BERT model in detail, proposed a BERT architecture based on topic information, and applied it to the pairwise semantic similarity detection task [30]. Du et al. combined LDA with Glove to propose a new text topic discovery method based on enhanced word embedding [31]. In summary, although relevant research on hybrid semantic feature learning has yielded certain outcomes, the current investigation on integrating topic information and deep semantic knowledge lacks systematicity and comprehensiveness. This paper proposes a distributed hybrid semantic representation method that combines a topic model integrating the concept knowledge base Probase with a weighted word embedding model based on Glove, thereby incorporating their respective feature information.

3. Methodology

3.1. Model Preprocessing

In the realm of natural language, preprocessing is a fundamental step in various text analysis tasks and plays a crucial role in subsequent semantic analysis. As shown in Figure 1, when utilizing natural language processing tools, English text preprocessing primarily encompasses special characters filtering, text segmentation, part-of-speech tagging, stop word removing, and lemmatization processing.

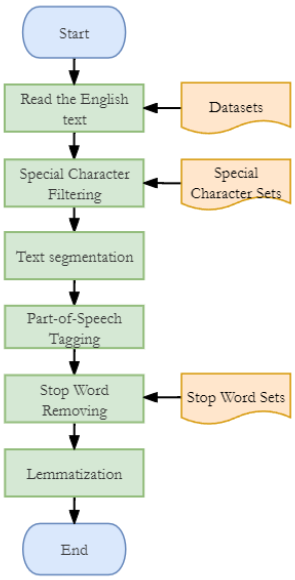


Figure 1. Model preprocessing flowchart.

3.1.1. Special Character Filtering

The text processing procedure often encounters English texts that contain special symbols, such as "◇", or Chinese punctuation marks inputted using the Chinese input method. These elements can significantly impact the segmentation and tokenization of English text. To address this issue, this paper constructs a dedicated character set by collecting commonly occurring special characters in the text. Subsequently, a regular expression matching approach is employed to batch-filter these special characters from the English text. Only after this filtering process can the subsequent step proceed with accurate segmentation.

3.1.2. Text Segmentation

The process of text segmentation involves utilizing natural language processing tools to divide the text into fundamental units. Given the necessity to consider both local

and global thematic coherence in this paper's topic analysis, an approach is adopted where the English text is segmented from a global perspective down to a local level. Initially, a highly accurate English segmentation tool is employed to segment the text into paragraphs. Subsequently, each paragraph is further divided into individual sentences. Finally, sentence segmentation takes place at the word level, resulting in the ultimate English text segmentation outcome.

3.1.3. Part-of-Speech Tagging

The present study employs a part-of-speech tagging tool to assign specific part-of-speech tags to each word in the segmented English text. The objective of part-of-speech tagging is to categorize words based on their linguistic meaning, morphology, and grammatical function. To accomplish part-of-speech tagging, we utilize Stanford University's natural language processing package known as Stanford CoreNLP [32]. This annotation tool incorporates an extensive set of rare word features comprising over 100,000 part-of-speech features that effectively cover the entire spectrum of English text for word tagging according to their parts of speech. Figure 2 illustrates an example demonstrating how this particular tool assigns tags to words for part-of-speech identification purposes. The label positioned above each word represents its corresponding part-of-speech category, for instance, NN denotes a noun while VB signifies a verb, etc.

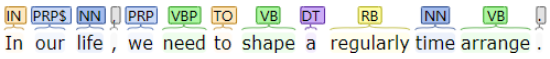


Figure 2. Example of part-of-speech tagging.

3.1.4. Stop Word Removing

Stop words refer to frequently occurring words in the text that have little semantic meaning, such as auxiliary verbs, modal particles, and prepositions. Traditional probabilistic topic models rely on co-occurrence frequency statistics of text words, making these high-frequency but low-meaning words significantly impact the obtained topic-term probability distribution. Therefore, it is essential to eliminate stop words from English texts during preprocessing to minimize noise interference caused by non-topic-related terms in compositions when using probabilistic topic models. A predefined set of stop words is employed for removing them during text processing to facilitate subsequent composition preprocessing operations.

3.1.5. Lemmatization

The purpose of lemmatization is to restore a word to its canonical form in the dictionary based on its part-of-speech tagging results after English text segmentation, thereby obtaining the corresponding root of the word. Specifically, lemmatization involves using a model to derive the base form of words in English text and mapping them back to their dictionary entries. This distinction between lemmatization and stemming is also significant: while stemming may yield incomplete or non-meaningful vocabulary, lemmatization guarantees that all results are complete words existing in the dictionary. Through this series of preprocessing operations, we obtain a vocabulary collection that best represents the English text.

3.2. Model Structure

The present study proposes a hybrid semantic representation model that integrates a concept knowledge base and weighted word embedding. It primarily encompasses an LDA topic model based on conceptual knowledge, a weighted word embedding model, and a feature fusion model. In the LDA topic model based on conceptual knowledge, the Probase concept knowledge base is employed as prior information for the LDA topic model, while incorporating an asymmetric approach prior to the probabilistic topic model approach. The prior conceptual knowledge is incorporated into the LDA topic model

to enhance the inherent semantic information of the topics, thereby facilitating a more robust textual topic analysis with profound interpretation. In order to mitigate the noise introduced by directly utilizing word embeddings based on deep language models, we propose a weighted word embedding model that combines TF-IWF and Glove word embedding models. The proposed model enables the adjustment of word embedding weights, thereby enhancing the accuracy of text representation based on word embeddings. Subsequently, to further enhance the precision and comprehensiveness of English text feature representation, a fusion strategy is devised that combines the LDA topic model based on conceptual knowledge with the weighted word embedding model for a hybrid semantic representation of text. The model framework diagram is shown in Figure 3.

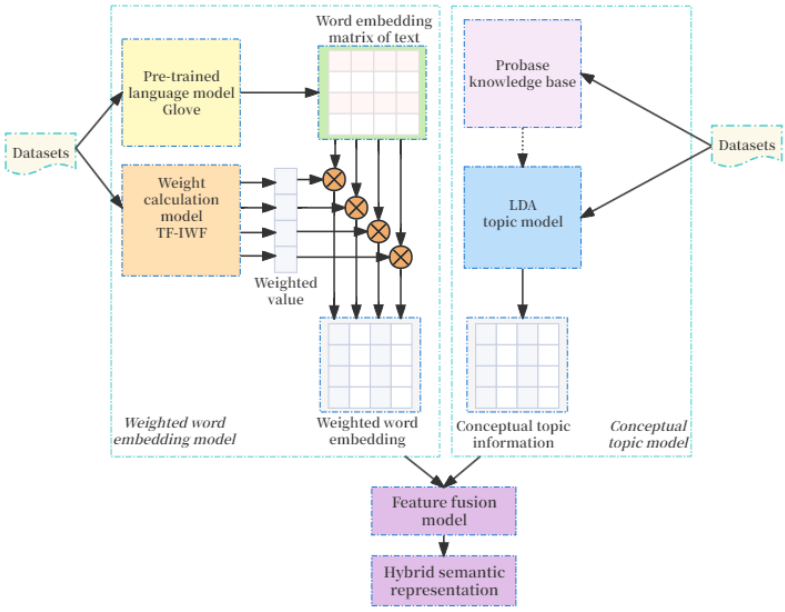


Figure 3. Model framework diagram.

3.2.1. LDA Topic Model Integrated Conceptual Knowledge

In the traditional LDA probabilistic generation model, θ_m and φ_k are the two most important sets of parameters in the model, respectively, representing the topic distribution of the document m and the lexical distribution of the kth topic. Both sets of parameters are subject to the Dirichlet distribution, that is, θ_m~Dirichlet(α), φ_k~Dirichlet(β), and α and β are the prior parameters of the Dirichlet distribution. α represents the a priori observation count of topic k in document m before observing any actual words in document m. β represents the prior observation count of vocabulary w in topic k before the vocabulary distribution of the kth topic is actually observed. When the LDA probabilistic topic model has no additional prior domain knowledge, α and β are set to equal values in each dimension. Generally, experience values are taken $\alpha = 50/K$, $\beta = 0.01$, which means that the model does not contain additional human prior knowledge and does not cause parametric shifts in the subject model.

This paper proposes a topic model that integrates conceptual knowledge, which essentially integrates the concept–instance information in the Probase concept knowledge base on the Dirichlet distribution parameters m and k of the LDA probabilistic topic model to provide concept prior knowledge for the probabilistic topic model. At this time, the various dimensions of m and k in the LDA topic model are no longer set to equal values. Instead, specific text conceptualization sets and concept clusters are generated for the English text corpus through the Probase concept knowledge base. Specific m and k are generated for the LDA topic model as prior parameters of Dirichlet distribution to increase human prior knowledge in the LDA probabilistic topic model, further expanding the deep underlying topic semantic space of the LDA topic model and improving topic coherence

and interpretability. Next, we will introduce, in detail, how to integrate the conceptual knowledge in the Probase concept knowledge base into the prior of the LDA topic model.

First of all, after the article m English text has been preprocessed, the corresponding composition root set $W^{(m)} = \left\{w_1^{(m)}, \cdots, w_N^{(m)}\right\}$ is obtained. Where N represents the total number of root sets in the m English text. Use the formula to calculate the probability that this document belongs to the concept c_i, as follows:

$$p\left(c_i \middle| W^{(m)}\right) = \frac{p\left(W^{(m)} \middle| c_i\right) p(c_i)}{p\left(W^{(m)}\right)} \propto p(c_i) \prod_{j=1}^{N} p\left(w_j^{(m)} \middle| c_i\right) \tag{1}$$

$$p(c_i) \propto \sum_{w_j^{(m)} \epsilon c_i} n\left(c_i, w_j^{(m)}\right) \tag{2}$$

In Equation (1), the symbol $\propto$ indicates that the probability is proportional to the given expression. This means that the probability is directly related to the value of the expression, up to a constant factor.

In the above formula, $p(c_i)$ is proportional to the sum of the frequencies of all words $w_j^{(m)}$ belonging to concept c_i in the mth English text. $p\left(w_j^{(m)} \middle| c_i\right)$ is the representative score of instantiating $w_j^{(m)}$ under concept c_i.

After the above formula, we can obtain the conceptual set C′ of the mth English text. After the conceptual set removes inactive concepts and deletes fuzzy concepts, we obtain the filtered concept set C″, and use the posterior probability. The probability value of Equation (1) sorts the concept set C″ in descending order. Among them, the fuzzy concept score of the Probase concept knowledge base refers to the measurement of the uncertain or unclear nature or meaning of a given concept c. It is calculated as follows:

$$Vag(c) = \frac{\sum_{e_1, e_2 \epsilon c} D(e_1, e_2)}{|c| * (|c| - 1)}, |c| > 1 \tag{3}$$

$$D(e_1, e_2) = 1 - \frac{\sum_{c_i = c_j} n(c_i, e_1) * n\left(c_j, e_2\right)}{\sqrt{\sum_{e_1 \epsilon c_i} n^2(c_i, e_1)} * \sqrt{\sum_{e_2 \epsilon c_j} n^2\left(c_j, e_2\right)}} \tag{4}$$

In the above formula, e_1, e_2 is an instance of concept c, $|c|$ is the number of instances contained in concept c, and $D(e_1, e_2)$ is the distance between instances e_1, e_2, using the conceptual distribution of instances to calculate their distance. Where $n(c_i, e_1)$ is the co-occurrence of the concept c_i and instance e_1.

Through the Probase concept knowledge base, the subtle relationships and minor changes between document vocabulary sets can be accurately controlled and an appropriate concept set can be obtained. Finally, we add the top three concepts ranked by probability in the concept set of each English text to the final concept set C of the corpus.

After obtaining the concept set C corresponding to the English composition corpus, we found that there are many identical or similar concepts in the concept set, and the number of concepts in the concept set is much larger than the number of topics set in the topic model. Therefore, after removing the same concepts from the concept set C, we use the K-Medoids algorithm to perform concept clustering on the concept set and obtain concept clusters corresponding to the number of topics. We convert the document vocabulary space into the document concept space through the Probase concept knowledge base and obtain concept clusters corresponding to the number of topics based on this. The topic model has a concept space based on the English text corpus and obtains conceptual prior knowledge of the text for the topic model.

Among them, the clustering distance between concepts is determined by the co-occurrence instances between concepts and the corresponding typicality scores. Each

concept in this article is represented by a vocabulary distribution vector. The specific formula is as follows:

$$dt(c_i, c_j) = 1 - consine\left(e_{c_i}, e_{c_j}\right) \tag{5}$$

$$e_c = \left\{(w_1, s_1), \cdots, \left(w_{|c|}, s_{|c|}\right)\right\} \tag{6}$$

The above Formula (5) calculates the semantic similarity distance between the lexical distributions of two concepts, where e_{c_i} represents the lexical distribution vector of the concept c_i. In Formula (6), w_i represents a concept instance object, and s_i represents the typicality score of the concept instance object w_i and the corresponding concept c. At this time, we use the Probase concept knowledge base to obtain concept clusters corresponding to the English composition corpus as prior knowledge for the LDA probabilistic topic model to prepare for subsequent prior parameters based on concept clusters.

Next, we incorporate conceptual knowledge into the prior parameters of the LDA topic model. The first is the prior parameter β of the subject-vocabulary, in each subject k, that is, the number of concept clusters, the prior parameter $\beta_k^{(w)}$ corresponding to each word w in the vocabulary:

$$\beta_k^{(w)} = \sum_{c_i \epsilon C_w \cap k} T(w, c_i) \tag{7}$$

In the above formula, $T(w, c_i)$ is the typical score of the concept c_i that the word w is obtained from, Probase. Among them, c_i is the concept intersection of the concept set C_w corresponding to the word w and the concept cluster k, which means that this concept set belongs to both the concept set C_w corresponding to the word w and the concept cluster k. Corresponding to the Dirichlet prior parameter β_k of each topic k, $\beta_k = \{\beta m_{w_1}, \cdots, \beta m_{w_v}\}$, where β_k is the Dirichlet symmetry prior parameter without additional prior knowledge. m_{w_i} is the normalized value of Formula (6), through which the Formula (7) is calculated. In the formula, max and min represent the maximum and minimum values in $\{\beta_k^{(w)} \Big| w = 1, \cdots, W\}$, respectively.

$$m_w = \frac{\beta_k^{(w)} - min}{max - min} \tag{8}$$

Document-topic prior parameter α use the following formula to calculate the topic probability distribution α_m of document m:

$$\alpha_m = td_{w_1}\beta_{w_1} + \cdots + td_{w_N}\beta_{w_N} \tag{9}$$

In the above formula, w_i, $i\epsilon\{1, \ldots, N\}$ represents the ith word in document m, and N represents the total number of words in document m. td_{w_i} represents the TF-IDF value of the word in the corpus and represents the importance weight of the word in the corpus. β_{w_i} represents the column value of the prior matrix corresponding to the word w_i in the topic-word distribution. Similarly, we normalize the prior distribution parameters of the document topic. α_m in Formula (9) represents the Dirichlet symmetry prior parameter without obtaining prior knowledge, m_i represents the normalized value, and max and min represent the maximum and minimum values in $\left\{\alpha_k^{(w)} \Big| k = 1, \ldots, K\right\}$, respectively.

$$m_k = \frac{\alpha_k^w - min}{max - min} \tag{10}$$

Through the above process, we can obtain the Dirichlet prior parameters α_m and β_k based on the Probase concept knowledge base, and then sample the topic using the following formula:

$$p(z_i = k|z_{\neg i}, w_i) \propto \frac{n_{k,\neg i}^{(t)} + \beta_t}{\left[\sum_{v=1}^{V} n_k^{(v)} + \beta_v\right] - 1} \cdot \frac{n_{m,\neg i}^{(k)} + \alpha_k}{\left[\sum_{z=1}^{K} n_m^{(z)} + \alpha_z\right] - 1} \tag{11}$$

where $n_k^{(v)}$ represents the number of term v appearing under topic k, $n_{k,\neg i}^{(t)}$ represents the number of term t appearing under topic k in the corpus except for the ith word, and β_t represents the Dirich of term t under topic k. α_k represents the Dirichlet prior value of topic k in document m. From the above formula, we can see that the sampling method proposed in this paper is basically consistent with the LDA thematic model. The focus of the text is to integrate the conceptual knowledge from the Probase concept knowledge base into Dirichlet's prior parameters to bring human conceptual prior knowledge to the model. The Probase-based probabilistic topic model constructed in this paper makes use of conceptual knowledge, combines similar concepts into concept clusters that conform to the English text corpus, integrates them into the priori of the LDA topic model, endowing the subject priori with certain conceptual meaning, and generates the potential topic representation of the English text with conceptual knowledge.

3.2.2. Weighted Word Embedding Model

Word embeddings derived from deep language models inherently introduce noise and fail to fully capture the significance of individual words. In this paper, we propose a novel weighted word embedding model that combines TF-IWF and Glove to enhance the embedding weight of high-value words, thereby improving the accuracy of text representation based on word embeddings.

The TF-IDF method assesses the significance of a term within a specific text in a given corpus. Its fundamental principle posits that the importance of a term is directly proportional to its frequency within the text, while it is inversely proportional to its occurrence across all texts in the corpus. In essence, when a word exhibits higher frequency within one particular text but lower frequency across other texts, it signifies a stronger correlation between the word and the text, thereby rendering it more representative of the textual characteristics. Therefore, most scholars are accustomed to using TF-IDF to measure the importance of words in text, and then combining it with other feature extraction methods to improve the accuracy of text representation. However, it is not reasonable to use IDF to measure the importance of words. For example, sometimes, there are some words with no obvious features, that is, unimportant words, which do not appear in most texts but lead to a large calculated value of IDF, and the model mistakenly thinks that the word is important, but in fact, such words are not of high importance. On the other hand, some words with high importance that are mentioned in a large number of texts will result in lower IDF values due to their occurrence in more texts, such as some words that frequently appear in a category or a field; such words are usually important in distinguishing the topic, but are given a lower weight. Different from IDF, Inverse Word Frequency (IWF) reduces the impact of similar texts in the text set on word weights and more accurately expresses the importance of words in the document to be searched. The calculation formula is as follows:

$$iwf_i = log\frac{\sum_{i=1}^{m} N_{w_i}}{N_{w_i}} \tag{12}$$

In the above formula, N_{w_i} is the word frequency of the word w_i in the text set and $\sum_{i=1}^{m} N_{w_i}$ represents the frequency of all words in the text set. According to Formula (12), if a certain word appears in multiple texts in the IWF model, but the total word frequency of the word is relatively small, the calculation result of IWF will be larger, indicating that the

word is relatively more important, which is also close to the fact, that is, the word is likely to be a significant feature of a certain type of text.

In addition, it is not reasonable to directly use TF to measure the importance of a word to the text in which it is located. For example, a certain word is highly representative of text features, but its frequency of occurrence in a certain text is low, which will result in too low TF, that is, the model underestimates the importance of the word. On the contrary, many modal particles and particles that have no actual meaning appear frequently, resulting in high TF, which means that the model seriously overestimates the importance of these words. In response to this problem, this paper deletes modal particles, particles, interjections, etc., in the text set of the preprocessing stage to improve the accuracy of the TF model. Based on the above analysis, we use TF-IWF as the word weight calculation model, as follows:

$$\mathrm{TF-IWF} = \frac{n_{i,j}}{\sum_k n_{k,j}} \times log \frac{\sum_{i=1}^{m} N_{w_i}}{N_{w_i}} \tag{13}$$

In Equation (13), the numerator $n_{i,j}$ denotes the frequency of the word w_i in text j, and the denominator $\sum\limits_k n_{k,j}$ denotes the sum of all vocabularies in the text j.

According to the above formula, calculate the TF-IWF distribution of each text in the corpus, that is, the word embedding weighted score, and normalize it according to the following formula. The normalized attention score is recorded as $h_m = (q_1, \cdots, q_l)$. The subscript m represents the text sequence number, and the subscript l represents the sequence number of the word in the text.

$$q_i = \frac{e^{q_i}}{\sum_{i=1}^{i=l} e^{q_i}} \tag{14}$$

In addition, the word embedding representation matrix of text m obtained through the deep language representation model Glove is denoted as v_m. The weighted score q_i of each word is multiplied by the corresponding word embedding vector v_i in the word embedding matrix in turn to obtain the weighted word embedding matrix of the text m, denoted as h_m. Where the word embedding vector of each text is recorded as h_i, the calculation formula is as follows:

$$h_i = q_i \times v_i \tag{15}$$

3.2.3. Feature Fusion Model

Text topic representation based on topic models has a strong ability to express global semantic information. In the LDA topic model based on concept knowledge proposed in this paper, the Probase concept knowledge base is used as a priori knowledge, and a priori concept knowledge is added to the topic model to improve the deep potential topic semantic information of the topic model. The word embedding vector based on the deep language model integrates contextual information, that is, the word embedding itself has integrated contextual semantic information and also contains semantic noise information. The weighted word embedding text representation model proposed in this article is obtained by assigning a certain weighted score to word embeddings and then accumulating them. In theory, it will inevitably incorporate more important contextual information.

The present section proposes a feature fusion strategy for the mixed semantic representation of English text. Firstly, we input the set of root words from the preprocessed text of the mth English composition into the constructed LDA topic model based on the Probase concept knowledge base and obtain the corresponding topic vector t_m. At the same time, the corresponding word embedding vector h_m is obtained through the weighted Glove word embedding model. By assigning a weight, the proportion of topic information based on conceptual knowledge and weighted word embedding information is adjusted to achieve the best English text vector representation. The calculation formula is as follows:

$$R_m = \lambda \cdot h_m + (1 - \lambda) \cdot t_m \tag{16}$$

In the above formula, λ is the weight adjustment parameter, its value range is [0,1], and R_m is the text representation vector. When λ = 1, $R_m = h_m$, the weighted word embedding is used as the representation vector of the text. When λ = 0, $R_m = t_m$, the LDA topic model based on concept knowledge is used as the representation vector of the text. When λ is other values, R_m is a hybrid semantic representation that fuses two different types of features.

4. Results and Discussion

In this section, in order to focus on text representation methods and evaluate the performance of the proposed model, we use the English composition topic analysis task to indirectly evaluate the hybrid semantic representation model proposed in this paper. We provide a detailed analysis of the conceptual knowledge-based LDA topic model and the weighted-based word embedding models and their fused performance.

4.1. Dataset

This study utilizes a diverse and comprehensive English composition corpus comprising data from the Chinese Learner English Corpus (CLEC), the International Corpus Network of Asian Learners of English (ICNALE), and the foreign English composition competition dataset available on Kaggle. Specifically, we selected 8000 English compositions across four distinct topics from CLEC, with approximately 2000 English compositions dedicated to each topic. Additionally, two composition topics were drawn from both ICNALE and Kaggle datasets, contributing another 8000 essays to our collection. The detailed distribution is presented in Table 1. The chosen English compositions span a wide array of subject areas, ensuring a rich and varied dataset for text analysis. Through the English composition dataset constructed above, a concept set and concept clustering based on the Probase concept knowledge base is generated for the topic model that integrates conceptual knowledge constructed in this article. Based on this, two important prior parameters in the topic model are obtained: the prior distribution parameters of topic vocabulary and the prior distribution parameters of document topic, providing rich conceptual prior knowledge for the LDA topic model. In terms of dataset division, the training and testing sets are proportioned at approximately 3:1 across all datasets, collectively encompassing over 40,000 words. The average English composition length within this corpus stands at around 150 words, further emphasizing the depth and breadth of the textual data employed in this research.

Table 1. Data sources for the dataset.

The Topic of Composition	Sources	On-Topic	Off-Topic
Practice Makes Perfect	CLEC	1483	517
Getting to Know the World Outside the Campus	CLEC	1315	685
How to make good use of college life	CLEC	1267	733
Chinese Traditional Festival	CLEC	1036	964
Whether it is important for college students to have a part-time job	ICCNALE	1013	987
Whether smoking should be completely banned at all the restaurants in the country	ICCNALE	1124	876
Write a response that explains how the features of the setting affect the cyclist. In your response, include examples from the essay that support your conclusion.	Kaggle	1528	472
Describe the mood created by the author in the memoir. Support your answer with relevant information from the memoir.	Kaggle	1609	391

4.2. Measure of Performance

The evaluation criteria for the model experimental results in this article use evaluation indicators widely used in various types of natural language processing analysis and machine learning in the world: accuracy, recall, and F-value. In the field of text analysis, the results are usually evaluated with the help of a confusion matrix, which consists of the true label and predicted label of the sample, as shown in Table 2.

Table 2. Confusion matrix.

Predicted Class \ True Class	Positive Sample	Negative Sample
Positive sample	TP	FP
Negative sample	FN	TN

The precision rate refers to the proportion of correct predictions among the samples predicted to be positive labels. In the field of information retrieval, it is also called the precision rate. The calculation formula is as follows:

$$P = \frac{TP}{TP + FP} \tag{17}$$

The recall rate indicates the proportion of samples that are predicted to be positive among the samples with positive true labels. The higher the recall rate requirement, the stronger the model's recall ability, that is, it is hoped to find all positive samples as much as possible. The calculation formula is as follows:

$$R = \frac{TP}{TP + FN} \tag{18}$$

In most cases, the model cannot meet the precision and recall requirements at the same time. In order to reflect the true situation of the model, the F value can be used as the evaluation index. The calculation formula is as follows:

$$F = \frac{\left(\alpha^2 + 1\right) * P * R}{\alpha^2 * (P + R)} \tag{19}$$

Among them, the larger the value of α, the greater the weight given to the recall rate; usually, the value is 1. At this time, the F value becomes the F1 value. The F1 value takes into account both precision and recall factors to play a mediating role, requiring both precision and completeness.

4.3. Analysis and Discussion

The experimental data of the English composition topic analysis method comprised 16,000 English compositions under eight topics selected from different corpora. This dataset included 10,375 on-topic and 5625 off-topic English compositions. The distributed vector representation of the word embedding model Glove was generated by training on the above corpus. In this study, the hybrid semantic representation method is evaluated for its effectiveness in the feature information fusion process, aiming to achieve optimal text representation results. By varying different values, the ratio of the conceptual knowledge fused topic model and the weighted word embedding model was adjusted, and experiments with different values of feature information fusion were conducted. By employing these strategies, we can build a robust argument demonstrating the individual and collective contributions of the different model components to the overall effectiveness of the hybrid semantic representation approach. The experimental results are shown in Table 3.

Table 3. Experimental results with different values.

λ	Precision	Recall	F1
$\lambda = 0$	90.63%	89.51%	90.07%
$\lambda = 0.1$	91.40%	89.95%	90.67%
$\lambda = 0.3$	92.76%	90.43%	91.58%
$\lambda = 0.5$	93.64%	91.07%	92.34%
$\lambda = 0.7$	94.17%	91.93%	93.03%
$\lambda = 0.9$	93.73%	91.28%	92.49%
$\lambda = 1$	93.22%	90.89%	92.04%

The experimental results show that as the proportion of the topic model that integrates conceptual knowledge and the weighted word embedding model changes, the text representation results obtained will fluctuate to a certain extent. The text representation achieves the best effect when $\lambda = 0.7$, and this configuration surpasses the effectiveness of both standalone models—the solely topic model integrated conceptual knowledge and the isolated weighted word embedding model—in terms of text representation quality. Notably, it also outperforms a naive combination approach where the two models are directly merged without optimization (i.e., at $\lambda = 0.5$), underscoring the importance of finely tuning the integration ratio for enhanced text representation.

In order to assess the efficacy of the English composition topic analysis methodology employing the hybrid semantic representation model proposed in this paper across diverse English composition corpora, a series of experiments were conducted on English compositions addressing various topics. The experimental results, as summarized in Table 4, illustrate the precision, recall rates, and F1 scores achieved by the hybrid semantic representation approach when applied to the analysis of distinct English composition topics. These metrics offer a comprehensive evaluation of the method's performance in English composition topic analysis.

Table 4. Experimental results of different topics.

The Topic of Composition	Precision	Recall	F1
Practice Makes Perfect	94.73%	87.12%	90.77%
Getting to Know the World Outside the Campus	93.65%	88.39%	90.94%
How to make good use of college life	94.88%	89.23%	91.97%
Chinese Traditional Festival	94.97%	89.68%	92.25%
Whether it is important for college students to have a part-time job	92.49%	87.61%	89.98%
Whether smoking should be completely banned at all the restaurants in the country	94.98%	86.95%	90.79%
Write a response that explains how the features of the setting affect the cyclist. In your response, include examples from the essay that support your conclusion.	90.64%	87.15%	88.86%
Describe the mood created by the author in the memoir. Support your answer with relevant information from the memoir.	91.12%	86.78%	88.90%

It can be seen from the experimental results that the English composition topic analysis method based on mixed semantic representation in this article has a better effect on the English composition test set of different topics and different corpora. The average F1 value of the English composition topic analysis method under eight topics exceeds 90%. Judging from the test results using English composition topic test sets of different lengths, the topic analysis effect of this article is basically not affected by the length of the question. The fundamental reason is that this article integrates the prior knowledge of the Probase knowledge base and enriches the semantic information of the topic. The accuracy of the English composition topic analysis of the four topics in the Chinese English Learner Corpus test set and the two topics in the Asian English Learner Corpus test set is slightly higher than the accuracy of the foreign English composition competition dataset. There may be

two reasons. One aspect may be that this article mainly uses the English compositions of Chinese students as the training set for training. Another aspect may be that foreign English composition topics are relatively open and contain scattered topic semantic information, making it difficult to perform topic clustering. Therefore, the above-mentioned test sets have relatively small differences in experimental results.

In order to verify the analytical effect of our newly devised hybrid semantic representation model, a series of meticulously designed comparative experiments were executed. This paper uses a hybrid semantic representation that combines the topic model and the Glove model (LDA+GloVe), a hybrid semantic representation that combines the topic model with conceptual knowledge and the Glove model (Improved LDA+GloVe), a hybrid semantic representation that combines the topic model and the weighted Glove model (Weighted GloVe+LDA), and a hybrid semantic representation that combines the topic model that integrates conceptual knowledge and the weighted Glove model (Our Model) to conduct comparative experiments on English composition topic analysis. In our newly devised hybrid semantic representation model, a series of meticulously designed comparative experiments were executed. The precision rate comparison experiment results are shown in Figure 4, the recall rate comparison experiment results are shown in Figure 5, and the F1 value comparison experiment results are shown in Figure 6.

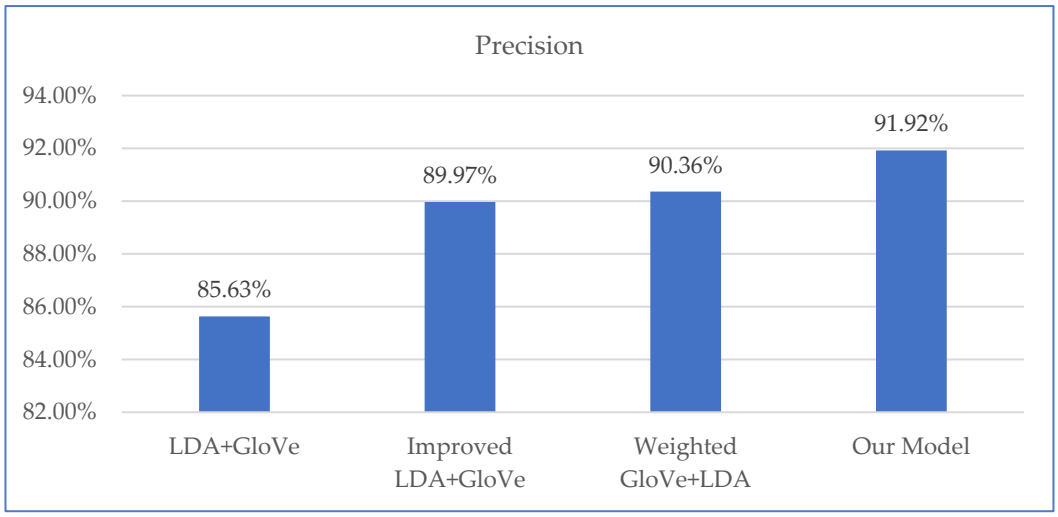


Figure 4. Experimental results of the precision comparison.

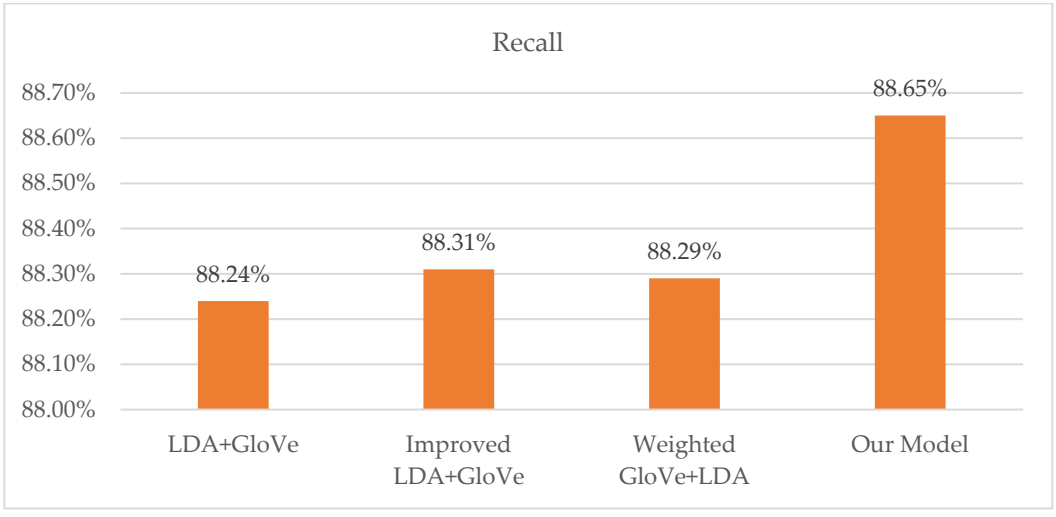


Figure 5. Experimental results of recall comparison.

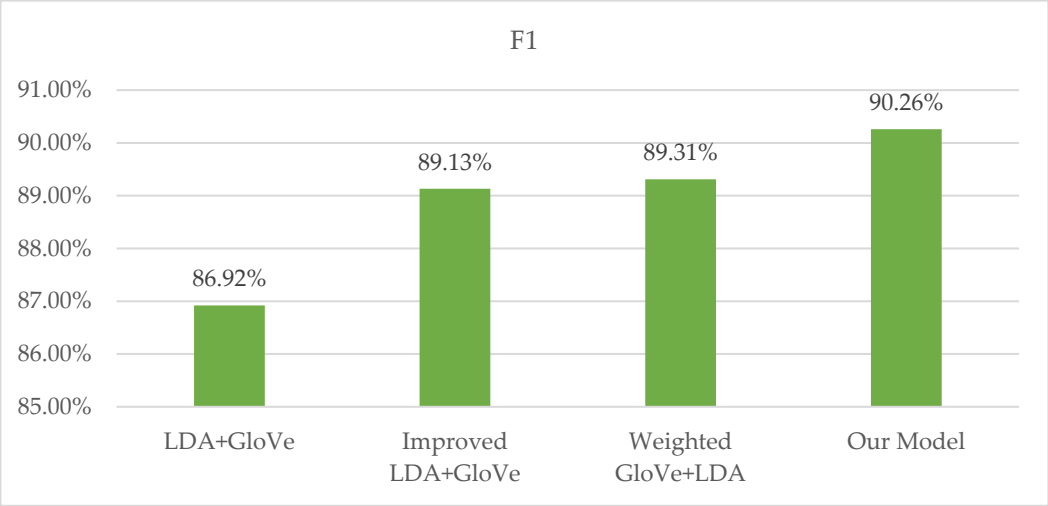


Figure 6. Experimental results of F1 comparison.

The experimental outcomes substantiate the effectiveness of our approach, a hybrid semantic representation model, which combines the topic model and the Glove word embedding model to analyze the topic of English composition. This initial phase is succeeded by the integration of an LDA topic model enriched with conceptual knowledge and weighted word embeddings, thereby facilitating a more exhaustive thematic semantic exploration. The hybrid semantic representation model constructed in this article maintains a stable recall rate; the English composition topic analysis accuracy rate of the test set is 91.92%, and the F1 value is 90.26%. These metrics collectively underscore the robustness and efficacy of our methodological innovation. Therefore, adding a topic model that integrates conceptual knowledge into the distributed semantic representation obtained by the weighted word embedding model Glove to cluster the topics of English compositions can effectively reduce the noise interference caused by non-topic words. At the same time, the LDA topic model that integrates conceptual knowledge also significantly improves the fine-grained semantic representation of topics in English compositions.

In order to verify the topic analysis method of English composition based on the mixed semantic representation proposed in this paper, a comparative experiment was conducted with the current typical topic analysis method of English composition. The current typical analysis methods include two aspects. First, use the word embedding Word2Vec model to generate a distributed vector representation and combine the word weight and the TF-IDF feature weight to form a formal representation of the sentence vector to implement English composition topic analysis. We call this method the "Word Embedding Model". Second, use the LDA topic model to extract the core topic word set and use the word embedding Word2Vec model to generate distributed word vectors to represent the expansion of the core topic words extracted from English compositions. Then, combined with the similarity between the composition and the title and the similarity between the subject words, the theme analysis of the English composition is obtained. We call this method the "LDA+Word Embedding Model". This paper proposes a hybrid semantic representation method for topic analysis of English compositions, which we call "Our Model". The above three methods were also tested using 16,000 student English compositions in the test set. The experimental results are shown in Table 5.

The experimental findings show that the "Word Embedding Model" uses the method of word embedding and model inverse document word frequency to extract key features in English compositions to a certain extent. Building upon this foundation, the "LDA+Word Embedding Model" further mitigates interference from non-topic vocabulary within the topic semantic space by selectively identifying and incorporating topical words from the texts. Our empirical validation solidifies the merit of our proposed hybrid semantic representation approach, which synergistically combines an LDA topic model augmented

with conceptual knowledge alongside a weighted word embedding model. This integration proves instrumental in substantially enriching the topic semantic representation of English compositions, thereby enhancing the precision of topic analyses. Consequently, our method not only refines the discernment of pertinent topics but also deepens the nuanced comprehension of topical content inherent in English compositions.

Table 5. Experimental results of topic analysis methods.

Method	Precision	Recall	F1
Word Embedding Model	90.13%	88.56%	89.34%
LDA+Word Embedding Model	90.74%	88.27%	89.49%
Our Model	91.92%	88.65%	90.26%

5. Conclusions

The present study proposes a hybrid semantic representation model that integrates a topic model incorporating conceptual knowledge with a weighted word embedding model. The primary innovations are as follows. Firstly, this paper employs the LDA topic model as a framework for conducting topic analysis on English compositions. It constructs an integrated LDA topic model by incorporating the Probase concept knowledge base, performs topic clustering, and subsequently obtains semantic representations of the topics. Secondly, this paper proposes the utilization of the weighted word embedding Glove model to enhance the representation of contextual semantic information in English compositions. Finally, the hybrid semantic representation of the text is obtained through the feature information fusion module designed in this paper. The experimental results demonstrate that the proposed hybrid semantic representation method, which combines a topic model integrating conceptual knowledge with a weighted word embedding model, significantly improves the performance of English composition topic analysis.

Author Contributions: Conceptualization, Z.Q. and G.H.; methodology, Z.Q. and Y.W.; formal analysis, X.Q. and J.W.; investigation, X.Q., Y.W., and J.W.; writing—original draft preparation, Z.Q.; writing—review and editing, Y.Z. and G.H.; supervision, Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (No.62066009), the Key Research and Development Project of Guilin (No.2020010308), the Guangxi Key Research and Development Project (No. Gui Ke AB22080047), the Project for Enhancing Young and Middle-aged Teacher's Research Basis Ability in Colleges of Guangxi (No.2022KY0799, No.2023KY0814), and the Fund of Guilin University of Aerospace Technology (No. XJ20KT17).

Data Availability Statement: The data that support the findings of this study are available from the corresponding authors upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Babić, K.; Martinčić-Ipšić, S.; Meštrović, A. Survey of Neural Text Representation Models. *Information* **2020**, *11*, 511. [CrossRef]
2. Minaee, S.; Kalchbrenner, N.; Cambria, E.; Nikzad, N.; Chenaghlu, M.; Gao, J. Deep Learning—Based Text Classification. *ACM Comput. Surv.* **2021**, *54*, 40. [CrossRef]
3. Otter, D.W.; Medina, J.R.; Kalita, J.K. A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 604–624. [CrossRef] [PubMed]
4. Patil, R.; Boit, S.; Gudivada, V.; Nandigam, J. A Survey of Text Representation and Embedding Techniques in NLP. *IEEE Access* **2023**, *11*, 36120–36146. [CrossRef]
5. Zhao, R.; Mao, K. Fuzzy Bag-of-Words Model for Document Representation. *IEEE Trans. Fuzzy Syst.* **2018**, *26*, 794–804. [CrossRef]
6. Jiang, Z.; Gao, S.; Chen, L. Study on Text Representation Method Based on Deep Learning and Topic Information. *Computing* **2019**, *102*, 623–642. [CrossRef]

7. Cheng, X.; Yan, X.; Lan, Y.; Guo, J. BTM: Topic Modeling over Short Texts. *IEEE Trans. Knowl. Data Eng.* **2014**, *26*, 2928–2941. [CrossRef]
8. Blei, D.M.; Griffiths, T.L.; Jordan, M.I. The Nested Chinese Restaurant Process and Bayesian Nonparametric Inference of Topic Hierarchies. *J. ACM* **2010**, *57*, 1–30. [CrossRef]
9. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed Representations of Words and Phrases and Their Compositionality. In Proceedings of the 27th Annual Conference on Neural Information Processing Systems 2013, Lake Tahoe, NV, USA, 5–8 December 2013; pp. 3111–3119.
10. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
11. Wu, W.; Li, H.; Wang, H.; Zhu, K.Q. Probase: A Probabilistic Taxonomy for Text Understanding. In Proceedings of the ACM SIGMOD International Conference on Management of Data, Scottsdale, AZ, USA, 20–24 May 2012; pp. 481–492.
12. Xu, J.; Cai, Y. Incorporating Context-Relevant Concepts into Convolutional Neural Networks for Short Text Classification. *Neurocomputing* **2020**, *386*, 42–53. [CrossRef]
13. Chauhan, U.; Shah, A. Topic Modeling Using Latent Dirichlet Allocation: A Survey. *ACM Comput. Surv.* **2022**, *54*, 1–35. [CrossRef]
14. Tian, H.; Wu, L. Microblog Emotional Analysis Based on TF-IWF Weighted Word2vec Model. In Proceedings of the IEEE 9th International Conference on Software Engineering and Service Science, Beijing, China, 23–25 November 2018; pp. 893–896.
15. Xun, G.; Li, Y.; Gao, J.; Zhang, A. Collaboratively Improving Topic Discovery and Word Embeddings by Coordinating Global and Local Contexts. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, 13–17 August 2017; pp. 535–543.
16. Joshi, A.; Fidalgo, E.; Alegre, E.; Fernández-Robles, L. DeepSumm: Exploiting Topic Models and Sequence to Sequence Networks for Extractive Text Summarization. *Expert Syst. Appl.* **2023**, *211*, 118442. [CrossRef]
17. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
18. Karras, C.; Karras, A.; Tsolis, D.; Giotopoulos, K.; Sioutas, S. Distributed Gibbs Sampling and LDA Modelling for Large Scale Big Data Management on PySpark. In Proceedings of the South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference, Ioannina, Greece, 23–25 September 2022; pp. 1–8.
19. Huang, J.; Li, P.; Peng, M.; Xie, Q.; Xu, C. Research on Subject Pattern Based on Deep Learning. *J. Comput. Sci.* **2020**, *43*, 827–855.
20. Wang, D.; Xu, Y.; Li, M.; Duan, Z.; Wang, C.; Chen, B.; Zhou, M. Knowledge-Aware Bayesian Deep Topic Model. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 14331–14344.
21. Andrzejewski, D.; Zhu, X.; Craven, M. Incorporating Domain Knowledge into Topic Modeling via Dirichlet Forest Priors. In Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, QC, Canada, 14–18 June 2009; pp. 25–32.
22. Chen, J.; Zhang, K.; Zhou, Y.; Chen, Z.; Liu, Y.; Tang, Z.; Yin, L. A Novel Topic Model for Documents by Incorporating Semantic Relations between Words. *Soft Comput.* **2019**, *24*, 11407–11423. [CrossRef]
23. Zhu, B.; Cai, Y.; Ren, H. Graph Neural Topic Model with Commonsense Knowledge. *Inf. Process. Manag.* **2023**, *60*, 103215. [CrossRef]
24. Liang, Y.; Zhang, Y.; Wei, B.; Jin, Z.; Zhang, R.; Zhang, Y.; Chen, Q. Incorporating Knowledge Graph Embeddings into Topic Modeling. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; pp. 3119–3126.
25. Shi, T.; Kang, K.; Choo, J.; Reddy, C.K. Short-Text Topic Modeling via Non-Negative Matrix Factorization Enriched with Local Word-Context Correlations. In Proceedings of the World Wide Web Conference, Lyon, France, 23–27 April 2018; pp. 1105–1114.
26. Ozyurt, B.; Ali Akcayol, M. A New Topic Modeling Based Approach for Aspect Extraction in Aspect Based Sentiment Analysis: SS-LDA. *Expert Syst. Appl.* **2021**, *168*, 114231. [CrossRef]
27. Panichella, A. A Systematic Comparison of Search-Based Approaches for LDA Hyperparameter Tuning. *Inf. Softw. Technol.* **2021**, *130*, 106411. [CrossRef]
28. Chen, Y.; Zhang, H.; Liu, R.; Ye, Z.; Lin, J. Experimental Explorations on Short Text Topic Mining between LDA and NMF Based Schemes. *Knowl. Based Syst.* **2019**, *163*, 1–13. [CrossRef]
29. Devlin, J.; Chang, M.; Lee, K. Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
30. Peinelt, N.; Nguyen, D.; Liakata, M. TBERT: Topic Models and BERT Joining Forces for Semantic Similarity Detection. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 7047–7055.
31. Du, Q.; Li, N.; Liu, W.; Sun, D.; Yang, S.; Yue, F. A Topic Recognition Method of News Text Based on Word Embedding Enhancement. *Comput. Intell. Neurosci.* **2022**, *2022*, 4582480. [CrossRef]
32. Manning, C.D.; Surdeanu, M.; Bauer, J.; Finkel, J.R.; Bethard, S.; McClosky, D. The Stanford CoreNLP Natural Language Processing Toolkit. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Baltimore, MD, USA, 23–24 June 2014; pp. 55–60.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.