J. COMBIN. MATH. COMBIN. COMPUT. 127b (2025) 5013--5028

Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

# A Study on Enhancing the Efficiency of Digital Content Generation and Distribution Using Natural Language Processing Techniques

Yueying Wang[1,2], Mingqi Li[3], Lin Sun[3], Qinrong Xu[1,✉]

[1] *School of Publishing, University of Shanghai for Science and Technology, Shanghai, 200000, China*

[2] *School of Energy and Power Engineering, University of Shanghai for Science and Technology, Shanghai, 200000, China*

[3] *School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai, 200000, China*

ABSTRACT

Under the background of the digital era, the self-media platform breaks the information barriers between the communicators and the receivers, effectively alleviating the information asymmetry problem between the two. Through observation and research, this paper finds that the current channels for receivers to obtain digital information can be divided into user-generated content (UGC), professional-generated content (PGC), and brand-generated content (BGC) according to the classification of the main body, but most of the managers are negligent in the management of these digital contents, and do not really utilize the value of their dissemination. Digital content generation and dissemination based on natural language processing (NLP) technology has become an important way to solve this problem. The method is based on the unified processing of a large amount of corpus, input Word2vec model and Skip-gram model two types of language models for training, with the obtained language model for the required text can be obtained word vectors, the different lengths of the text will be unified vectorization. By introducing evaluation indexes such as dissemination efficiency, content quality and coverage, the effect of generated content can be measured objectively. The value of generating digital content to improve the dissemination efficiency is verified through the evaluation of the actual effect.

*Keywords:* natural language processing; digital content generation; word2vec model; word vector; dissemination efficiency

✉ Corresponding author.
*E-mail address:* lwyx_01@126.com (Q. Xu).

Received 25 February 2024; Revised 05 April 2024; Accepted 15 December 2024; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-279

## 1. Introduction

Digital content industry is an important component of the cultural industry, and is also a refraction of the state of development of the cultural industry in the digital era. With the deepening of the integration of the digital economy and the real economy, the overall human demand for the total amount and richness of digital content continues to increase, and the massive demand for digital content supply pulls the application of digital content generation technology to the ground. Digital content generation technology has set off a content creation boom, reshaping and even subverting the production mode and consumption mode of digital content.

Digital content generation technology refers to the use of artificial intelligence technology to automatically generate content, and the emergence of machine learning technology provides strong support for digital content generation technology [1]. The outstanding advantages of machine learning technology in digital content creation, represented by natural language processing, can effectively improve the content production efficiency, to a certain extent, the human resources from the content production link to free, and then find more energy to invest in creative thinking, strategy development, to help create a high-quality digital content ecology [2-4]. The generated digital content has the characteristics of authenticity, diversity, and controllability, which helps enterprises and individuals to improve the efficiency of content production, and provide more rich and diversified, dynamic and interactive content, which has a wide range of application prospects [5-6].

The digital content industry is regarded as a new knowledge-integrated industry that combines high technology and new ideas, covering a number of subfields, and natural language processing techniques play an important role in relation to enhancing the efficiency of digital content generation in the field. Literature [7] utilizes natural language processing and image generation techniques to create text-sourced visuals and voice-overs to efficiently generate multimedia-format content, demonstrating the potential of this approach in areas such as automated media content creation. Literature [8] proposes a natural language processing architecture for extracting large entities and relational networks, which enables rapid access to scientific knowledge graphs in academic domains and effectively facilitates the management, analysis, dissemination and processing of scientific knowledge. Literature [9] designed a data enhancement method based on natural language processing techniques to improve the text generation performance of long text and short text classifiers by establishing generalized text transformation rules, which was successfully validated on different types of datasets. Literature [10] examined the application of natural language processing techniques in the field of music information retrieval, using NLP tools to accomplish symbolic music generation and information retrieval tasks by calculating symbolic sequences, which accelerated the efficiency of music data processing. Literature [11] shows that the generation of high-quality test questions in the field of education is a complex task, for this reason, a test question generation method based on natural language processing and topic modeling is proposed, and attention and pointer generator mechanisms are introduced, which show good performance in digital content generation. Literature [12] analyzes the answer generation mechanism including question analysis, answer extraction and paragraph retrieval in Q&A systems, and proposes the use of natural language processing techniques to achieve coherent and accurate automatic answers based on the full recognition of contextual data. Literature [13] describes the business process of generating e-learning resources based on natural language processing technology and packages it into independent and reusable web services to effectively solve the problem of shortage of learning resources in e-training. Literature [14] establishes an automated movie script generation model, which automatically generates movie scripts and storytelling systems by processing and recognizing a large amount of movie textual data and visual information in order to extract deep features.

This paper focuses on the main problem of how NLP technology affects the generation and dissemination of digital content, takes the Word2vec model, skip-gram model and NLP word embedding model as the core theoretical basis, and comprehensively uses questionnaire survey, manipulation variable test and other methods to try to answer the following questions: First, based on the embodied perspective of NLP technology, the logical basis of the impact of NLP technology on the generation and dissemination of digital content is analyzed. Second, the technical evolution of NLP is reviewed and sorted out, and the enhancement of the efficiency of digital content generation and dissemination by three digital content types, namely, UGC, PGC, and BGC, is explored.

## 2. NLP-based digital content generation technology

### *2.1. Natural Language Processing (NLP) Technologies*

Nowadays, with the rapid development of artificial intelligence and data acquisition devices, the efficiency of digital content in collecting, organizing, storing, and generating and utilizing has been greatly improved. A large amount of high-quality, multi-scale relevant data is rapidly accumulating, which provides an important source of clues and data basis for the training of big language models. However, the utilization efficiency of textual data is generally low in the training process of big language models, and a large amount of information contained in textual data is still to be mined.

Natural Language Processing (NLP) is an emerging discipline and technology that has been gradually formed after the introduction of computers, when people try to explore ways to process natural language in an automated way to meet the application needs in all kinds of specific situations [15]. NLP itself is an interdisciplinary that integrates the knowledge and achievements of multiple fields such as computer science, artificial intelligence, linguistic science, logic, psychology, and so on. NLP technology has made some progress in semantic analysis, lexical annotation, entity recognition, machine translation, sentiment analysis, automatic digest, syntactic analysis, and co-reference disambiguation, etc., and it has a broad space for exploration.

Word embedding is a collective term for a set of language modeling and feature learning methods technical processes in natural language processing (NLP) that map words, phrases, or sentences from a vocabulary to specific locations in a vector space of high-dimensional real numbers. Methods for generating such mappings include neural networks, probabilistic models, dimensionality reduction of word covariance matrices, interpretable knowledge bases, and other methods. Word and phrase embeddings when used as the underlying input representation can improve the performance of many downstream natural language processing tasks. With the exploding demand for digital content, traditional methods of digital content generation suffer from inefficiency and high cost. NLP techniques have the ability to transform unstructured text into numerical, structured information and enable computers to understand human language through methods such as machine learning and deep learning. Through the NLP framework, this paper aims to achieve the dual goals of improving the efficiency of digital content generation and optimizing the dissemination effect.

### *2.2. Word Vector Acquisition and Application*

2.2.1. Word2vec modeling. Before the birth of word embedding methods such as Word2vec model, natural language processing processes usually represented words as single, discrete numbers. Among them, the most intuitive representation is One-hot Representation, which represents words as high-dimensional vectors composed of 0 and 1, the size of the dimension is the size of the vocabulary in the corpus, and each word has a value of 1 in its specific dimension, and the rest of the positions are filled with 0. Obviously, this method cannot fully express the semantic information, and it will result in the dimensionality explosion in the computation process.

Both models of Word2vec model contain three layers, i.e., input layer, projection layer and output layer. Among them, the structure of CBOW model and Skip-gram model is shown in Fig. 1, which predicts the current word $w(t)$ on the premise that the context $w(t-2)$, $w(t-1)$, $w(t+1)$, $w(t+2)$ of the current word $w(t)$ is known, while the Skip-gram model predicts the current word $w(t)$ on the premise that its context $w(t-2)$, $w(t-1)$, $w(t+1)$, $w(t+2)$ is known.

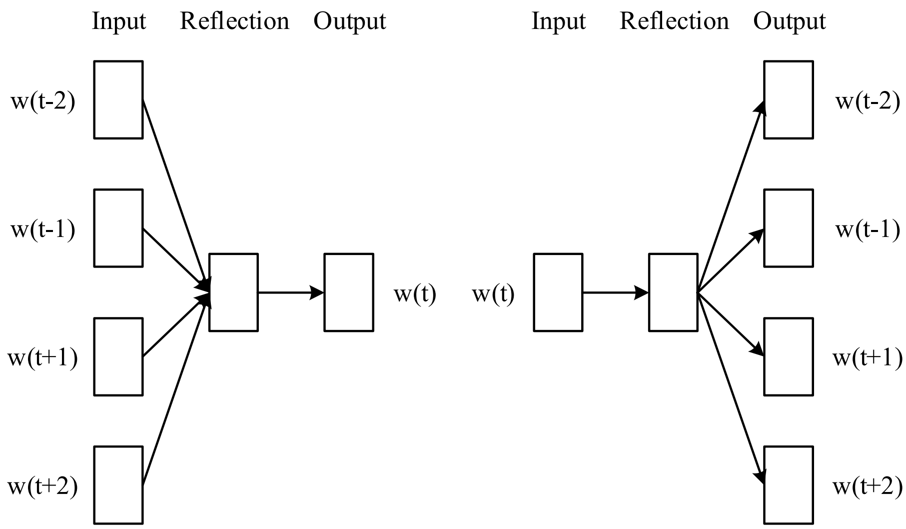


**Fig. 1.** Structure diagram of CBOW and Skip-gram model

2.2.2. Acquisition of word vectors. In this paper, we choose one of the Word2vec Skip-gram models to embed words in the digital content generation system, Skip-gram is a simple but practical NLP word embedding model, the specific structure is shown in Figure 2. To obtain the word vector of a word, the unique heat code of the word is entered in the input layer, the linear mapping of the code is entered in the first hidden layer $W \times x + b$ ( $x$ is the unique heat code of the word, $W$ and $b$ are the parameters), and the third layer can be regarded as a classifier, using Softmax regression. The training samples are one-hot coded word vectors of such word pairs, and the output of the final model is a probability distribution, i.e., the probability of occurrence of each word of the vocabulary list in the contextual word fetching window of the input word.

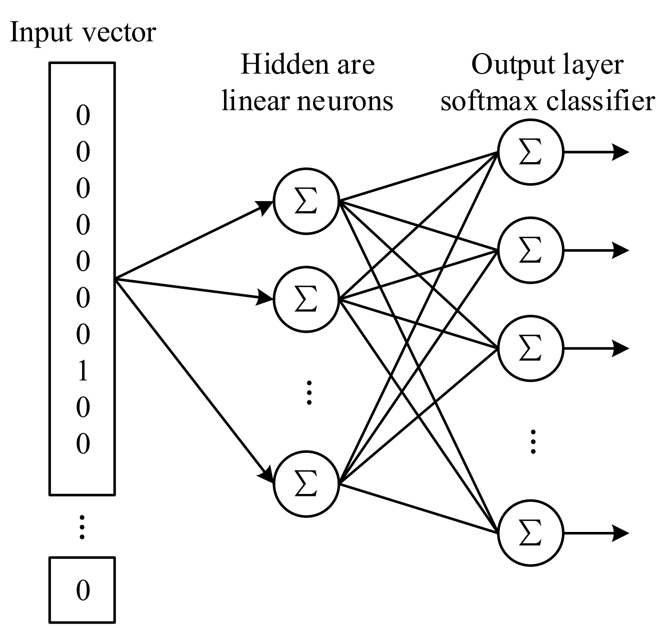


**Fig. 2.** Structure diagram of Skip-gram model

The idea of the application of this model is similar to that of the autocoder in that a neural network is first constructed based on the training data. After training this network model using the corpus built at work, we do not utilize it for new context prediction tasks, but need the parameters learned from the training data, i.e., the weight matrices of the hidden layer, and the word vectors are obtained from the transformation of these weight parameters.

2.2.3. Application of word vectors. In order to test the validity of the Skip-gram model trained by the corpus, and the resulting word vectors, nouns with continuity patterns can be selected to calculate the semantic similarity, and in this paper, different numerical contents are selected to complete the validity verification.

Word vectors are vectors that convert the words they represent into a particular dimension, but Chinese sentences are composed of numerous words, to get sentence vectors composed of multiple words, the following methods can be used:

1) Accumulation. The cumulative method is done by superposing the word vectors of all the non-disabled words in the sentence. If the sentence has $n$ non-disabled word, the word vector of the sentence is obtained by the following means:

$$V_{sentence} = V_{word\,1} + V_{word\,2} + \ldots\ldots + V_{word\,n} \tag{1}$$

2) Averaging. The averaging method is similar to the cumulative method method in that it also involves superimposing all the non-discontinued word vectors in a sentence, but in the end, the number of non-discontinued words needs to be added at the superimposed vectors in terms of the number of non-discontinued words. The word vectors of a sentence are obtained by the following means:

$$V_{sentence} = \frac{(V_{word\,1} + V_{word\,2} + \ldots\ldots + V_{word\,n})}{n} \tag{2}$$

Based on the above method, the sentence vectors in the corpus can be obtained, and the digital content generation task can be better accomplished, as well as the construction of the digital content generation system.

# 3. Strategies for improving the efficiency of digital content generation and dissemination

## *3.1. Theoretical basis of NLP technology to enhance communication efficiency*

The enhancement of the efficiency of digital content dissemination by NLP technology is a real phenomenon, and NLP technology empowerment is an inevitable trend of digital content dissemination. From intermediary to subject, from tool to environment, NLP technology plays an increasingly important role in digital content dissemination activities. The development of products such as ChatGPT and Wenxin Yiyan as digital content generative artificial intelligence all reflect the latest achievements of NLP technology, which have revolutionized digital content dissemination [16]. If digital content generation is regarded as a powerful technology purely from the perspective of instrumental rationality, it obviously loses the holistic and humanistic nature of scrutinizing social phenomena. In the new era, the symbiosis between NLP technology and human beings has become a new way of living in the society in general, and the integration and development of NLP technology and digital content dissemination will become an inevitable choice.

3.1.1. Logic of practice. From the perspective of media technology evolution over time, the invention of a new technology always dominates and becomes the core technology that shapes the media environment at that time, but this does not mean that the old technology disappears, and the old and new technologies still shape the macro-technological environment in a coexisting and symbiotic relationship. On a static plane, the different attributes and functions of each medium bring about different sensory biases. The practice of embodiment of digital content dissemination is precisely accompanied by the dynamic evolution of the media environment and the sensory bias of media attributes, and the body is the subject of the practice of technological embodiment. As a matter of fact, the current media environment is an integrated system of multiple media symbols, and spoken, written, printed, electronic and digital media are all important technological elements of digital content dissemination.

Embodied practice is a social activity in which people acquire information, generate information, and disseminate digital content in the objective world, and digital content exists and flows, either implicitly or explicitly, in the interaction between the media environment and bodily practices. Taking the relationship between body and media as an entry point to understand digital content dissemination again, the relationship between NLP technology-body practice-digital content dissemination is shown in Figure 3. Digital content dissemination in different media environments presents different types of characteristics, from synchronous interactive digital content dissemination, asynchronous unidirectional digital content dissemination, to digital content dissemination in mimetic environments, and multimodal digital content dissemination. Among them, multimodal digital content dissemination implies embodied dissemination practices in a media environment of multiple perceptions co-constructed in the form of NLP technologies.

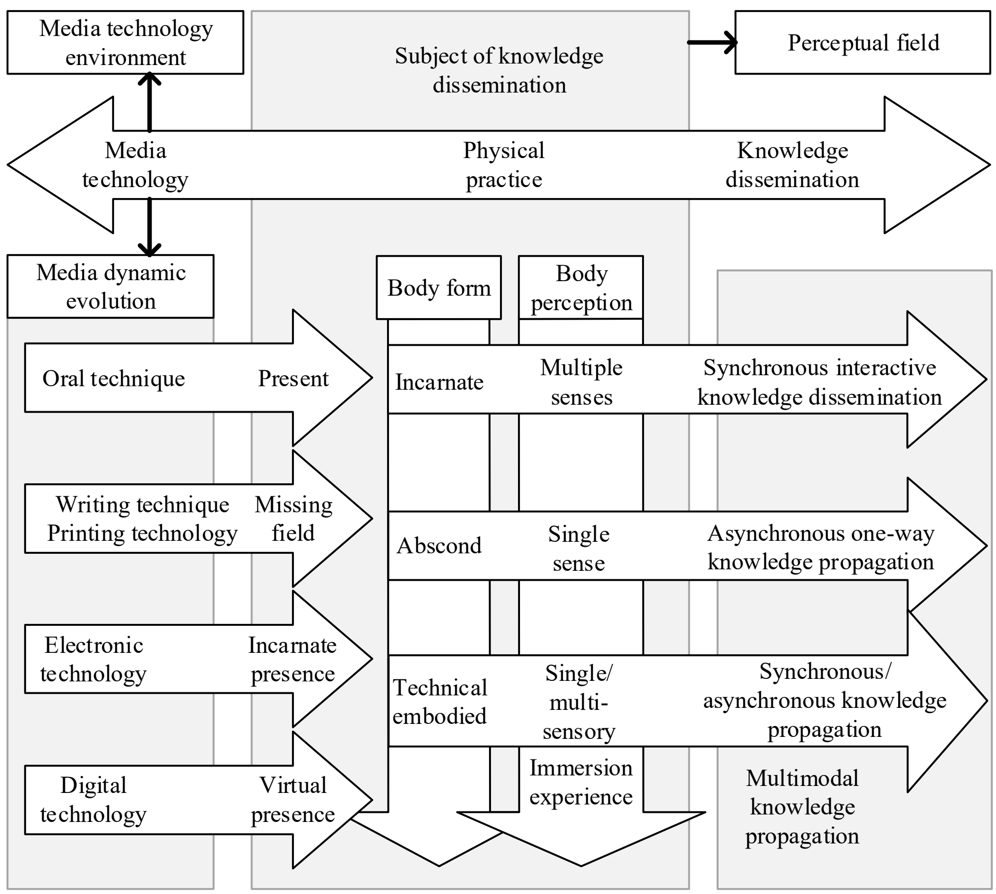


**Fig. 3.** Embodied practice of digital content dissemination under NLP technology

The form of the body in the physical practice of digital content dissemination has undergone multiple shifts due to different technological environments, from multi-sensory interaction with physical participation, to single-sensory input mediated by disembodied knowledge carriers, to multi-sensory involvement in technological embodiment and virtual immersive experience, the evolution of the media environment of digital content dissemination has continuously refreshed and reshaped the perceptual field of the body. Looking at digital content dissemination from the perspective of the relationship between technology and the body, we can find that digital content dissemination activities have never been separated from the process of interaction between the media and the body, and that the association between the body and technology has a natural attribute.

3.1.2. Presentation. The digital environment for digital content generation and dissemination empowered by NLP technology is generated and presented in the synergistic co-construction of digital media interface, digital application scene, and digital learning space. The interface-scene-space is a mutually complementary and symbiotic relationship as shown in Figure 4.

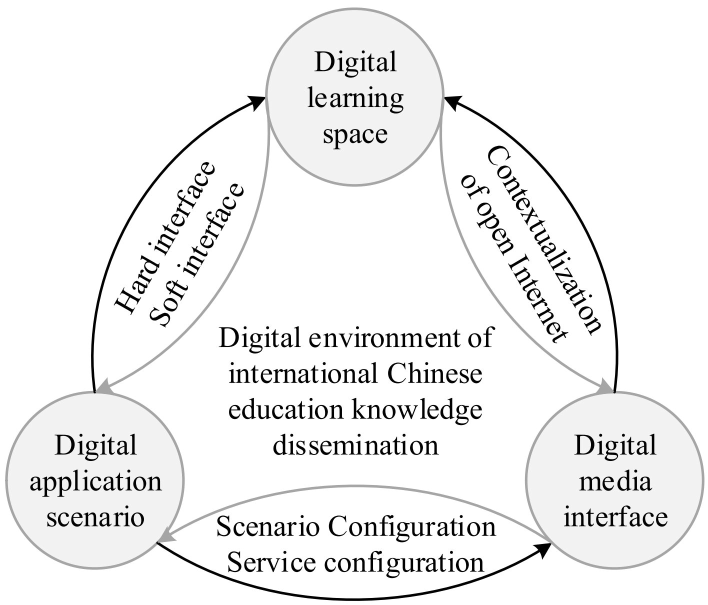


**Fig. 4.** Digital environment presentation of digital content generation and dissemination

1) Media interface for digital content distribution. The interface for the dissemination of digital content can be explained by Michael Heim's dual definition of interface, that is, the interface is both the peripheral device of the computer and the display screen, and the activity of the person who is connected to the data through the display. Digital content dissemination presents knowledge content in front of the educational subject through the "interface" of digital media, which not only has the materiality of connecting the physical space to enable people to achieve the state of "getting started" of the media, but also has the digitality of gathering knowledge resources in the virtual space. In other words, the dissemination of digital content must be achieved through the interface between people and media. In communication, "interface is usually the point of contact between the receiver of a message and the message". The media interface of digital technology is an intermediary system that connects the existence of the body with the objective world, connecting the body with the knowledge content and resources stored in the digital environment, and the body carries out knowledge learning, knowledge understanding, knowledge production and meaning expression through the digital media interface.

2) Application Scene of Digital Content Dissemination. In the digital era, the scene is the constituent element of digital space and embodied environment, and also the environmental element of technology embodied perception, which is related to the spatial environment and physical perception environment created by digital technology. The application scene of digital content dissemination has changed from traditional and single communication scene to virtual and diversified communication scene. The specific connotation of the scene includes the scene configuration within the digital media interface, the service adaptation of the digital application scene, and the virtual and real

transformation of the technology embodied scene. For the subject of digital content dissemination, the knowledge resources and knowledge services in the learning scene should be in line with the needs and expectations, so that he or she can obtain a perception experience that matches his or her needs in the scene.

The digital media interface and the form of knowledge interaction of the learning scene should match the perception ability of the subject. From the point of view of the virtual and real transformation of the technological embodiment scene, the scene transformation is the body in the physical space, virtual space and social space and other multi-dimensional space of the transition, individuals in the technological embodiment of the “transitions” in the perception, experience, cognition and behavioral results of the differences. Scene transformation is realized in the digital media environment, and the digital media environment needs the support of interwoven scenes, and the two are mutually exclusive. The digital scene transformation of digital content generation and dissemination implies that the body realizes the seamless connection of dissemination time, the extensive connection of dissemination roles, and the deep coupling of dissemination content in the scene transformation of real and virtual space.

### *3.2. Digital Content Generation and Distribution Model for NLP*

In the digital era, the dissemination carrier and dissemination path of digital content have changed significantly, and MOOC, short videos, online Q&A communities, etc. have greatly simplified the process of people's access to digital content and reflected the interactivity with users' needs, i.e., according to the feedback of users, the digital content and methods provided are continuously improved [17]. The emergence of NLP technology has aroused a new thinking about the relationship between digital content generation and dissemination, on the one hand, digital content generation supported by NLP technology, and on the other hand, digital content dissemination supported by NLP technology. However, in order to understand the digital content dissemination mode of NLP, it is necessary to clarify the changes in the ecosystem of digital content, the dissemination subject, and the dissemination path, so that corresponding innovative strategies can be proposed in a targeted manner.

3.2.1. Digital content ecosystem. The digital content ecosystem supported by NLP technology is a systematics theory covering multiple levels of knowledge subjects, knowledge objects and knowledge environments. Its core is to satisfy the digital content needs of the object, and to collaborate and coexist within the digital content system and in different digital content populations, which promotes the generation and flow of digital content through interaction with each other. The interaction of digital content between the virtual and the real makes the connotation and extension of the digital content ecosystem change significantly. To fully understand the digital content ecosystem of NLP, it is necessary to clarify the subject of digital content, the digital content environment, and the inherent operation mechanism, and then clarify the relationship between the components of the system, as well as their roles in the generation and dissemination of digital content.

1) Digital Content Subject. The main body of digital content is the core component of the digital content ecosystem, including not only the producers and users of digital content, but also the distributors of digital content. Due to the different digital content ecosystems in which they are located, digital content subjects can be divided into different types, such as the digital content subjects in the meta-universe digital content ecosystem, which include three types of real people, virtual people and digital people. Artificial Intelligence as a digital content subject fully demonstrates its characteristics of adapting to the digital content environment, and ChatGPT, as a representative of NLP technology, also embodies its ability to adapt to the environment in which it is located, and according to the process of formation and dissemination of digital content, the relationship between its digital content subjects is shown in Figure 5.

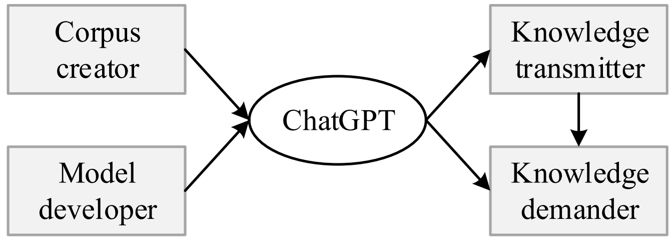


**Fig. 5.** Relationships between subjects of digital content

2) Digital content environment. The environment in which NLP technology is located is completely open, and can be divided into political, economic, social, legal, and cultural environments on a macro level, i.e., macro factors affecting the research and development and application of NLP technology. The digital content environment of NLP technology is shown in Figure 6, and can be divided into enterprises, technologies, platforms (mediums), and users at the application level. Enterprises form NLP technology products with specific functions through technological research and development, and users through a specific platform (media) to realize the use of NLP technology, access and generation of digital content.

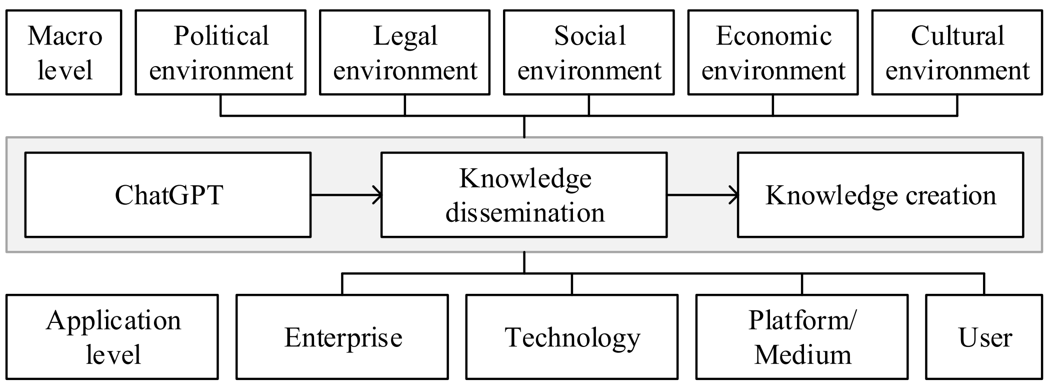


**Fig. 6.** Digital content environment of NLP technology

3) Operation Mechanism. Taking ChatGPT in NLP technology as an example, the operation mechanism of digital content ecosystem includes both the generation of digital content and the transfer, absorption and diffusion of digital content, which can be summarized into two aspects of digital content generation and dissemination, and its operation mechanism is shown in Figure 7. Digital content generation is the result of the joint action of all knowledge contributors, which are mainly corpus creators and model developers, on the basis of which the matching of supply and demand is realized through the big language model.

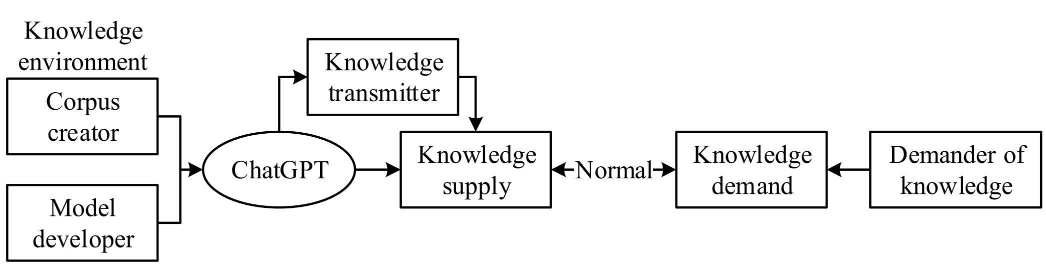


**Fig. 7.** The operation mechanism of digital content system based on ChatGPT

3.2.2. NLP-based digital content distribution paths. Digital content dissemination is the process of diffusion through carriers and accessed by demanders. The digital content dissemination based on NLP technology has a relatively single digital content dissemination subject, which is not only manifested in one-to-one digital content interaction, but also in a single digital content creation, i.e., the generation of digital content is only accomplished through NLP technology. These digital contents can either be the answers to the questions raised by the demanders or the recommendations of related materials, such as books and other learning materials that meet students' reading ability, interests and reading history. It can be seen that NLP technology simplifies the digital content dissemination path, and the digital content demander can obtain the required knowledge only through NLP, and at the same time, as the algorithms and semantic information continue to improve, the digital content it provides will be more accurate.

## 4. Analysis of the efficiency of digital content generation and dissemination

In order to verify the quality of digital content generated by NLP technology, and the effectiveness of improving the dissemination efficiency, practical experiments can be carried out to obtain the word vectors of digital content using the Skip-gram model, respectively. Digital content types include UGC, PGC, and BGC. Among them, UGC stands for user-generated content, PGC stands for professional-generated content, and BGC stands for brand-generated content.

### *4.1. Analysis of the efficiency of digital content dissemination*

4.1.1. Sample selection. Digital content distribution has become the preferred choice of many enterprises due to its low cost and stronger effect. Both large enterprises and small and medium-sized enterprises use valuable digital content on social platforms or in the network to attract and maintain customers. Currently, China's Internet penetration rate reaches 75.0%, and as long as one has access to the Internet, he or she has access to digital content. Therefore, in order to ensure the comprehensiveness and validity of the sample, there is no special restriction on the survey respondents. In addition, and to ensure the credibility of the sample, each IP address can only fill out the questionnaire once.

In this paper, the digital content types UGC, PGC, AND BGC are taken as independent variables, and the mediating variables are consumer perception, including perceived trustworthiness, perceived usefulness, and perceived risk, and the efficiency of digital content dissemination is the dependent variable. During the period from October 1 to October 7, 2024, 150 questionnaires were distributed online using social platforms such as WeChat and Weibo, and 124 valid questionnaires were collected. On this basis, SPSS26.0 was used to test the reliability and validity of the scales in the questionnaire, and the reliability and validity test results were passed, which verified the rationality and reliability of the questionnaire.

4.1.2. Manipulated variable tests. Widely used in research in the field of psychology, manipulation testing is a method of ensuring the successful manipulation of an expected variable through the measurement of the manipulated variable and other relevant factors. Manipulation tests are considered an important means of assessing the validity of an experimental manipulation and are an important symbol of experimental rigor. In this paper, the study is a contextualized experimental study, so it is necessary to conduct a manipulation test to ensure the validity of the manipulation of the type of digital content generated by NLP technology.

In this paper, the test of the validity of the manipulated variables is conducted through two interrogatives of the purpose of the subject of information dissemination, i.e., entertainment-financial interest, and hobby-occupation, i.e., the consumer's perception of the commercial purpose of the publisher's release of information. There are three information dissemination subjects here, and the results of one-way ANOVA and LSD post hoc tests are shown in Table 1. The results show that the commerciality difference between UGC and PGC is -1.3643, and the commerciality difference between PGC and BGC is -0.6885, i.e., the perceived commercial purpose of brand-generated content is greater than that of profession-generated content is greater than that of user-generated content. Therefore, in the contextual experiment in this paper, users can distinguish the digital content produced by different subjects based on the material, and the manipulation of digital content types is successful, and the set contextual material can be used in the research.

**Table 1.** Purpose-perceived differences in digital content types

| Dependent variable | (I) Types of digital content | (J) Types of digital content | Mean difference (I-J) | Standard error | SD | Sig. |
|---|---|---|---|---|---|---|
| Digital content dissemination efficiency | UGC | PGC | -1.3721* | 0.251 | 0.402 | 0.000 |
| | | BGC | 2.2034* | 0.243 | 0.389 | 0.000 |
| | PGC | UGC | 1.3721* | 0.247 | 0.371 | 0.000 |
| | | BGC | -0.7018* | 0.255 | 0.325 | 0.000 |
| | BGC | UGC | -2.2034* | 0.247 | 0.381 | 0.000 |
| | | PGC | 0.7018* | 0.256 | 0.283 | 0.000 |
| * The significance level of the mean difference was 0.05. | | | | | | |

4.1.3. Impact of NLP digital content generation on dissemination efficiency. The methods of difference test are chi-square test, independent samples t-test, paired samples t-test and ANOVA, in this paper, the independent variable is a category variable and the group is greater than 2, so we can only choose ANOVA for difference test.

1) Digital content generation type and user dissemination willingness. The ANOVA results of digital content type on user dissemination willingness are shown in Table 2, from which it can be seen that the significance level of the difference of digital content type on user dissemination willingness is 0.000, and the result is less than 0.05, which indicates that there is a significant difference in the impact of different types of digital content generated by NLP technology on user dissemination willingness. Subsequently, the variance chi-square of the factor of user communication willingness was tested, as shown in the table, the significance level of the variance chi-square test of user communication willingness is 0.125, which is greater than 0.05. This indicates that the variance of user communication willingness is chi-square in all the groups, and it can be tested post hoc using the LSD method.

**Table 2.** One-way analysis of variance between digital content type and communication intention

| Results of one-way analysis of variance | | | | | | |
|---|---|---|---|---|---|---|
| Test item | | Sum of squares | Degree of freedom | Mean square | F | Sig. |
| User communication intention | Intergroup | 51.325 | 2 | 25.372 | 12.025* | 0.000 |
| | Intra-class | 875.118 | 122 | 3.305 | 14.188* | 0.000 |
| | Total | 926.443 | 124 | 28.677 | 11.341* | 0.000 |
| Homogeneity test of variance | | | | | | |
| Test item | Levin statistics | Degrees of freedom 1 | | Degrees of freedom 2 | | Sig. |
| User communication intention | 2.834 | 2 | | 122 | | 0.125 |
| * The significance level of the mean difference was 0.05. | | | | | | |

The results of the LSD method test are shown in Table 3, which shows that the mean difference between the effects of UGC and PGC on users' willingness to disseminate is 0.55085, $p=0.001<0.05$, which indicates that UGC is more capable of increasing users' willingness to disseminate digital content than digital content generated by PGC. The mean difference between the effects of PGC and BGC on users' willingness to disseminate is 0.29442, $p=0.032<0.05$, indicating that PGC is more effective than BGC in improving the dissemination of digital content.

**Table 3.** Multiple comparative analysis of users' willingness to disseminate digital content

| Dependent variable | (I) Types of digital content | (J) Types of digital content | Mean difference (I-J) | Standard error | SD | Sig. |
|---|---|---|---|---|---|---|
| Digital content dissemination efficiency | UGC | PGC | 0.55085* | 0.152 | 0.352 | 0.001 |
| | | BGC | 0.81432* | 0.157 | 0.347 | 0.000 |
| | PGC | UGC | -0.55085* | 0.150 | 0.365 | 0.001 |
| | | BGC | 0.29442* | 0.155 | 0.334 | 0.032 |
| | BGC | UGC | -0.81432* | 0.161 | 0.328 | 0.000 |
| | | PGC | -0.29442* | 0.156 | 0.277 | 0.032 |
| * The significance level of the mean difference was 0.05. | | | | | | |

2) Digital content generation type and user perception. The results of the variance of digital content generation type on user perception are shown in Table 4, from which it can be seen that the significance level of the variance of perceived credibility, perceived usefulness, and perceived risk is 0.000, which is less than 0.05, indicating that there is a significant difference in the impact of different types of digital content on perceived credibility, perceived usefulness, and perceived risk.

**Table 4.** One-way ANOVA results for digital content type on user perception

| Test item | | Sum of squares | Degree of freedom | Mean square | F | Sig. |
|---|---|---|---|---|---|---|
| Perceived credibility | Intergroup | 36.732 | 2 | 19.725 | 10.057 | 0.000 |
| | Intra-class | 761.341 | 122 | 1.687 | | |
| | Total | 798.073 | 124 | | | |
| Perceived usefulness | Intergroup | 50.811 | 2 | 27.068 | 15.834 | 0.000 |
| | Intra-class | 774.351 | 122 | 1.934 | | |
| | Total | 825.162 | 124 | | | |
| Perceived risk | Intergroup | 46.501 | 2 | 23.981 | 14.856 | 0.000 |
| | Intra-class | 737.072 | 122 | 1.672 | | |
| | Total | 783.573 | 124 | | | |

### *4.2. Evaluation and Analysis of Generative Effectiveness*

4.2.1. Generation steps. The steps of digital content generation based on NLP technology include data acquisition and preprocessing, key information extraction, content generation and content touch-up:

1) Data Collection and Preprocessing. In the process of digital content generation, it is first necessary to collect a large amount of text data related to a specific topic and preprocess these data. In this regard, data collection is the collection of textual data related to the topic from multiple sources. Text cleaning is the removal of Hypertext Markup Language (HTML) tags, special characters, etc., to retain plain text content. Segmentation is to split the text into words or phrases. Deactivation Removal is to remove common deactivated words to reduce the impact of noise on subsequent analysis. Lexical annotation is to label the lexical properties of each word for subsequent key information extraction.

2) Key Information Extraction. Key information extraction is a key step in generating digital content, which identifies important concepts and key words from the text.

3) Digital Content Generation. Digital content generation is the process of automatically generating digital content based on the extracted key information, which can be realized by using NLP generator, combined with templates and grammar rules. Among them, template design is to design multiple content templates to provide a framework for subsequent filling. Key information filling is to fill the templates using important concepts and key words obtained from key information extraction to generate content paragraphs. Grammar rule application is to combine grammar rules to ensure that the generated content is smooth and accurate.

4) Digital content touch-up. The generated digital content may have grammatical errors and incoherence, and content touch-ups are needed to improve the quality of the content.

4.2.2. Quality analysis of generated content. For the assessment of the effectiveness of digital content generation quality, you can choose a content quality score. The assessment is performed using the open source tool Natural Language Data Kit (NLTK), which evaluates the generation quality scores of individual sentences or multiple sentences (e.g., entire documents and paragraphs) by means of functions that come with the tool.

In order to assess the quality of digital content generated by the previously proposed NLP technique, the distribution of the quality scores of content generated using the NLTK is shown in Figure 8. It can be seen that a high percentage of content with high quality scores is extracted, with a focus on the percentage of content with 0.5-1.0 quality scores reaching more than 60%, indicating that the NLP technique achieves high-quality digital content generation.

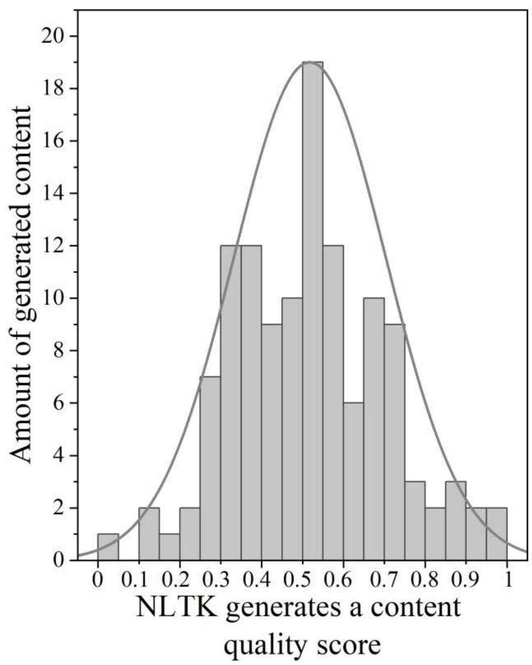


**Fig. 8.** NLTK generates a content quality score distribution

4.2.3. Changes in coverage. The coverage of digital content generated by the NLP technique is measured using keyword coverage, and the change in coverage over time is shown in Figure 9. In the figure, 0 to 10 represents the progress of the digital content generated by NLP, and a larger number represents more learning progress. As the generated content unfolds, the keyword coverage of the generated content increases from 0.25 to 1.06, and the coverage increases gradually. It can be concluded that after the steps of data acquisition and preprocessing, key information extraction, and content generation content embellishment, the content of NLP technology can be generated quickly and efficiently, which provides a new solution for digital content generation and automation applications.

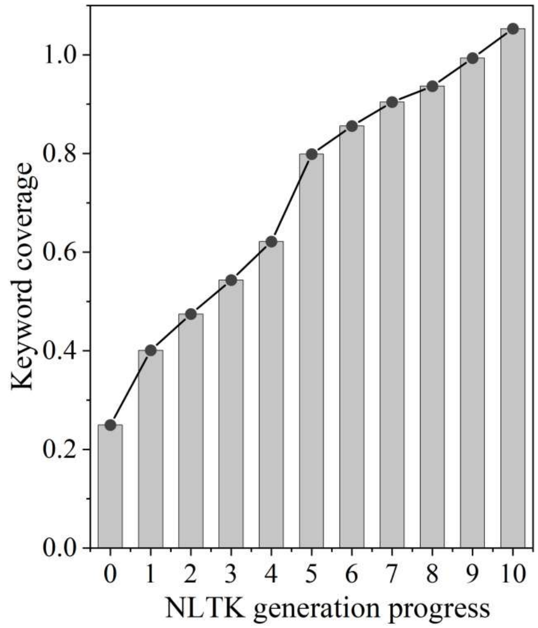


**Fig. 9.** The trend of coverage over time

4.2.4. Lift rate of content dissemination. By posting the digital content generated under the synergy of NLP technology on 10 social platforms (S1-S10), the dissemination rate (click-through rate) of the content generated using NLP is compared with that using the traditional model, respectively. The results of the enhancement rate of digital content dissemination are shown in Table 5. From the table, it can be seen that the enhancement rate of content dissemination after digital content generation using NLP technology increases by an average of 20.056% across the 10 platforms as compared to the dissemination effect using the traditional method. This indicates that NLP technology digital content generation is effective, with high matching and dissemination efficiency.

**Table 5.** Content dissemination increase rate

| Social platform | Traditional digital content creation | NLP digital content generation mode | Difference value | Propagation lift rate/% |
|---|---|---|---|---|
| S1 | 21825 | 22726 | 901 | 4.128 |
| S2 | 32971 | 46210 | 13239 | 40.153 |
| S3 | 10462 | 15516 | 5054 | 48.308 |
| S4 | 35578 | 41271 | 5693 | 16.001 |
| S5 | 26280 | 33326 | 7046 | 26.811 |
| S6 | 27560 | 32012 | 4452 | 16.154 |
| S7 | 1088 | 1377 | 289 | 26.563 |
| S8 | 28260 | 33524 | 5264 | 18.627 |
| S9 | 15518 | 16376 | 858 | 5.529 |
| S10 | 19394 | 20510 | 1116 | 5.754 |
| Mean | 21894 | 26285 | 4391 | 20.056 |

## 5. Conclusion

In the era of digital technology, academics have explored and researched the dissemination mode, creation and generation methods of digital content. Based on the results of previous research, this paper explores the influence of natural language processing (NLP) technology on digital content generation and dissemination under the perspectives of communication theory, embodiment theory and media environment theory, and makes some innovations in the following aspects:

First, on the basis of clarifying the logic of embodied practice of technology-body-communication, the historical evolution, digital reality and embodied practice of NLP technology's influence on digital content generation and dissemination are discussed in depth. The introduction of interdisciplinary perspectives on the theoretical construction and practical elaboration of digital content dissemination provides doctrinal support for further qualitative research and empirical testing of this study.

Secondly, this study carries out a systematic research on digital content communication based on NLP technology embodiment, highlights the important roles of technological embodiment practice and embodied perception of the communication subject, identifies the relationship between the digital embodied environment, embodied perception, and the communication effect, and constructs and verifies the mechanism of NLP technology's influence on digital content communication.

Thirdly, the relationship between embodied environment, embodied perception, digital content generation and communication effect is taken as an orientation, and the dilemmas arising from the integration of NLP technology into digital content generation and communication practice are analyzed, and the optimization path of communication is proposed from the perspectives of concept, environment, subject, and governance, so as to provide strategic references to the digital content's digitalization, high-quality, and connotative development.

## References

[1] Sančanin, B., & Penjišević, A. (2022). Use of artificial intelligence for the generation of media content. Social Informatics Journal, 1(1), 1-7.

[2] Zubair, F., & Jafir, A. (2021). Natural Language Generation: Recent Advances and Applications. International Journal of Advanced Engineering Technologies and Innovations, 1(1), 50-74.

[3] Kumar, N., Kumar, M., & Singh, M. (2016). Automated ontology generation from a plain text using statistical and NLP techniques. International Journal of System Assurance Engineering and Management, 7, 282-293.

[4] Loukili, S., Fennan, A., & Elaachak, L. (2023, May). Applications of Text Generation in Digital Marketing: a review. In Proceedings of the 6th International Conference on Networking, Intelligent Systems & Security (pp. 1-8).

[5] Kulkarni, C. (2022). Ethical implications of large language models in content generation. Journal of Artificial Intelligence, Machine Learning & Data Science, 1(1), 62-67.

[6] Kim, T. W. (2023). Application of artificial intelligence chatbots, including ChatGPT, in education, scholarly work, programming, and content generation and its prospects: a narrative review. Journal of educational evaluation for health professions, 20.

[7] Jadhav, D., Agrawal, S., Jagdale, S., Salunkhe, P., & Salunkhe, R. (2024, October). AI-Driven Text-to-Multimedia Content Generation: Enhancing Modern Content Creation. In 2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) (pp. 1610-1615). IEEE.

[8] Dessì, D., Osborne, F., Recupero, D. R., Buscaldi, D., & Motta, E. (2021). Generating knowledge graphs by employing natural language processing and machine learning techniques within the scholarly domain. Future Generation Computer Systems, 116, 253-264.

[9] Bayer, M., Kaufhold, M. A., Buchhold, B., Keller, M., Dallmeyer, J., & Reuter, C. (2023). Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers. International journal of machine learning and cybernetics, 14(1), 135-150.

[10] Le, D. V. T., Bigo, L., Herremans, D., & Keller, M. (2024). Natural language processing methods for symbolic music generation and information retrieval: A survey. ACM Computing Surveys.

[11] Wang, H. C., Chiang, Y. H., & Chen, I. F. (2024). A method for generating course test questions based on

natural language processing and deep learning. Education and Information Technologies, 29(7), 8843-8865.

[12] Upadhyay, P., Agarwal, R., Dhiman, S., Sarkar, A., & Chaturvedi, S. (2024). A comprehensive survey on answer generation methods using NLP. Natural Language Processing Journal, 8, 100088.

[13] Fragoso-Diaz, O. G., López-Caballero, V., Rojas-Pérez, J. C., Santaolaya-Salgado, R., & González-Serna, J. G. (2021). On the generation of e-learning resources using business process, natural language processing, and web services. IT Professional, 23(2), 40-44.

[14] Dharaniya, R., Indumathi, J., & Kaliraj, V. (2023). A design of movie script generation based on natural language processing by optimized ensemble deep learning with heuristic algorithm. Data & Knowledge Engineering, 146, 102150.

[15] Dastan Hussen Maulud,Siddeeq Y. Ameen,Naaman Omar,Shakir Fattah Kak,Zryan Najat Rashid,Hajar Maseeh Yasin... & Dindar Mikaeel Ahmed. (2021). Review on Natural Language Processing Based on Different Techniques. Asian Journal of Research in Computer Science1-17.

[16] Abdulaziz Altamimi. (2024). Harnessing the integration of chat GPT in higher education: the evaluation of stakeholders sentiments using NLP techniques. Discover Sustainability(1),509-509.

[17] Shiming Ma. (2024). Research on Digital Media Content Production and Dissemination Mode Based on Artificial Intelligence. Applied Mathematics and Nonlinear Sciences(1).