

# A study of the depth calculation of the thematic distribution of English children's literature and its guidance for educational content selection

Ying Jin<sup>1</sup>, 

<sup>1</sup> Faculty of Foreign Languages and Business, Jiaozuo Normal College, Jiaozuo, Henan, 454000, China

## ABSTRACT

English children's literature has strong application value in educational content selection. This study takes classic English children's literature texts as the research object, and constructs a semantic theme mining model based on the implicit Delicacy Distribution (LDA). Through keyword weight analysis and theme probability distribution calculation, multi-dimensional theme clustering and visual characterization of literary works are realized. According to the 2378 English children's literature collected in the corpus, the LDA model was used to extract five core themes: "Adventure and Fantasy", "Friendship and Teamwork", "Growth and Self-Identity", "Family and Affection", and "Nature and Animals". A semester-long controlled experiment was conducted with third-grade students in an elementary school in Guangdong Province, designing graded English teaching content based on the results of topic distribution. Through the questionnaire survey, vocabulary test and reading ability assessment, it was found that students in the experimental group significantly outperformed the control group in terms of active interest in learning (12.42% increase in mean value) and independent learning ability (15.67% increase in test scores) ( $p < 0.05$ ). The study shows that the educational content adaptation method based on the LDA theme model can effectively optimize the selection strategy of teaching resources, and provide a theoretical basis and practical path for the precise matching of literary themes and cognitive development stages in children's English teaching.

**Keywords:** English children's literature, LDA model, topic clustering, visual representation

## 1. Introduction

English is a language subject that needs to be precipitated and accumulated in daily life, which requires a large number of literary works as the basic support in the learning process, so as to utilize literature to help students build an open eye and magnified thinking ability [1-3]. Children's literature is mostly

 Corresponding author.

E-mail address: [HNjinying111@163.com](mailto:HNjinying111@163.com) (Y. Jin).

Received 25 February 2024; Revised 10 June 2024; Accepted 05 November 2024; Published Online 16 April 2025.

DOI: [10.61091/jcmcc127b-193](https://doi.org/10.61091/jcmcc127b-193)

© 2025 The Author(s). Published by Combinatorial Press. This is an open access article under the CC BY license

(<https://creativecommons.org/licenses/by/4.0/>).

short and involves relatively basic words, which can greatly reduce the students' reading difficulty, and it includes a large number of cartoon pictures, so the combination of graphics and text can help students better understand the original text, and also stimulate the students' reading interest [4-7]. In the process of cultivating English quality education in such an open and comprehensive way, it can help students fully recognize English from the aspects of foreign culture, knowledge and life, understand English and get close to English, so that they can naturally apply English to real life [8-10].

In addition, English learners tend to lack self-confidence in the process of English learning due to the small vocabulary mastered and weak expression ability in the primary stage of learning, and introducing foreign classic children's literature into teaching is a good way to build up self-confidence [11-13]. Therefore, we can look for some foreign literary masterpieces, wonderful novels and other literary works suitable for teenagers, effectively used in primary and secondary English teaching, and become part of the teaching resources for students [14-16]. In view of this, the research on the distribution of knowledge topics in English children's literature is carried out by using the depth calculation method in the hope of giving appropriate suggestions for the selection of teaching content based on the distribution of knowledge topics in order to improve the instructional design ability of teachers [17-18].

The article explains the principle of the implicit Delicacy Distribution (LDA) model and constructs a three-level probabilistic topic model based on the bag-of-words model, the generative probability model and the topic model. The research focuses on the deep computation of topic distribution of English children's literature and its educational application, and innovatively introduces the LDA model into the field of educational content design. The LDA model automatically extracts implicit topics from massive texts, revealing semantic associations and content tendencies among the works. It captures the multidimensional semantic structure of the text through probability distribution, breaks through the framework of theme classification by artificial generalization, and provides data support for the systematic screening of educational resources. On the other hand, the linkage between the theme distribution results and teaching practice realizes the closed-loop verification from "text characterization" to "learning behavior optimization".

## **2. The significance of English children's literature in teaching and the model of theme mining**

### *2.1. The Necessity of Applying English Children's Literature in English Teaching and Learning*

Children's literature is a general term for all kinds of literary works created specifically for children as readers. Children's literature has its own distinctive features: it is full of children's interest, full of fantasies, storytelling, simple and intuitive plots, and concise and easy-to-understand language, etc. As a genre with unique aesthetic forms, children's literature has different aesthetic styles in different places. Children's literature as a unique aesthetic form of literature, due to the country, ethnicity, religion, politics, etc., its occurrence and development of the trajectory is very different, children's literature in different places present different aesthetic style, because children's literature is created for children, suitable for children's level of appreciation and aesthetic requirements, so in the elementary school English teaching, we should strengthen the use of English children's literature.

Teaching with several children's literature helps students improve their English cultural literacy. The formation of a language is inseparable from its specific geographic environment and cultural background, and in order to truly master a language, it is also inseparable from the study of the cultural background knowledge of this language. Children's literature is specially created for children and suitable for young people to read, and their creation cannot be separated from the social environment and cultural background in which the authors live, and they will surely reflect this social and cultural background. Therefore, the use of children's literature should be strengthened in the teaching process so that students can further understand the cultural background from which the English language is

produced, improve their ability to perceive and accept the English language and culture, and appreciate the real charm of the English language.

## 2.2. Principles of the LDA model

The Latent Dirichlet Distribution (LDA) is an effective semantic topic mining model in the field of natural language learning. It is used to infer the topic distribution of a document. In topic modeling, a topic represents a concept, an aspect, expressed as a series of related feature words, and the topic is the conditional probability of the feature words. It can give the topic of each document set in the form of a probability distribution, and can give the feature words of each topic in the form of a probability distribution.

LDA tries to solve a certain type of scenario problem by constructing bag-of-words model, generative probability model and topic model to discover abstract topics in a document set.

2.2.1. Bag-of-words model. Bag-of-words model i.e. a document is a collection of words, LDA topic model is based on the bag-of-words model to recognize phrases in a document. In a document only focuses on whether a phrase appears in the document or not, regardless of the order in which it appears. The same word has different probabilities in different topics. Putting all the words in a bag without considering the lexicon and word order, i.e., each word is independent, counts each word and also counts the number of times each word occurs. The bag-of-words model only considers the weights of all the words, which are related to the frequency of occurrence of the word in the text.

2.2.2. Generating probabilistic models. The LDA generative probability model assumes that each class of data obeys a Dirichlet distribution. Since there are multiple feature words representing a topic, the distribution of a topic and multiple feature words is a multinomial Dirichlet distribution, and similarly, the distribution between a document and a topic is a multinomial Dirichlet distribution. The probability density function of the Dirichlet is:

$$f(x_1, x_2, \dots, x_k; \alpha_1, \alpha_2, \dots, \alpha_k) = \frac{1}{B(\alpha)} \prod_{i=1}^k x_i^{\alpha_i-1} \quad (1)$$

where  $B(\alpha)$  is the Beta distribution.

$$B(\alpha) = \frac{\prod_{i=1}^k \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^k \alpha_i\right)} \quad (2)$$

$$\sum_{i=1}^k x_i = 1 \quad (3)$$

General Meaning the  $k$ -dimensional Dirichlet distribution expression is:

$$D(P|\alpha) = \frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K P_k^{\alpha_k-1} \quad (4)$$

The method assumes that each word is extracted from a possible hidden topic behind it. For each document in the corpus, LDA defines the following generation process: (1) for each document, extract a topic from the topic distribution; (2) extract a word from the word distribution corresponding to the topic extracted above; and (3) repeat the above process until every word in the document has been traversed.

2.2.3. Thematic models. Topic modeling is a statistical model used in fields such as machine learning and natural language processing to discover abstract topics in a series of documents. The LDA approach allows the generated documents to contain multiple topics. The process of generating topics and topic feature words for this model is as follows: first determine a number of topics  $K$ , determine a parameter  $\theta$ , select  $N$  feature words as topic representatives, and generate a dictionary from the list of feature words. Each topic obeys a Dirichlet distribution with parameter  $\theta$  for the document set, and each feature word obeys a Dirichlet distribution for that topic.

where  $\theta$  is a vector of topics, each column of the vector meaning the probability of each topic appearing in the document, which is a non-negative normalized vector;  $p(\theta)$  is the distribution of  $\theta$ , specifically the Dirichlet distribution, i.e., the distribution of distributions; the  $N$  featured words of each topic extracted from the full set  $w_n$  of words;  $z_n$  denotes the selected topic, and  $p(z|\theta)$  denotes the probability distribution of the topic  $z$  given  $\theta$  the topic, specifically the value of  $\theta$ , i.e., the value of  $p(z=i|\theta)=\theta_i$ ;  $p(w|z)$  denotes the distribution of keywords given the topic  $z$ .

Prepare  $M$  comment document as a dataset, and get  $N$  words and manually selected  $K$  topics from the document set by word splitting. The LDA method first selects a topic vector  $\theta$  to determine the probability of each topic being selected  $P$  (topic | document set), and then, when generating each word, selects a topic  $z$  from a topic distribution vector  $\theta$ , and generates a word according to the probability of word in the topic  $z$ , the probability of the word in the topic  $P$  (the eigen-word topic  $z$ ) distribution to generate a word. LDA obeys the Dirichlet distribution when selecting both topics and feature words. The Dirichlet prior concentration parameter of its topic distribution  $p(z|\theta)$  is parameter  $\alpha$ , and the topic-feature-word distribution  $p(w|z)$  is parameter  $\beta$ . The  $\alpha$  parameter indicates the document-topic density, and the higher the value of  $\alpha$  it, the more specific is the topic distribution of each document. The  $\beta$  parameter indicates the topic-word density of,  $\beta$  the higher the value, the more specific the word distribution of each topic. The model is shown in Fig. 1.

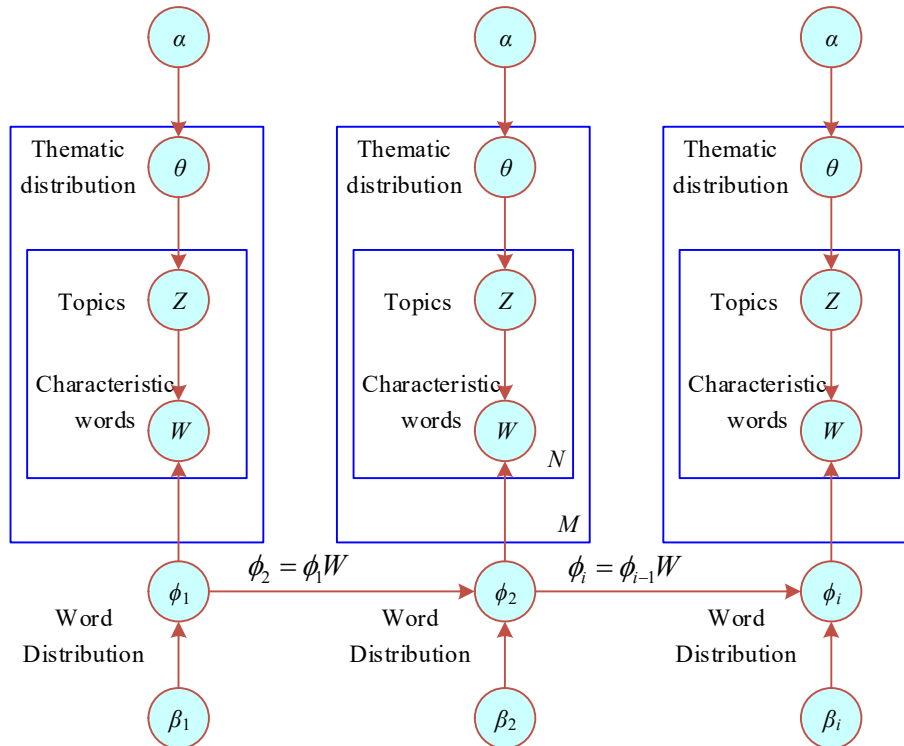


Fig. 1. LDA document model

This paper clusters the comment texts in each time interval based on the LDA topic model,

summarizes the multiple clustered topics, and obtains the full set of topics that the audience pays attention to, which lays the foundation for analyzing the core evolution trajectory.

### 2.3. Text Mining and LSA Modeling

2.3.1. Text Mining and Topic Modeling. Text mining technology is a natural language processing technology that aims to automatically extract valuable information and knowledge from a large amount of text data, which includes techniques such as text categorization, sentiment analysis, entity recognition, keyword extraction, and topic modeling. Text is a kind of unstructured data, and topic modeling, as an important tool for text mining, provides an effective way to process these unstructured data.

Since topic modeling is a text mining technique based on probabilistic graphical models, its main role is to automatically identify the implied topics from a large-scale text corpus and associate each document with these topics. The core idea of topic modeling is to represent each document as a probability distribution of topics, and at the same time to represent each topic as a probability distribution of words. In this way, for a new document, its possible topics can be determined by calculating the similarity with the probability distribution of each topic. The ultimate goal of topic modeling is to learn a topic-word matrix and a document-topic matrix, where the topic-word matrix represents the probabilistic relationship between each topic and word, and the document-topic matrix represents the probabilistic relationship between each document and topic.

2.3.2. Latent semantic analysis models. In order to overcome the shortcomings of LSA models that cannot capture the probabilistic structure of word co-occurrence, this paper proposes a probabilistic latent semantic analysis (PLSA) model based on probabilistic graph models. It is a topic model based on probabilistic graph model for topic modeling and analysis of text.

The basic idea of PLSA is to consider each word in a document as a random variable and model the topic as a latent variable, the model assumes that each document can be composed of a number of topics and each topic is a probability distribution of a number of words. During training, PLSA estimates the model parameters by maximizing the likelihood function to determine the distribution of topics for each document and the distribution of words for each topic. The whole model represents the process of document generating topics and topics generating words, thus obtaining document-word co-occurrence data.

Assuming that the set of  $n$  document is denoted as  $D = \{d_1, d_2, \dots, d_n\}$ , and the set of all words in the document is denoted as  $W = \{w_1, w_2, \dots, w_m\}$ ; setting  $K$  topics, denoted as set  $Z = \{z_1, z_2, \dots, z_k\}$ , and the multinomial probability distribution  $P(d)$  denotes the probability of generating a document  $d$ , the parameters of the iterative learning model through the EM algorithm can be computed to obtain  $P(z|d)$  and  $P(w|z)$ . where the  $P(z|d)$  multinomial conditional probability distribution denotes the probability of generating a topic  $z$  for document  $d$ , and the  $P(w|z)$  multinomial conditional probability distribution denotes the probability of generating a topic  $z$  probability of generating word  $w$ .

Although the PLSA model overcomes the disadvantage that the LSA model is unable to capture the probabilistic structure of word co-occurrence, its  $P(z|d)$  parameters increase linearly as the number of documents increases, which can easily lead to the overfitting problem of the model, and the PLSA model is computed based on a fixed word-topic distribution and topic-document distribution, which cannot assign probability values to the unknown words in the new text, which results in cannot accurately infer the topic distribution of the new document, so the probabilistic latent semantic analysis model is also not applicable to the analysis task of topic modeling for a large number of English children's literature.

#### 2.4. LDA Modeling for Topics in English Children's Literature

Due to the large total amount of data in the English Children's Literature dataset and the relatively complex compositional structure of different categories of English children's literature, the correlation between words cannot be captured using the LSA model, while the PLSA model will lead to the overfitting problem of the model with the increase of the number of works, and cannot accurately infer the thematic distribution of new children's literature texts. Based on this, the article uses the LDA model to solve this problem. LDA considers a document as a collection composed of a set of words, regardless of the order and sequential relationship between the words, and for English children's literature, a literary work can contain multiple topics. The schematic of the LDA model is shown in Fig. 2.

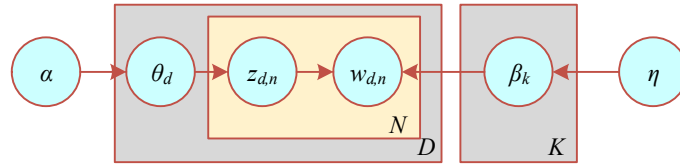


Fig. 2. LDA model

In Fig. 2,  $N$ ,  $D$ , and  $K$  indicate the number of repetitions of the variables included in the box; the gray solid circles indicate the observable values, and the hollow circles indicate the implied random variables; and the arrows indicate the dependencies of the variables.

The process of extracting themes of literary works based on the LDA model can be divided into two steps. First, as shown on the left side of Fig. 2, for each literary work, a theme is randomly selected from the theme distribution; then, as shown on the right side of Fig. 2, for each word in the work, a word is randomly selected from the word distribution of the corresponding theme. This process can be viewed as a generative process of extracting each word from the topic distribution and the word distribution.

The parameter estimation of the LDA model consists of two parts, the prior distribution of the topic distribution and the prior distribution of the word distribution. Both the prior distribution of the topic distribution and the prior distribution of the word distribution use the Dirichlet distribution, which is a probability distribution of multivariate continuous random variables.

In Figure 2,  $w$  is an observable variable and all others are hidden variables. Suppose a corpus  $D$  is a collection of texts of English children's literature; where each work  $d$  is a sequence of  $N$  words, denoted as  $W = [w_1, w_2, \dots, w_n]$ . The process of generating documents in a corpus  $D$  is as follows.

1) Select a topic probability distribution  $\theta \sim \text{Dir}(\alpha)$  from the topic Dirichlet probability distribution;

2) According to the corresponding word probability distribution in the obtained topic probability distribution, randomly sample the words in the document  $p = (w_{d,n} | \theta, \beta)$ ;

3) Repeat the above process to generate documents.

The joint probability distribution function of the LDA model is shown in equation (5).

$$p(\beta, \theta, z, w) = \left( \prod_{j=1}^N p(\beta_j | \eta) \right) \cdot \left( \prod_{d=1}^D p(\theta_d | \alpha) \prod_{n=1}^N p(z_{d,n} | \theta_d) \prod_{d=1}^D p(z_{d,n} | \beta_{i,k}, z_{d,n}) \right) \quad (5)$$

### 3. The application of LDA in English children's literature

#### 3.1. Theme extraction of works based on LDA modeling

3.1.1. Data collection. English children's works are selected based on the three principles of "dissemination", "literary" and "educational". "Communicative" refers to the ability of children's works in English to be widely disseminated, easy to understand and acceptable. This principle emphasizes that the language of the work should be clear and concise, in line with the cognitive level and language habits of children, and ensure that the information can be smoothly conveyed to the target audience. "Literary" emphasizes English children's works in the art form. The excellence of the language includes the beauty of the language, the attractiveness of the plot, the vividness of the characters, and the overall artistic conception. "Educational" is an indispensable core principle of English children's works, which aims to convey knowledge, values and social skills through stories to promote the holistic development of children. According to the three principles, 2378 English children's literature works that meet the requirements were collected in the corpus of English children's literature, and the theme extraction was carried out based on the LDA model.

3.1.2. Theme Extraction in English Children's Literature. In order to analyze the different themes and their variability in the pedagogical applications of English children's literature, this study conducted LDA topic modeling on a collection of 2378 English children's literature. The researcher of the article used the Latent Dirichlet Allocation module in python and set the number of topic models K to 20. With the number of topic models iterating, the study found that the perplexity was 673.21 at the number of topics of 5, and the model fit was optimal. Combining the topic clustering results and the keywords in each topic, the topic distribution of English children's literature can be summarized, and the topic distribution of English children's literature under LDA topic modeling is shown in Table 1.

**Table 1.** Thematic distribution of English children's literature

| Theme                       | Number of works | Keyword                                                                                 |
|-----------------------------|-----------------|-----------------------------------------------------------------------------------------|
| 1: Adventure and fantasy    | 528             | Adventure, magic, quest, dragon, treasure, forest, map, courage, danger, mythical       |
| 2: Friendship and team      | 863             | Friend, loyalty, trust, Teamwork, help, kindness, share, bond, support, together        |
| 3: Growth and self-identity | 1641            | Grow, learn, identity, Courage, challenge, mistake, Confidence, change, accept, journey |
| 4: Family and affection     | 752             | Family, parent, love, home, sibling, protect, memory, tradition, reunion, care          |
| 5: Nature and animals       | 665             | Animal, forest, tree, bird, river, wild, protect, freedom, companion, nature            |

The themes of English children's literature are divided into five categories: "Adventure and Fantasy", "Friendship and Team", "Growth and Self-Identity", "Family and Affection", and "Nature and Animals". The LDA clustering visualization results are shown in Figures 3 and 4.

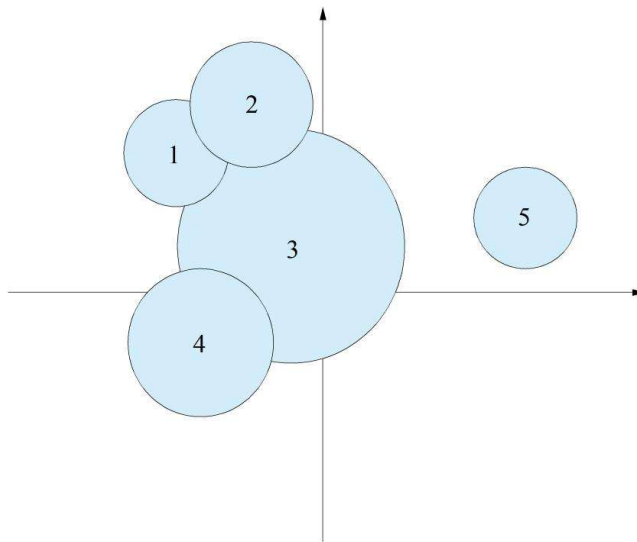


Fig. 3. LDA clustering visualization results(K=5)

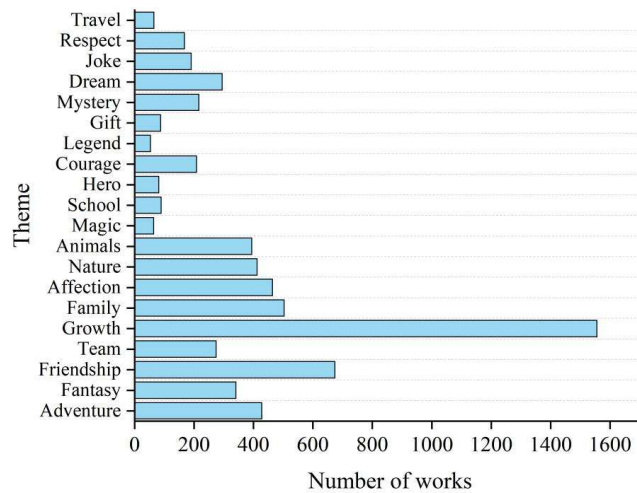


Fig. 4. LDA clustering visualization results(K=20)

3.2. Application of textual analysis of English children's literature

3.2.1. Keyword weighting analysis. Table 2 lists the 10 most highly weighted words in the corpus for the five themes that make up English-language children's literature.

Table 2. The weight of topics and keywords in the corpus

| Theme   | Keyword1  | Keyword2 | Keyword3 | Keyword4 | Keyword5  | Keyword6 | Keyword7   | Keyword8  | Keyword9  | Keyword10 |
|---------|-----------|----------|----------|----------|-----------|----------|------------|-----------|-----------|-----------|
| Theme1  | Adventure | Magic    | Quest    | Dragon   | Treasure  | Forest   | Map        | Courage   | Danger    | Mythical  |
| Weight  | 0.064     | 0.039    | 0.008    | 0.003    | 0.009     | 0.005    | 0.004      | 0.027     | 0.020     | 0.004     |
| Theme12 | Friend    | Loyalty  | Trust    | Teamwork | Help      | Kindness | Share      | Bond      | Support   | Together  |
| Weight  | 0.091     | 0.009    | 0.028    | 0.039    | 0.117     | 0.047    | 0.027      | 0.007     | 0.017     | 0.094     |
| Theme3  | Growth    | Learn    | identity | Courage  | Challenge | Mistake  | Confidence | Change    | Accept    | Journey   |
| Weight  | 0.090     | 0.009    | 0.008    | 0.009    | 0.007     | 0.007    | 0.007      | 0.007     | 0.007     | 0.004     |
| Theme4  | Family    | Parent   | Love     | Home     | Sibling   | Protect  | Memory     | Tradition | Reunion   | Care      |
| Weight  | 0.105     | 0.009    | 0.083    | 0.019    | 0.008     | 0.006    | 0.011      | 0.004     | 0.006     | 0.056     |
| Theme5  | Animal    | Forest   | Tree     | Bird     | River     | Wild     | Protect    | Freedom   | Companion | Nature    |

|        |       |       |       |       |       |       |       |       |       |       |
|--------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Weight | 0.085 | 0.019 | 0.005 | 0.003 | 0.002 | 0.003 | 0.031 | 0.014 | 0.006 | 0.066 |
|--------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|

The words with the highest weights in Theme 1 "Adventure and Fantasy" were "Advance" and "Magic" with 0.064 and 0.039, "Friend", "Kindness" and "Together" in Theme 2 "Friendship and Team" with the highest frequency with keyword weights of 0.091, 0.047 and 0.094, respectively, and "Growth" in Theme 3 "Growth and Self-Identity" The weight in turn is 0.090, which is the most important keyword in the theme. The weights of "Family" and "Love" in Topic 4 "Family and Affection" were the highest, with 0.105 and 0.083, while the weights of "Animal" and "Nature" in Topic 5 "Nature and Animals" were 0.085 and 0.066, respectively.

3.2.2. Probability of Text Attribution Theme. To further analyze the distribution of these themes, we counted the probability of each text belonging to its major theme and presented the results in Table 3. Table 3 shows the values of the probabilities of different English children's literature texts belonging to their main themes.

**Table 3.** The probability that the corpus Chinese text belongs to the main topic

| Theme                       | Literary work                                                | Chance | Literary work                                               | Chance |
|-----------------------------|--------------------------------------------------------------|--------|-------------------------------------------------------------|--------|
| 1: Adventure and fantasy    | "Alice's Adventures in Wonderland"<br>(Lewis Carol 1865)     | 0.963  | "The Lion, the Witch and the Wardrobe"<br>(C.S. Lewis 1950) | 0.932  |
| 2: Friendship and team      | "Charlotte's Web"<br>(E.B. White 1952)                       | 0.594  | "The Wonderful Wizard of Oz"<br>(L. Frank Baum 1900)        | 0.588  |
| 3: Growth and self-identity | "Holes"<br>(Louis Sachar 1988)                               | 0.841  | "A Wrinkle in Time"<br>(Madeleine L'Engle 1962)             | 0.720  |
| 4: Family and affection     | "Little House on the Prairie"<br>(Laura Ingalls Wilder 1935) | 0.723  | "The Railway Children"<br>(E. Nesbit 1906)                  | 0.882  |
| 5: Nature and animals       | "The Wind in the Willows"<br>(Kenneth Grahame 1908)          | 0.906  | "The Story of Doctor Dolittle"<br>(Hugh Lofting 1920)       | 0.857  |

The results of the analysis show that Alice in Wonderland and The Legend of Narnia: The Lion, the Witch and the Wardrobe are two classic works of English children's literature that belong to this theme, with a high probability of 0.963 and 0.932 respectively, because the fantasy elements are easy to identify, although there are other themes covered in the works, when fantasy is mostly magical, they belong to the "adventure and fantasy" theme.

And "Charlotte's Web" is about the cross-species friendship between Wilbur the pig and Charlotte the spider, showing the power of life and mutual assistance. Its natural and animal elements can not be ignored, but the overall work reflects the friendship between partners, and generally belongs to the theme of "friendship and team", with a probability of 0.594; "The Wizard of Oz" is the same, it is about the protagonist and the scarecrow, the tin man, and the lion team adventure, the work is full of adventure and fantasy factors, but the main interpretation is the true meaning of teamwork, so it still belongs to the theme of "friendship and team", with a probability of 0.588.

"The Hole" tells the story of the teenager Stanley digging out the family secrets in adversity, although there is a plot of family and family affection, but the main highlight of the work is the story of the teenager completing self-redemption and growth, which undoubtedly belongs to the theme of "growth and self-identity", with a high probability of 0.841;

"Cabin on the Prairie" tells the survival story of a pioneer family in the western grasslands, conveying tenacity and family bonds. The probability of belonging to the theme of "family and affection" is 0.723, and the probability of belonging to the theme of "family and affection" is 0.882.

There is a high probability that works in the "Nature and Animals" theme belong to this theme, because the factors about nature and animals are easy to identify. "The Wind in the Willows" tells the story of the idyllic life of moles, river rats and other animals on the banks of the Thames, and celebrates nature and friendship, with a probability of 0.906, while "Doctor Doolittle" is about a doctor who can talk to animals and leads animal friends on a rescue journey around the world, with a probability of 0.857.

#### **4. Experiments in designing educational content based on English children's literature**

Based on the previous in-depth calculation of the thematic distribution of English children's literature, this paper applies the textual analysis of the collected English children's literature to the actual English education content.

#### 4.1. *Experimental Objects and Experimental Procedures*

4.1.1. **Experimental Objects.** The third grade students of an elementary school in Guangdong Province were selected as the research object to apply this paper's depth calculation based on the thematic distribution of English children's literature in the English teaching content.

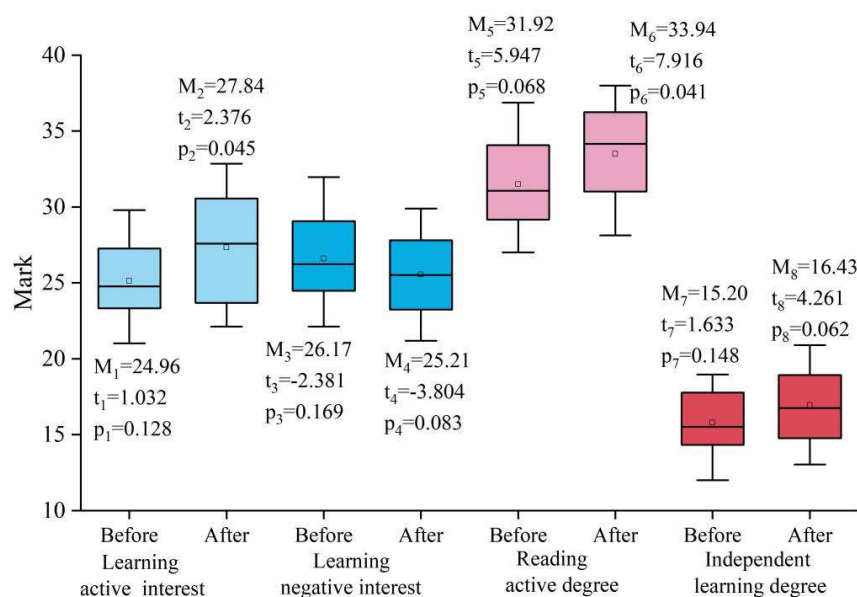
There are six classes in the third grade in this school, and two classes are selected from these six classes as the experimental group and the control group respectively. There were 82 students in the experimental group, of which 42 were boys and 40 were girls, and 81 students in the control group, of which 40 were boys and 41 were girls. The students in the experimental group were taught English education content applying English children's literature based on English literature, and the control group was taught conventional traditional English teaching content.

4.1.2. **Experimental steps.** According to the requirements of the experimental method, this experiment is divided into the following three steps: first, experimental preparation, collecting pre-test data on the English learning interest of the two groups of students; second, experimental implementation, the experimental group teaches the English teaching content designed with the application of English children's literature, and the control group teaches the regular English curriculum of the school; third, experimental evaluation, the experimental group and the control group are issued with the scales and the data of English learning interest of the students in the two groups are collected before and after the experiment. The experimental evaluation was conducted by distributing a scale to the experimental group and the control group to collect data on the students' interest in learning English before and after the experiment.

The English Learning Interest Evaluation Scale includes four dimensions: "positive interest in learning", "negative interest in learning", "positive reading degree", and "self-directed learning degree". There are five options for each question in the scale, the reverse questions are scored according to the order of options 5-1, and the positive questions are scored according to the order of options 1-5, and the higher the score, the better the effect. The sum of the scores of the four dimensions represents the student's overall level of interest in English.

#### 4.2. *Comparative analysis of data within the group before and after the experiment*

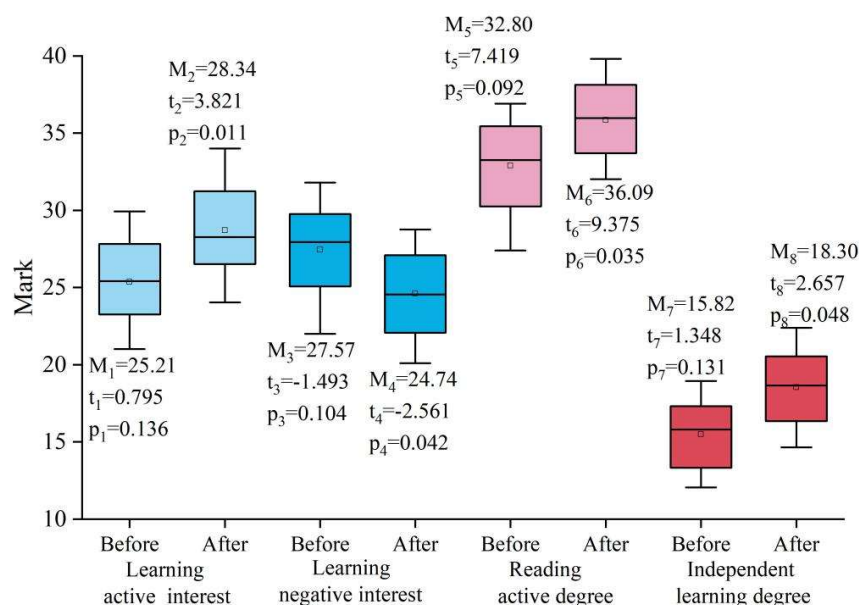
4.2.1. **Comparative analysis of within-group data before and after the experiment in the control group.** A comparison of the data on students' English learning interest before and after the experiment in the control group is shown in Figure 5.



**Fig. 5.** Comparison of students' learning interest of the control group

From the above figure, we know that the average value of positive interest in learning in the control group before the experiment is 24.76,  $p=0.128>0.05$  results of the pre-experimental test there is no significant difference, the control group's positive interest in learning in the post-experimental test the average value of 27.84,  $p=0.045<0.05$  results of the post-experimental test there is a significant difference; pre-experimental test of learning in the negative interest in the average value of 26.17, after the experimental  $p=0.169>0.05$  result there is no significant difference, the experimental post-test learning negative interest mean is 25.21, the experimental post-test  $p=0.083>0.05$  result there is no significant difference; the experimental pre-test control group reading positivity mean is 31.92, the experimental pre-test  $p=0.068>0.05$  result there is no significant difference, the experimental post-test reading positivity Mean value of 33.94, experimental post-test  $p=0.041<0.05$  result there is a significant difference; experimental pre-test control group independent learning degree mean value is 15.20, experimental pre-test  $p=0.148>0.05$  result there is no significant difference, experimental post-test control group independent learning degree mean value is 16.43, experimental post-test  $p=0.062>0.05$  result there is no significant difference.

4.2.2. Comparative analysis of intra-group data before and after the experiment in the experimental group. The data comparison of students' English learning interest before and after the experiment in the experimental group is shown in Figure 6.



**Fig. 6.** Comparison of students' learning interest of the experiential group

From the above table, it can be seen that the positive interest in learning pre-test mean of 25.21, pre-test  $p=0.136>0.05$  results did not exist significant difference, the post-test mean of 28.34, post-test  $p=0.011<0.05$  results there is a significant difference; the negative interest in learning pre-test mean of 27.57, pre-test  $p=0.104>0.05$  results did not There is a significant difference, negative interest in learning post-test mean of 24.74, post-test  $p=0.042<0.05$  result there is a significant difference; positive level of reading pre-test mean of 32.80, pre-test  $p=0.92>0.05$  result there is no significant difference, positive level of reading post-test mean of 36.09, post-test  $p=0.035<0.05$  result there is a significant difference; the degree of independent learning experimental pre-test mean was 15.82, experimental pre-test  $p=0.131>0.05$  result there is no significant difference, experimental post-test mean of the degree of independent learning was 18.30, experimental post-test  $p=0.048>0.05$  result there is a significant difference.

4.2.3. Comparative analysis of preexperimental intergroup data between the experimental group and the control group. A comparison of the data on students' English learning interest after the experiment between the experimental group and the control group is shown in Table 4.

**Table 4.** Comparison of students' interest in English learning after the experiment

| Dimension                   | Group              | N  | M±SD       | t     | p     |
|-----------------------------|--------------------|----|------------|-------|-------|
| Learning active interest    | Control group      | 81 | 27.84±5.12 | 2.376 | 0.045 |
|                             | Experimental group | 82 | 28.34±5.73 | 3.821 | 0.011 |
| Learning negative interest  | Control group      | 81 | 25.21±5.27 | 3.804 | 0.083 |
|                             | Experimental group | 82 | 24.74±4.63 | 2.561 | 0.042 |
| Reading active degree       | Control group      | 81 | 33.94±4.93 | 7.916 | 0.041 |
|                             | Experimental group | 82 | 36.09±4.07 | 9.375 | 0.035 |
| Independent learning degree | Control group      | 81 | 16.43±4.68 | 4.261 | 0.062 |
|                             | Experimental group | 82 | 18.30±4.82 | 2.657 | 0.048 |

From the above table, we know that there is a significant difference between the experimental group in the four dimensions of positive interest in learning, negative interest in learning, positive degree of reading and independent learning after the application of English children's literature. Compared with

the students in the control group, the learning effect of the students in the experimental group after the experimental teaching was significantly improved by applying the teaching content of thematic distribution of English children's literature.

## 5. Conclusion

In this study, the mining method of LDA theme modeling is used to identify and calculate the themes of English children's literature, and based on the results of the theme distribution to design the graded English teaching content, which is applied to the actual teaching practice, and the following conclusions are obtained.

1) When the number of topics is 5, the confusion degree is 673.21, and the model fitting degree reaches the best. The LDA model extracted five themes: "Adventure and Fantasy", "Friendship and Teamwork", "Growth and Self-Identity", "Family and Affection", and "Nature and Animals".

2) After the pedagogical content application of English children's literature works, the mean values of the four dimensions of students' positive interest in learning, negative interest in learning, positive degree of reading, and independent learning in the experimental group are 28.34, 24.74, 36.09, and 18.30, respectively, with p-values less than 0.05, and the results are all significantly different.

The thematic distribution of English children's literature constructed in the article helps to cultivate students' motivation and initiative in learning as well as independent thinking ability, and provides a reference for the reform of English curriculum content.

## References

- [1] Bland, J. (2019). Teaching English to Young Learners: More Teacher Education and More Children's Literature!. *Online Submission*, 7(2), 79-103.
- [2] Kaşlıoğlu, Ö., & Ersin, P. (2018). Pre-service teachers' beliefs about literature integration in English language teaching classrooms. *Journal of Language and Linguistic Studies*, 14(3), 213-232.
- [3] Rainey, E. C. (2017). Disciplinary literacy in English language arts: Exploring the social and problem-based nature of literary reading and reasoning. *Reading Research Quarterly*, 52(1), 53-71.
- [4] Koss, M. D., & Daniel, M. C. (2018). Valuing the Lives and Experiences of English Learners: Widening the Landscape of Children's Literature. *TESOL Journal*, 9(3), 431-454.
- [5] Shuqair, K., & Dashti, A. (2019). Teachers' Perceptions of the Use and Effectiveness of Children's Literature in the EFL Classrooms of the Primary Schools of Kuwait. *English Language Teaching*, 12(7), 87-97.
- [6] Putri, M. A., Harto, S., & Moecharam, N. Y. (2017). Promoting character education in EFL classroom: Using children's literature as a teaching material. *Journal of English and Education*, 5(2), 163-169.
- [7] Akins, M., Tichenor, M., Heins, E., & Piechura, K. (2018). Teachers' knowledge of children's literature: What genres do teachers read?. *Reading Improvement*, 55(2), 63-68.
- [8] Yang, C. (2023). Research on the Effect of Teachers' Use of Extracurricular Materials on Improving Students' Comprehensive Quality. *Journal of Education, Humanities and Social Sciences*, 8, 223-228.
- [9] Chen, Q. (2024). Integration of In-Class and Extracurricular English Reading in Elementary Schools under the "Double Reduction" Policy. *Journal of Education and Educational Research*, 8(3), 193-195.
- [10] Jie-Feng, W., Xin-Fei, W., & He, X. (2021, March). The Evaluation of College English Extracurricular Reading Independent Learning Model Based on Big Data. In *2021 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)* (pp. 451-455). IEEE.
- [11] Erdem, M. (2016). Literature in English language teaching. *European Journal of Language and Literature Studies*, 2(1), 157-162.

- [12] Tehan, P., Yuksel, D., & Inan, B. (2015). The place of literature in an English language teaching program: What do students think about it. *The reading matrix: An International online journal*, 15(2), 45-52.
- [13] Cheetham, D. (2015). Extensive reading of children's literature in first, second, and foreign language vocabulary acquisition. *CLELE journal*, 3(2), 1-23.
- [14] Renandya, W. A., Krashen, S., & Jacobs, G. M. (2018). The potential of series books: How narrow reading leads to advanced L2 proficiency. *LEARN Journal: Language Education and Acquisition Research Network*, 11(2), 148-154.
- [15] Durrant, P., & Brenchley, M. (2019). Development of vocabulary sophistication across genres in English children's writing. *Reading and writing*, 32(8), 1927-1953.
- [16] Jinxiu, J., & Zhengping, Z. (2016). Principles and Implementation of Reading Activities in Primary School English Class. *English Language Teaching*, 9(12), 74-78.
- [17] Jautze, K., van Cranenburgh, A., & Koolen, C. (2016). Topic Modeling Literary Quality. In DH (pp. 233-237).
- [18] Li, D., Wu, K., & Lei, V. L. (2024). Applying Topic Modeling to Literary Analysis: A Review. *Digital Studies in Language and Literature*, 1(1-2), 113-141.