# Character Representation and Application Analysis of English Language and Literature Based on Neural Network

Yao Song
School of Foreign Languages,
Henan University of Animal Husbandry and Economy,
Zhengzhou, 450000, China

***Abstract*—The development of computer technology has promoted the continuous progress of Natural language processing technology and the great development of ideology and culture, and also prompted literary workers to create a large number of literary works. This poses a new challenge to the application of Natural language processing technology. Text analysis and processing is realized by Natural language processing technology. In the information society, the amount of data is increasing exponentially, and the number of literary works produced is also rapidly increasing. In order to gain a comprehensive understanding of domestic and foreign history and culture, some Chinese readers are not only satisfied with reading Chinese works from ancient and modern times, but also hope to read and understand foreign literary works. Current mainstream methods for literary character analysis are manual, making the results highly subjective and inefficient for large-scale literary works. To address this problem, this study proposes a character representation and analysis method based on neural networks using English novels as an example. By preprocessing data and utilizing the word dependency relationship to represent character vectors and calculate similarity, the study uses the Skip-gram model to train character vectors and K-means for clustering. An AGA-BPNN model is proposed for character and gender prediction and classification, with a 95.42% accuracy rate achieved in character prediction classification, and an average accuracy, recall, and F1 score of 0.953, 0.962, and 0.962, respectively, in gender prediction and classification. The results demonstrate the effectiveness of the method and propose a new approach for novel character analysis.**

***Keywords*—Neural network; English; literary image; character vector; similarity calculation**

## I. Introduction

In literary works, fiction is one of its important forms of expression. In novel analysis, the analysis of novel characters, including gender, personality, etc., is the basic work to help readers understand novels [1]. Analyzing and researching the characters in novels can help readers understand the characteristics of characters, social environment and the author's thought expression in literary works from the aspects of society, history and literary value [2-3]. In the traditional analysis of characters in novels, the mainstream method is mainly manual. As a result, the analysis results are highly subjective. In addition, the efficiency is low and it takes a long time to analyze large literary works [4]. In China, due to the fact that most modern Chinese literary works are protected by copyright, it is relatively difficult to obtain a large number of Chinese literary texts and conduct analysis and research on them. However, some English literary works are no longer within the copyright protection period. In order to facilitate the processing, analysis, and research of these works, English novels are used as a novel corpus for study and analysis. With the help of Natural language processing technology, analyzing literary works and extracting useful information from them can help readers better read literary works. Using computer to analyze characters in novels requires the use of Natural language processing related technologies to transform the expression of characters in novels into data that can be understood by computers. Neural network (NNs) is an algorithm model that imitates the structure of the human brain and realizes high-precision, high-efficiency distributed parallel information processing. The advantage of NNs is that it can process large-scale data in parallel, and the accuracy and operation speed of the model will not be greatly affected. In classification problems, NNs has important and wide applications [5]. To this end, the study takes English literary works as an example, and builds a model based on the Skip-gram model, K-means and neural network to analyze the characters in the novel. The main purpose of the research is to apply NLP technology and NNs technology to large-scale novel character analysis, so as to improve the efficiency of novel character analysis, strengthen readers' understanding of the work, and help readers better understand the connotation and history of the novel's heritage.

## II. Related Works

In a literary work, characters are one of the three elements of a novel, the core part of a literary work, and an important content that embodies the literary value of a novel. Analyzing and researching the characters in novels can help readers understand the characteristics of characters, social environment and the author's thought expression in literary works from the aspects of society, history and literary value. Therefore, the research on the analysis of literary characters is very common. Shutan MI took the literary characters that appeared in the tenth grade Russian literature class as examples, such as Andrei Bolkonsky and Nikolai Rostov, to reveal the parallelism among literary characters, so as to ensure the effective cognition of student's sex [6]. Ravela took

"White Boys Shuffle" as an example to analyze the experience and growth of the protagonist in the novel, and discussed the racial ideology in the novel, believing that the novel has to some extent promoted the global conceptualization of race [7]. Rebel analyzed the works of three famous writers, and compared the content including the structure, theme, genre and ideology of the works, so as to dynamically analyze the development process of nineteenth-century literature. The analysis results show that ideological disputes have a greater impact on the fate of characters in Turgenev's novels [8]. From a narrative perspective, Bai and others analyzed the protagonist's image and character in the work of Nobel Prize winner William Faulkner - A Rose for Emily, to help readers better understand the thoughts expressed in "A Rose for Emily" [9]. Nischik took the novel "Penelope Piad" as an example to compare the mythical characters in Canadian mythology and Greek mythology, and reimagined the characters in "Odyssey" to reveal the genders contained in the concept of ancient mythology [10]. In a study by Ni, the language, use of behavior, and descriptions of the characters in Miracle are analyzed, and the techniques, intentions, and types used by literary characters to convey between words are explored. After summarizing, the researchers believe that their types include assertiveness, empathy, instruction, and expression [11]. Starkowski believes that in the study of literary works, more energy and attention should be devoted to the study of secondary characters. And taking Dickens' "Still There" as an example, he discusses the inadequacy of the traditional interpretation of secondary characters and the reflection of secondary characters in the novel on the society in the work [12]. In a study by Indrasari et al., taking Bronte's novel Wuthering Heights as an example, they used sociological methods and qualitative description methods to conduct an in-depth analysis of the middle-class female roles in the novel in the 19th century, thereby deepening the understanding of these roles [13].

NNs is an algorithm model that imitates the structure of the human brain and realizes high-precision, high-efficiency distributed parallel information processing. The advantage of NNs is that it can process large-scale data in parallel, and the accuracy and operation speed of the model will not be greatly affected. Therefore, NNs has important and wide applications in classification problems. For example, CNN, which performs well in image recognition and classification, and BPNN, which appears frequently in automatic data classification research, etc. In recent years, scholars have made more and more in-depth research on neural networks. Plonka et al. analyzed the structure of one-dimensional ReLU DNN, and proposed a recursive algorithm to remove parameter redundancy in the deep ReLU neural network (ReLU DNN), thereby optimizing the network structure of ReLU DNN and improving its performance [14]. Raghu et al. discussed the similarity between CNN and visual transformer, and based on this similarity, discussed the possibility and operability of complementary fusion between the two [15]. Jiang designed a photonic device evaluation model based on an improved DNN, and discussed the model's learning of device geometric features in the context of photonics, and the robustness of the model under large-scale data [16]. In order to better understand the problem solving ability and interpretability of problem solving strategies of deep neural networks, Samek et al. conducted a comprehensive discussion and analysis of relevant research literature in recent years. And its application fields, application effects, and application prospects were analyzed and discussed [17]. Chung et al. proposed a neural population geometry method, and used this method to explain and analyze the structure and function of ANN. The actual application effect of the method was tested. The test results verified the effectiveness of the method [18]. Based on the latest research contents, Zhou et al. reviewed the construction, optimization and application of graph neural network (GNN), and discussed the application and performance of GNN variants, pointing out the direction for the development of GNN [19]. Wright et al. designed a deep physical neural network (DPNN) trained by backpropagation, which can effectively reduce the energy consumption required in scientific and engineering applications [20]. Ghosh et al. provided a comprehensive description of the basic structure, foundation, research progress, and main application fields of CNN, and pointed out the important contributions of CNN in the field of artificial intelligence [21]. Kong et al. proposed an Audio Neural Network (PANN) for audio recognition in large-scale data. Tests show that the model outperforms recent state-of-the-art audio recognition models [22].

In summary, CNN can perform image recognition and classification, and has excellent performance in automatic data classification research. In Computer language, ReLU DNN carries out vector training for fictional characters through Natural language processing, and uses mathematical methods to express characters, so that people can be counted and analyzed. The above content indicates that character analysis in literary works is very important, as it is the foundation for readers to understand the novel. However, the existing analysis of characters in literary works is based on a small number of novel characters, which is inefficient, and the analyzed literary works are also extremely limited. Therefore, this study applies neural networks to the analysis of characters in English literary works, and efficiently analyzes a large number of characters in novels through Big data technology and NLP technology. Thus, it can help readers of different novels more conveniently understand the content of the novel, as well as the corresponding characters' history and culture in the novel.

## III. Representation and Application of Literary Characters Based on Neural Network

Neural network technology is the foundation of all subsequent research in this paper. The main content of this chapter is to construct a corpus and preprocess the data using the corpus. At the same time, the Skip gram model is used to train character and feature word vectors, thereby calculating the training of novel character vectors and character similarity. Then, K-means is used to cluster and analyze feature vectors, and an adaptive mutation improved genetic algorithm (AGA) is proposed to obtain the optimal parameters of BPNN and improve model performance. Finally, the feature vectors are input into AGA-BPNN to achieve prediction and classification of feature features and gender.

### *A. Corpus Construction and Data Preprocessing*

In literary works, fiction is one of its important forms of expression. In novel analysis, analyzing the characters in the novel, including gender, personality, etc., is the fundamental work to help readers understand the novel. In recent years, with the rise and development of AI, it is very feasible to use NLP technology to perform automatic, intelligent, and efficient clustering, classification, and analysis of novel characters, which has aroused the interest of some researchers. To this end, the paper takes English literary works as an example, and builds a model based on a neural network to analyze the characters in the novel. The first is to collect data in Project Gutenberg (PG) and build a corpus. PG is a text database that contains a large number of novels of different genres and authors. After the corpus is constructed, the corpus data is preprocessed to facilitate subsequent natural language processing (NLP). The preprocessing of literary works includes lexical and syntactic analysis, clustering of names, etc., as shown in Fig. 1.

The part of speech tagging and Named-entity recognition in Fig. 1 are implemented in this study using conditional random field model (CRF). If there are two groups of random variables $x$ and $y$, when input $x$, the conditional probability distribution of the $p(y|x)$ output is expressed as $y$. $y$ can be regarded as a Markov random field at this time, but $p(y|x)$ is a conditional random field. At this time, the conditional probability is obtained by formula (1).

$$p(y|x,\theta)=\frac{1}{Z(x,\theta)}\exp\left(\sum_{c\in C}\theta_c^T f_c(x,y_c)\right) \quad (1)$$

In formula (1), $\theta_c$ is a weight vector. $\theta$ is the weight vector in any potential energy function. $f(\ )$ is the activation function. $c$ is the largest set of conditional random fields, and $Z(x,\theta)$ is the normalization term, which can be expressed as formula (2).

$$Z(x,\theta)=\sum_y \exp\left(\sum f_c(x,y_c)^T \theta_c\right) \quad (2)$$

In the vocabulary tagging process of the English novel corpus, if the structure of $x$ and $y$ is the same, a linear chain CRF will be formed, as shown in Fig. 2.

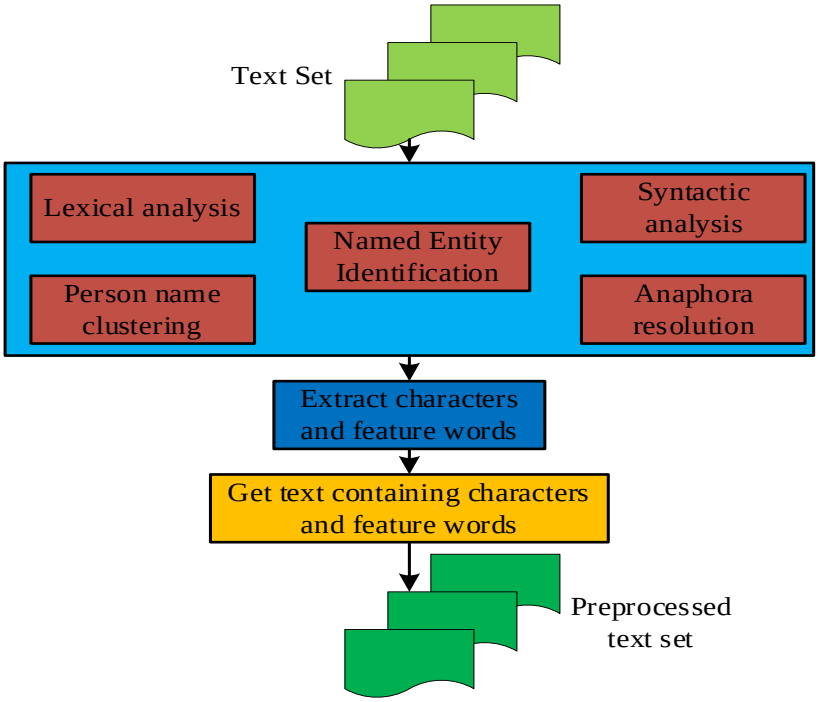



Fig. 1. Pretreatment of literary works.

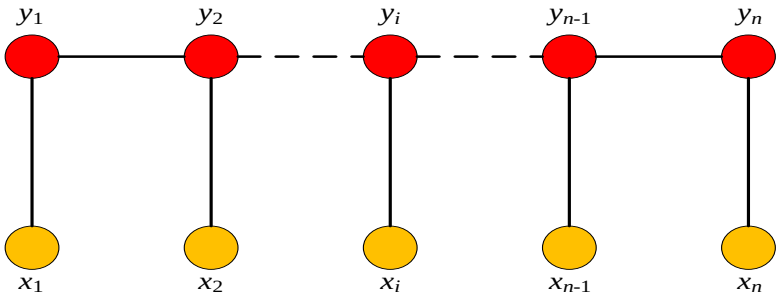



Fig. 2. Linear chain CRF.

The conditional probability at this time is obtained by formula (3).

$$p(y|x,\theta) = \frac{1}{Z(x,\theta)}\exp\left(\sum_{t=1}^{T-1}\theta_1^T f_1(x,y_t)+\sum_{t=1}^{T-1}\theta_2^T f_2(x,y_t,y_{t+1})\right) \quad (3)$$

In formula (3), $f_1(x,y_t)$ is a position $t$-related state feature, and $f_2(x,y_t,y_{t+1})$ is a state transition feature that can be represented by a state transition matrix. In Natural language processing, specific words are divided into five categories according to the dependency grammar relationship between specific words and Chinese characters, namely possessive case (poss), direct object (dobj), noun subject (nsubj), adjective subject (amod) and passive noun subject (nsubjpass). The character words and feature words in the corpus are extracted and expressed as binary phrases (characters and feature words). According to the dependency grammar relationship between feature words and Chinese characters, the top 4 words in each category are selected. The extraction results are shown in Table I. It can be seen that the feature words in Table I are closely related to the behavior or attributes of the characters.

A word, the basic unit of a language, with its own sound, meaning, and syntactic function, expresses ideas and information that are universally and intuitively understood by its native speakers. In terms of novel, a word plays an essential part in composing an outstanding masterpiece, portraying striking characters, and sculpturing a unique style.

The scope of English vocabulary is wide and has different types. For example, from formal to informal, simple to complex, standard to non-standard, common to rare, written to colloquial, general to specific, conceptual to associative, elegant to rough, monosyllabic to polysyllabic, Anglo Saxon to Latin. The artful and skillful choice of words is crucial for achieving special effect and getting desired texture in novels. However, the author does not randomly choose vocabulary, but deliberately chooses vocabulary after carefully considering the purpose, theme, and readers.

Word frequency refers to the times of words appeared in literary works. From Table I, it can be clearly observed that most of the top 4 word in each category are from Anglo-Saxon. English vocabulary mainly comes from Anglo Saxon, French, Latin, and Greek. The vocabulary derived from Anglo Saxon forms the basic vocabulary of English speakers, who have known these words since childhood and frequently use them in daily life. These words are short, simple, basic daily words ---- monosyllables and disyllables, indicating that authors tend to use common words to portray characters, create a brisk rhythm and convey the theme to readers.

Verbs occur as part of the predicate of a sentence and carry markers of grammatical categories, such as person, number, tense, aspect, and mood. Verbs can be stative ones, referring to state of affairs; as well as dynamic ones, referring to actions and events. Table I shows that the most commonly used verbs are simple but dynamic, and many of them refer to characters' movements. The frequently used verbs add rhythmic value to novels and make the characters in novels full of vivid energy.

Nouns are commonly divided into two categories ---- the concrete nouns and the abstract nouns. The former refers to a real, physical thing with a definite indication vividly and clearly; while the latter mostly refers to a quality, state, action, perception or any other implicit concepts. It is eye-catching that top 4 words in poss (possessive case) are all concrete nouns, which transmit specific and accurate meaning and thought to readers and produce vivid impressions.

Adjective can be used to express psychological, physical, auditory, visual, color, evaluative, and emotive attributes. Moreover, they can be used to portray characters and convey the thematic meaning of novels. The abundant adoption of adjective can assist in the vivid portrayal of the characters, enrich the setting of novels, as well as push forward the development of the story. Frequently used adjectives present the beauty of language and charm of novels before readers.

### *B. Character Vector Training and Character Similarity Calculation in Novels*

After preprocessing the corpus, feature words with dependent syntactic relationships can be obtained. At this point, the Skip-gram model is used to train character word vectors and feature word vectors. In a text data set $D$, the probability distribution model is expressed as formula (4).

$$p(D=1|w,c) = \frac{1}{1+e^{-v_w \cdot v_c}} \quad (4)$$

TABLE I. TOP 4 WORDS IN EACH CATEGORY

| Category | Terms | Word frequency | Category | Terms | Word frequency |
|---|---|---|---|---|---|
| poss | hand | 323364 | a mod | old | 89334 |
| | the face | 225147 | | young | 54381 |
| | life | 119574 | | major | 46453 |
| | mind | 101435 | | great | 33680 |
| dobj | tell | 183454 | nsubjpass | call | 19405 |
| | take | 80144 | | bear | 19241 |
| | meet | 58442 | | make | 15403 |
| | reply | 50554 | | know | 14428 |
| nsubj | say | 2122164 | - | - | - |
| | be | 754572 | | - | - |
| | go | 469388 | | - | - |
| | come | 369643 | | - | - |

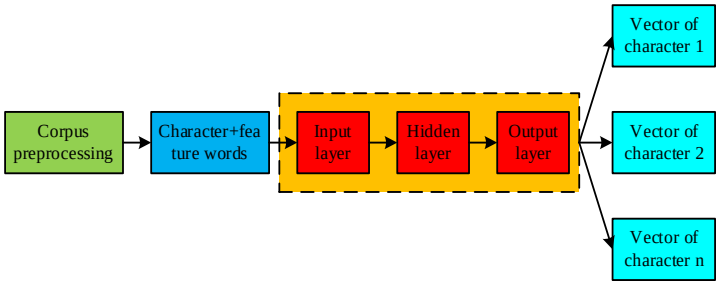
In the formula (4), $w$ and $c$ represent character words and feature words, respectively, and the word pairs formed by the two are $(w,c)$. $v_w$ and $v_c$ represent the vectors of character words and feature words, respectively. $p(D=1|w,c)$ represents the probability of word pairs in $D$. In Equation (4), $v_w$ and $v_c$ are the main learning parameters of the model. In order to maximize the target word pair in $D$, using the objective function of formula (5) to constrain it.

$$\arg\max_{v_w,v_c} \sum_{(w,c)\in D} \log \frac{1}{1+e^{-v_w \cdot v_c}} \quad (5)$$

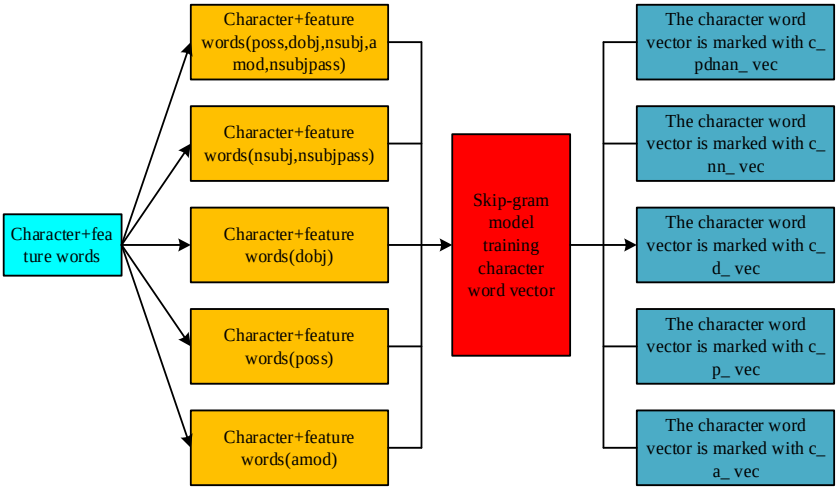
In formula (5), by adjusting the value in the data set $w,c$, the value of $v_w = v_c$ and $v_w \cdot v_c$ is large enough. To use the Skip-gram model to train character word vectors. The process is shown in Fig. 3(a), where n represents the number of character words in the data set.

According to the parts of speech of the five specific words mentioned above, they are divided into five categories. Among them, the first group combines the above characteristic words to represent a vector, where the vector is denoted as c_ Pdnan_ Vec, using this method to represent character vectors; The second group uses nsubj and nsubjpass, with the vector recorded as c_ Nn_ Vec; The third group is through dobj, and the vector at this time is recorded as c_d_vec; The fourth group is through poss, and the vector at this time is recorded as c_p_vec; The fifth group is through amo to represent the characters in the novel, and the vector at this time is recorded as c_a_vec. Fig. 3(b) shows the representation and classification of five groups of character vectors in the process of training character word vectors by the Skip-gram model. In Fig. 3, the trained character vectors all have 200 dimensions. After extracting all characters and related feature words, the character vector is represented by the word heat method, denoted as c_ Pdnan_ Hot. In this method, the characteristic words of any novel character only have one chance to appear.

After the character vector is generated, to calculate the character similarity of the characters in the novel. The principle of character similarity calculation is that in the two novels A and B, if the context of the characters in A is similar to that of the characters in B, it is considered that these two characters are similar, and their vectors in multidimensional space are also similar. The paper adopts the cosine similarity method to calculate the similarity between person vectors. Assuming that the vectors of each dimension of a character in a novel $P$ are expressed as $[P_1, P_2, \ldots, P_{200}]$, and the vectors of each dimension of another character in a novel $Q$ are expressed as $[Q_1, Q_2, \ldots, Q_{200}]$, the cosine similarity value between the two is calculated by formula (6).




(a) Skip-gram model training process of character word vector




(b) Representation and classification of five groups of character vectors

Fig. 3. Word vector training.

$$Sim(P,Q)=\cos(\theta)=\frac{P\cdot Q}{\|P\|\|Q\|} \quad (6)$$

In formula (6), $\theta$ represents the angle between the vectors of the two novel characters $Q$ and $P$ in the vector space. Based on the above content, the extraction of novel character vectors and the calculation of novel character similarity are completed.

### C. Clustering and Classification of Novel Characters Based on Neural Network

In order to obtain the relationship between the characters in the novel more clearly, the study uses K-means to perform cluster analysis on the character vectors. The clustering steps of K-means are shown in Fig. 4.

In the clustering of K-means, the $k$ samples in the data set are divided into $k$ clusters by selecting an initial cluster center; and then iteratively optimizes the cluster center and refines the grouping. If the division of clusters is expressed as $(B_1, B_2, \ldots B_k)$, then the iterative goal of K-means is to minimize the value of formula (7).

$$E=\sum_{i=1}^{k}\sum_{x\in B_i}\|x-\mu_i\|_2^2 \quad (7)$$

In formula (7), $\mu_i$ represents the centroid of the $i$-th cluster, that is, the $B_i$ is the mean vector of all sample data in the $i$-th cluster, which can be calculated by formula (8).

$$\mu_i=\frac{1}{\|B_i\|}\sum_{x\in B_i}x \quad (8)$$

If the input data sample in K-means is expressed as $A=\{A_1, A_2, \ldots, A_m\}$, a sample $k$ is randomly selected in the data set $A$ as the initial $k$ centroid, expressed as $\{\mu_1, \mu_2, \ldots, \mu_k\}$. At this time, the distance between a $i$-th sample $x_i$ and the $j$-th centroid $\mu_j$ is shown in formula (9).

$$d_{ij}=\|x_i-\mu_j\|^2 \quad (9)$$

$d_{ij}$ is the distance between all centroids and the sum $x_i$ according to formula (9). Selecting the smallest centroid of $d_{ij}$ as the new cluster center, and updating the cluster, such as in formula (10).

$$B_{\lambda i}=B_{\lambda i}\cup\{x_i\} \quad (10)$$

In formula (10), it represents the $B_{\lambda i}$ th cluster after updating in the data set. $i$ At this time, the centroid is recalculated. When the vectors of all centroids in the data set are no longer changing, it indicates that the clustering effect at this time is optimal, and the cluster division result is output at this time, see formula (11).

$$B=\{B_1, B_2, \ldots, B_k\} \quad (11)$$

After the character clustering is completed, a model needs to be built to automatically classify novel characters, including character classification and gender classification. Among them, character traits are an important content for readers to understand the characters in novels, and also an important entry point for readers to analyze characters in novels. In psychology, there are many ways to analyze the character of a character. The paper selects the most authoritative and commonly used Myers-Briggs type index to analyze the character of the characters in the novel. The gender prediction is classified according to the characteristics of the characters, such as titles, actions, language, and emotions. The classification model is constructed using a neural network. In the work of automatic data classification, commonly used neural networks include perceptrons, BPNNs, and RBFNNs. This paper builds a classification model based on BPNN, and optimizes BPNN to improve classification performance. Firstly, an improved genetic algorithm with adaptive mutation (AGA) is proposed to obtain the optimal parameters of BPNN. The adaptive mutation probability of AGA is calculated by formula (12).

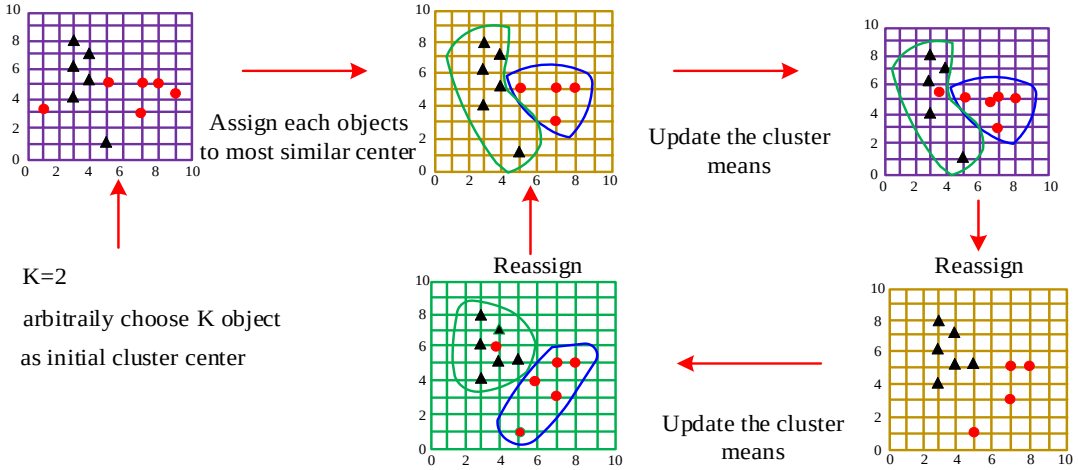



Fig. 4. Clustering steps of K-means.

$$P = \frac{P_1 + P_2}{2} = \frac{\left[P_0 - \left(P_0 - P_{\min}\right) \cdot m\right]}{M} + \frac{\left[P_0 \cdot \max\limits_{X_K \in \Omega} F\left(X_K\right) / \bar{F}\right]}{2} \quad (12)$$

In formula (12), $M$ is the total number of the updates of the algorithm. $m$ is the current iterative update number of the algorithm. $P_0$ is the initial variation probability. $P_{\min}$ is the minimum mutation rate. $\bar{F}$ is the average fitness value $\max\limits_{X_K \in \Omega} F\left(X_K\right)$ of the current population of the algorithm. In any update iteration of the algorithm, the probability of any individual in the population $a_j$ being selected for mutation in this iteration can be calculated by formula (13).

$$z\left(a_j\right) = \frac{f\left(a_j\right)}{\sum_{j=1}^{d} f\left(a_j\right)} \quad (13)$$

In formula (13) of $f\left(a_j\right)$, $a_j$ is the fitness value, and $d$ is the number of all individuals in the current population in AGA. Through the formula (13), the mutation probability of all individuals in the current population is calculated, and according to the accumulated probability value, it is judged which individuals can be inherited to the next generation. By this method, the prematureness of the algorithm can be avoided, and the deficiency of the weak global optimization ability of GA can be made up. Compared with the fixed mutation probability strategy adopted by traditional genetic algorithms, the adaptive mutation probability of genetic algorithms can adjust the mutation probability of the population according to the situation. This can enable the algorithm to better jump out of local optimization and enhance its global optimization ability. Therefore, using AGA to optimize BPNN can find the best parameters of BPNN better and faster and improve the performance of the model. Input the character vector into AGA-BPNN to realize the prediction and classification of character and gender. Based on the above content, the representation of novel characters based on neural network is realized, and then the intelligent and automatic analysis of novel characters is realized.

## IV. Effectiveness Analysis of Character Representation Method Based on AGA-BPNN

### A. Effectiveness of c_pdnan_vec Vector Representation

Based on the Skip-gram model, the study extracts and classifies novel character vectors, uses cosine similarity to calculate the similarity between different novel characters, and then uses K-means to cluster the extracted novel character vectors. In addition, the novel character vector is input into AGA-BPNN for training and learning, and the gender and personality of the novel characters are predicted and classified through AGA-BPNN. Based on the above strategies, the intelligent and automatic analysis and representation of novel characters is realized. The English novels in PG are collected as experimental data to test the effectiveness of the proposed method. First, the validity of c_pdnan_vec character vector representation is verified. Four assumptions commonly used by literature researchers are used to verify c_pdnan_vec, c_nn_vec, c_d_vec, c_p_vec, c_a_vec, c_pdnan_hot. The above six kinds of character vectors are used to represent the similarity between people in the four categories of hypotheses, and the similarity evaluation accuracy of each vector in the four categories of hypotheses is compared. The results are shown in Table II. It can be seen that on the four major assumptions, the average accuracy of the c_pdnan_vec vector representation is 1.00, 0.85, 0.76, and 0.62, and it performs best among the six vector representations. The c_a_vec vector representation has the worst performance. In the first type of hypothesis, the performance of the c_pdnan_hot vector representation is close to that of the c_pdnan_vec vector representation, but on the second to fourth types of hypotheses, the performance of the c_pdnan_vec vector representation is significantly better than that of the c_pdnan_hot vector. This is because c_ pdnan_ Vec vector representation is more comprehensive and can reflect the character characteristics from many aspects. The test results verify the correctness of the c_pdnan_vec vector representation.

### B. K-means-based Novel Character Cluster Analysis Effect

The paper uses the K-means algorithm to cluster novel characters. In order to verify the application effect of K-means in novel character clustering, it is compared with K-MEDOIDS, fuzzy C-means algorithm (FCM) and density-based clustering algorithm (DBSCN). Purity and clustering accuracy (ACC) are used to compare the performance of several methods on novel character clustering. See Table III for the Purity values of several methods. It can be seen that in the five experiments, the average value of K-means Purity is 0.75, which is 0.12, 0.11, and 0.13 exceed that of K-MEDOIDS, FCM, and DBSCN, respectively. These data show that the clustering effect of K-means is better. The distance between the readers and the characters in the novels can be shortened, and the distance between the readers and the author can also be shortened. Readers are more likely to feel that they are invited to experience the whole story.

The ACC are shown in Table IV. In the five experiments, the average ACC of K-means is 0.86, which is 0.05, 0.05 and 0.08 exceed that of K-MEDOIDS, FCM and DBSCN respectively. The above results indicate that K-means has a better application effect in novel character clustering and is more suitable for clustering analysis of novel characters, verifying the correctness of the K-means algorithm. This can achieve the goal of vividly portraying characters and creating a certain rhetorical effect.

TABLE II. SIMILARITY EVALUATION ACCURACY OF EACH VECTOR IN THE FOUR CATEGORIES OF ASSUMPTIONS

| Vector representation | Number of experiments | Four categories of assumptions | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| c_pdnan_vec | 1 | 1.00 | 0.86 | 0.76 | 0.68 |
| | 2 | 1.00 | 0.83 | 0.77 | 0.65 |
| | 3 | 1.00 | 0.87 | 0.75 | 0.62 |
| | Average | 1.00 | 0.85 | 0.76 | 0.65 |
| c_nn_vec | 1 | 0.80 | 0.68 | 0.18 | 0.25 |
| | 2 | 0.82 | 0.66 | 0.20 | 0.21 |
| | 3 | 0.75 | 0.64 | 0.17 | 0.25 |
| | Average | 0.79 | 0.66 | 0.18 | 0.24 |
| c_d_vec | 1 | 0.25 | 0.25 | 0.17 | 0.27 |
| | 2 | 0.20 | 0.26 | 0.19 | 0.29 |
| | 3 | 0.23 | 0.28 | 0.18 | 0.25 |
| | Average | 0.23 | 0.26 | 0.18 | 0.27 |
| c_p_vec | 1 | 0.50 | 0.50 | 0.48 | 0.14 |
| | 2 | 0.54 | 0.51 | 0.52 | 0.12 |
| | 3 | 0.45 | 0.53 | 0.53 | 0.11 |
| | Average | 0.50 | 0.51 | 0.51 | 0.12 |
| c_a_vec | 1 | 0.00 | 0.00 | 0.32 | 0.00 |
| | 2 | 0.00 | 0.00 | 0.30 | 0.00 |
| | 3 | 0.00 | 0.00 | 0.31 | 0.00 |
| | Average | 0.00 | 0.00 | 0.31 | 0.00 |
| c_pdnan_hot | 1 | 1.00 | 0.45 | 0.50 | 0.53 |
| | 2 | 1.00 | 0.43 | 0.49 | 0.52 |
| | 3 | 1.00 | 0.42 | 0.54 | 0.53 |
| | Average | 1.00 | 0.43 | 0.51 | 0.53 |

TABLE III. PURITY VALUE OF SEVERAL METHODS

| Number of experiments | Purity value | | | |
|---|---|---|---|---|
| | K-means | K-MEDOIDS | FCM | DBSCN |
| 1 | 0.75 | 0.63 | 0.64 | 0.55 |
| 2 | 0.73 | 0.66 | 0.52 | 0.54 |
| 3 | 0.75 | 0.58 | 0.68 | 0.63 |
| 4 | 0.78 | 0.69 | 0.70 | 0.71 |
| 5 | 0.74 | 0.59 | 0.65 | 0.65 |
| Average | 0.75 | 0.63 | 0.64 | 0.62 |

TABLE IV. ACC VALUE OF SEVERAL METHODS

| Number of experiments | Purity value | | | |
|---|---|---|---|---|
| | K-means | K-MEDOIDS | FCM | DBSCN |
| 1 | 0.85 | 0.80 | 0.76 | 0.76 |
| 2 | 0.91 | 0.82 | 0.85 | 0.75 |
| 3 | 0.87 | 0.84 | 0.82 | 0.72 |
| 4 | 0.82 | 0.83 | 0.77 | 0.84 |
| 5 | 0.86 | 0.78 | 0.86 | 0.82 |
| Average | 0.86 | 0.81 | 0.81 | 0.78 |

### *C. Effect Analysis of AGA-BPNN Classification Model*

To verify the application effect of AGA-BPNN model in novel character gender prediction classification, this paper borrowed the Myers Briggs type index to analyze the personality characteristics of novel characters in personality prediction classification. Recording several personalities of the characters as 1~5. Compare the AGA-BPNN, GA-BPNN, and BPNN models to classify character personalities under different sample sizes, and verify the optimization effect of AGA on BPNN, as shown in Fig 5. In Fig. 5, when the number of samples ranges from 0 to 80, the prediction accuracy of the three models for character personality is high. When the number of samples exceeds 80, the prediction accuracy of BPNN drops significantly, the accuracy of GA-BPNN model decreases to a certain extent, while the accuracy of AGA-BPNN model hardly changes. Overall, the accuracy of the AGA-BPNN model reaches 95.42% when performing character prediction and classification of novel characters; while the accuracy of the GA-BPNN model reaches 90.66%, which is 4.76% lower than the AGA-BPNN model; the accuracy of the BPNN model reaches 86.53% %, 8.89% lower than the AGA-BPNN model. It can be seen that in the prediction and classification of characters in novels, the accuracy of the AGA-BPNN model is significantly better than the other two models.

The same data is used to evaluate the gender prediction and classification ability of the above three models. Accuracy, Recall and F1 are selected as evaluation indicators. Fig. 6

shows the gender prediction and classification performance of the three models. It can be seen that the average accuracy, average recall and average F1 values of AGA-BPNN are 0.953, 0.962 and 0.929, respectively, which exceed GA-BPNN and BPNN models. The above results show that the model proposed in the study has better performance in gender prediction. This is because after the optimization of the AGA algorithm, the accuracy and convergence of the BPNN model have been improved, thus improving the accuracy of the BPNN model for novel character classification. The average accuracy of the BPNN model reached 0.929, which has higher accuracy compared to the ReLU DNN network structure proposed by Plonka G et al. and the DPNN network proposed by Wright L G et al. To sum up, the methods proposed in the paper can accurately realize intelligent novel character representation and analysis, thus helping readers better understand the novel.

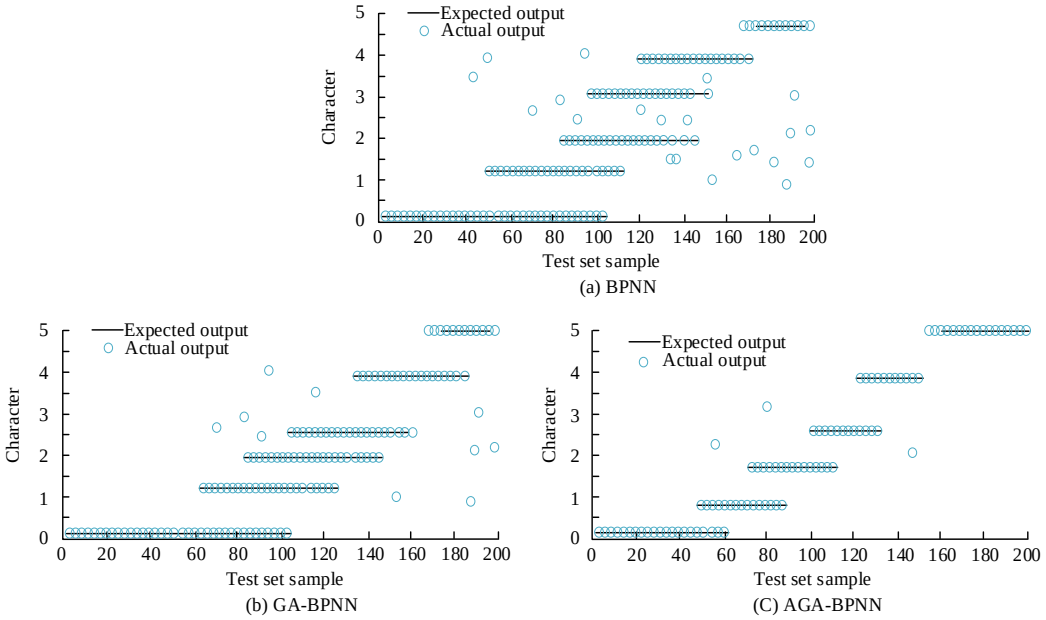



Fig. 5. Accuracy of AGA-BPNN classification of character.

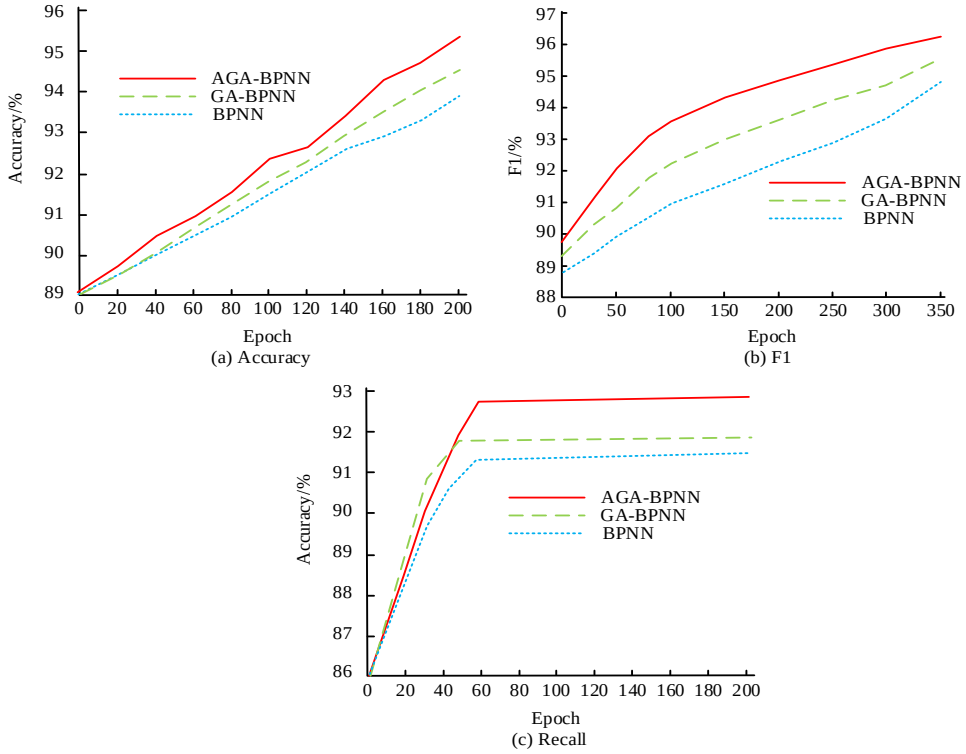



Fig. 6. Gender prediction and classification ability of three models.

## V. CONCLUSION

Character analysis in literary works is very important, and it is the basis for readers to understand novels. However, the existing analysis of characters in literary works is based on a single or a small number of novel characters, which is inefficient. Aiming at this problem, the paper is based on the Skip-gram model and the AGA-BPNN model to realize the vector representation. In the experiment, on the four assumptions, the average accuracy of c_pdnan_vec vector representation is 1.00, 0.85, 0.76, 0.62. Therefore, c_ Pdnan_ Vec vectors perform the best in terms of comprehensiveness among the six vector representations. The performance of VEC vector representation is the worst, which verifies the validity of c_pdnan_vec vector representation. The average value of K-means Purity is 0.75, 0.12, 0.11, and 0.13 exceed K-MEDOIDS, FCM, and DBSCN, respectively. The average value of ACC is 0.86, which is 0.05, 0.05, and 0.08 exceed K-MEDOIDS, FCM, and DBSCN, respectively. This confirms the effectiveness of K-means in novel character clustering work. When predicting and classifying characters in novels, the accuracy of the AGA-BPNN model reaches 95.42%, which is 4.76% exceed that of the GA-BPNN model and 8.89% exceed that of the BPNN model. When the sample size is between 0 and 80, the three models have higher prediction accuracy for character traits. When the number of samples exceeds 80, the prediction accuracy of BPNN decreases significantly, and the accuracy of GA-BPNN model decreases to a certain extent, while the accuracy of AGA-BPNN model remains almost unchanged. The accuracy of the AGA-BPNN model is significantly better than the other two models in predicting and classifying the personalities of novel characters. In the gender prediction classification of novel characters, the average accuracy rate, average recall and average F1 value of AGA-BPNN are 0.953, 0.962 and 0.929, respectively, which are exceed those of GA-BPNN and BPNN. The average accuracy of the BPNN model reached 0.929, which has higher accuracy compared to the ReLU DNN network structure proposed by Plonka G et al. and the DPNN network proposed by Wright L G et al. Compared to Bai X et al.'s approach of analyzing novel characters from a narrative perspective, research based on the Skip gram model and AGA-BPNN model provides a better understanding of character characteristics and article ideas. Therefore, the method proposed in the paper can accurately realize intelligent novel character representation and analysis, thereby helping readers to better understand the novel. However, the novel text data used in the experiment are all English novels, so the experiment only verifies the effect of the proposed method in the analysis of characters in English novels. In-depth research is needed in the future to verify the application effect of the proposed method in the analysis of characters in novels in other languages.

## REFERENCE

[1] Dara C, Simanjuntak M B. Representation of Standard Language on The Dilan Characters in The Novel" Dilan 1990". LITERACY: International Scientific Journals of Social, Education, Humanities, 2022, 1(2): 57- 68.

[2] Wiryadiningsih K, Indiaatmoko B. The Literary Style of Javanese Female Characters in the Novel Jemini by Suparto Brata. Seloka: Jurnal Pendidikan Bahasa dan Sastra Indonesia, 2020, 9(2): 147-158.

[3] Shalimova DV, Shalimova I V. Peter Newmark's Translation Procedures as Applied to Metaphors of Literary Texts (Based on Stephen King's Works). Bulletin of Kemerovo State University, 2020, 22(1):278-287.

[4] Boulogne P. And now for something completely different... Once again the same book by Dostoevsky: A (con) textual analysis of early and recent Dostoevsky retranslations into dutch. Cadernos de Tradução, 2019, 39: 117 -144.

[5] Yang GR, Wang X J. Artificial neural networks for neuroscientists: a primer. Neuron, 2020, 107(6): 1048-1070.

[6] Shutan M I. The parallelism of the characters in the literature class in the 10th grade: Andrei Bolkonsky and Nikolai Rostov. Literature at School, 2020(1, 2020):68-78.

[7] Ravela C. "Abandoning This Sinking Ship America": The Classical Bildungsroman, Minor Characters, and the Negative Dialectic of Race in Paul Beatty's White Boy Shuffle. Genre, 2020, 53(1):27-52.

[8] Rebel G M. Out of Time Characters in Literary Works Of 1859: "Family Happiness" By Lev Tolstoy, "Oblomov" By Ivan Goncharov, "A House of Gentlefolk" By Ivan Turgenev. Bulletin of Udmurt University Series History and Philology, 2020, 30(5):859-869.

[9] Bai X, Zhang X, Li Y. An Analysis of Emily's Characters in A Rose for Emily from the Perspective of Narration. Journal of Language Teaching and Research, 2020, 11(4): 611-615.

[10] Nischik R M. Myth and Intersections of Myth and Gender in Canadian Culture: Margaret Atwood's Revision of the Odyssey in The Penelopiad. Zeitschrift für Anglistik und Amerikanistik, 2020, 68(3): 251-272.

[11] Ni W. The illocutionary acts of the characters in wonder A novel by RJ Palacio. International Journal of Linguistics Literature and Culture, 2020, 6(3):36-40.

[12] Starkowski K H. "Still There": (Dis)engaging with Dickens's Minor Characters. NOVEL A Forum on Fiction, 2020, 53(2):193-212.

[13] Indrasari DN, Rahman F, Abbas H. Middle Class Women Role in the 19th Century as Reflected in Bronte's Wuthering Heights. ELS Journal on Interdisciplinary Studies in Humanities, 2020, 3(2):214-218.

[14] Plonka G, Riebe Y, Kolomoitsev Y. Spline representation and redundancies of one-dimensional ReLU neural network models. Analysis and Applications, 2022, 21(01):127-163.

[15] Raghu M, Unterthiner T, Kornblith S, Zhang CY, Dosovitskiy A. Do vision transformers see like convolutional neural networks? Advances in Neural Information Processing Systems, 2021, 34: 12116-12128.

[16] Jiang J, Chen M, Fan J A. Deep neural networks for the evaluation and design of photonic devices. Nature Reviews Materials, 2021, 6(8): 679-700.

[17] Samek W, Montavon G, Lapuschkin S, Anders CJ, Müller K R. Explaining deep neural networks and beyond: A review of methods and applications. Proceedings of the IEEE, 2021, 109(3): 247- 278.

[18] Chung SY, Abbott L F. Neural population geometry: An approach for understanding biological and artificial neural networks. Current opinion in neurobiology, 2021, 70: 137-144.

[19] Zhou J, Cui G, Hu S, Zhang ZY, Yang C, Liu ZY, Wang LF, Li CC, Sun M S. Graph neural networks: A review of methods and applications. AI open, 2020, 1: 57-81.

[20] Wright LG, Onodera T, Stein MM, Wang TY, Schachter DT, Hu Z, McMahon P L. Deep physical neural networks trained with backpropagation. Nature, 2022, 601(7894): 549-555.

[21] Ghosh A, Sufian A, Sultana F, Chakrabarti A, De D. Fundamental concepts of convolutional neural network. Recent trends and advances in artificial intelligence and Internet of Things, 2020: 519-567.

[22] Kong Q, Cao Y, Iqbal T, Wang WW, Plumbley M D. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 2880-2894.