J. COMBIN. MATH. COMBIN. COMPUT. 127a (2025) 487–502

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

Combinatorial Press

Improving the LDA model applied to news form innovation in cultural communication

Yiyu Chen[1,✉], Lin Ma[1]

[1] *College of Humanities and Social Sciences, Beijing Institute of Petrochemical Technology, Beijing, 102617, China*

ABSTRACT

The development of media technology profoundly affects the presentation mode, dissemination rate and scale of news information, which in turn reshapes the business chain and business landscape of the entire news media industry. Based on the analysis of the shortcomings of the LDA model, this paper proposes an improved LDA model with binomial distribution, and applies it to the analysis of the evolution of news topics. The model introduces binomial distribution to enhance the discriminative ability of lexical items, and parallelises it to improve the classification effect of news topics. In order to effectively obtain the relevant features of cultural communication in news documents, this paper introduces BERT to obtain word embedding and word vector matrix, and then realises the generation of theme word structure and theme words. The performance of the improved LDA model is verified through the THUCNews dataset, and the news topic morphology is visualised and analysed with the example data, and its morphological evolution, as well as the degree of contribution to cultural communication, is studied. The theme consistency score of the improved LDA model is -13.39 when the word generation probability is 1, which is 19.14% higher than that of the traditional LDA model. The intensity of the 'cultural policy' news format theme increases 14.44 times from 2010 to 2023, and the mean value of the 'cultural governance' news format theme's contribution to cultural dissemination reaches 0.091. Based on the innovation and evolution of news forms, we can empower more communication channels for culture and spirit, so as to enhance people's cultural self-confidence and national cultural soft power.

Keywords: improved LDA model; binomial distribution; word vector matrix; news pattern theme; cultural dissemination

1. Introduction

With the rapid development of social economy, our society has begun to enter a kind of all-media era, and this has become an important development trend. It is not only manifested in the audience's increasingly strong demand for information, but also manifested in the more diversified means of information

✉ Corresponding author.
E-mail address: IyoBang@163.com (Y. Chen).

Received 10 August 2024; Revised 12 November 2024; Accepted 30 January 2025; Published Online 15 April 2025.
DOI: 10.61091/jcmcc127a-029

dissemination, therefore, the news field of cultural communication in order to obtain a certain degree of development, we must carry out a variety of innovations, and make specific improvements from the form [1-4]. However, due to the traditional development model is more influential, in the current news form of innovation has encountered more problems, and the lack of effective ways of innovation [5-7]. The current news form has certain limitations on its development, which does not match the needs of the all-media era, so morphological innovation by virtue of innate advantages is a task that must be accomplished by the current news communication, which is of great help to enhance the competitiveness of the news communication field [8-11].

In the all-media era, the innovation of news form is manifested in several aspects. In content production, editorial centrism is migrating to audience centrism, and all kinds of convergent technologies that keep emerging must be summoned on the basis of being used for people, serving audiences and satisfying their needs [12-15]. The audience's emotional point of view becomes a key consideration, through the care and expression of individual destiny in the big era, to achieve the cognitive conversion and transition from micro-destiny to macro-issues. In terms of interaction design, the point of innovation is shifted from result to process. Through the restoration and reconstruction of the scene, the process in which the audience can deeply participate is created [16-19]. This process, without violating the authenticity of the news, has both the anamorphic nature of the news scene and the ideological nature that transcends the news event [20-21].

The development of technology drives the change of news production and dissemination mode, realising the change of news value, and the new form of news dissemination represented by ‘aggregation’ and ‘personalization’ is gradually becoming the mainstream trend. Aiming at the problems of information loss and topic confusion in the existing LDA model, the article introduces binomial distribution to improve the LDA model. The binomial distribution is used to enhance the judgement ability of the model on word items, so as to obtain the corresponding topics more clearly. When constructing the news topic analysis model, the LDA model is used to obtain the feature vectors of the news pattern topics, and the BERT module is used to generate the word vector matrix. The THUCNews dataset is chosen as the research data source, and the performance of the improved LDA model is analysed through the theme consistency and average pointwise mutual information, and also combined with the example data to analyse the news pattern theme evolution and the degree of contribution to cultural communication.

2. Modelling

With the development of technology and the popularisation of the Internet, the era of integrated media has become the mainstream trend in the field of news dissemination. Under the background of this era, the cultural communication path of news media also presents brand-new characteristics and trends. News media in the era of integrated media is no longer limited to traditional paper media, radio and television, but carries out cultural communication through various channels such as the Internet, social media, mobile applications, etc., which expands the breadth and depth of communication and also realises the innovative development of news forms. This chapter constructs a news theme analysis model applied to cultural communication based on the improved LDA model, aiming to explore the trend of news form theme evolution under the perspective of cultural communication.

2.1. *Improvement of the LDA model*

2.1.1. LDA model. Latent Delicacy Allocation (LDA) is an unsupervised learning topic model, which models the text through the corpus context, following the three-layer Bayesian model rule of ‘document-topic-feature word’, reflecting the potential topics through the probability distribution of words, and discovering hot topics [22].

Relying on a variety of different types of documents, which constitute a corpus, namely:

$$W=\left(w_1,w_2,\ldots,w_M\right) \tag{1}$$

And a document can be represented by an ordered sequence of words:

$$d_n=w_m=\left(w_{m1},w_{m2},\ldots,w_{mm_1}\right) \tag{2}$$

The process of generating word sequences is carried out by randomly assigning a topic to each word in each document and then generating words by selecting topics based on a probability distribution. By repeating the process of generating words, the distribution of topics of documents and the distribution of words under each topic can be inferred.

Setting the number of topics as K, $\vec{\alpha}\to\overrightarrow{\theta_m}\to z_{m,n}$ this process is expressed in the generation of the m rd document, Mr into the topic distribution, the formula is:

$$\overrightarrow{\theta_m}=Dir\left(\overrightarrow{\theta_m}|\vec{\alpha}\right) \tag{3}$$

And then by the n nd word of the subject $Z_{m,n}$ in the $\overrightarrow{\theta_m}$ -document, denoted by:

$$z_{m,n}=Mult\left(z_{m,n}\left|\overrightarrow{\theta_m}\right.\right) \tag{4}$$

where $\vec{\alpha}\to\overrightarrow{\theta_m}$ is a Dirichlet distribution and $\overrightarrow{\theta_m}\to z_{m,n}$ is a Multinomial distribution, so the union of these two steps is a Dirichlet-Multinomial conjugate structure obtained:

$$p\left(\vec{z}\,|\Theta\right)=\prod_{m-1}^{M}\prod_{k=1}^{K}\vartheta_{m,k}^{r(k)} \tag{5}$$

$\vec{\beta}\to\vec{\varphi}_k\to w_{m,n}\left|=z_{m,n}\right.$ denotes the word distribution numbered $k=z_{m,n}$ subject is selected from the word distribution $\vec{\varphi}_k$ of the k nd subject and sampled to generate vocabulary $w_{m,n}$.

Among them:

$$\vec{\varphi}_k\sim Dir\left(\vec{\varphi}_k\left|\vec{\beta}\right.\right) \tag{6}$$

$$w_{m,n}\sim Mult\left(w_{m,n}\left|\vec{\varphi}_k\right.,z_{m,n}\right) \tag{7}$$

The same Dirichlet-Multinomial conjugate structure is obtained after the union of these two steps:

$$p\left(\vec{w}\,|\vec{z},\phi\right)=\prod_{k=1}^{K}\prod_{t=1}^{V}\varphi_{k,t}^{n^{(i)}} \tag{8}$$

From this Dirichlet-Multinomial covariance structure it is possible to obtain the formula for generating the probability of a word appearing in a particular document in the LDA topic model as:

$$p\left(\vec{z},\vec{w}\left|\vec{\alpha},\vec{\beta}\right.\right)=p\left(\vec{w}\left|\vec{z},\vec{\beta}\right.\right)p\left(\vec{z}\,|\vec{\alpha}\right) \tag{9}$$

Among them:

$$p\left(\vec{w}\left|\vec{z},\vec{\beta}\right.\right)=\int p\left(\vec{w}\,|\vec{z},\phi\right)p\left(\phi\left|\vec{\beta}\right.\right)d\phi=\prod_{k=1}^{K}\frac{\Delta\left(\overrightarrow{n_k}+\vec{\beta}\right)}{\Delta\left(\vec{\beta}\right)} \tag{10}$$

$$p\left(\vec{z}\,|\vec{\alpha}\right)=\int p\left(\vec{z}\,|\Theta\right)p\left(\Theta|\vec{z}\right)d\Theta=\prod_{m-1}^{M}\frac{\Delta\left(\overrightarrow{n_m}+\vec{\alpha}\right)}{\Delta\left(\vec{\alpha}\right)} \tag{11}$$

Bringing the above two equations into equation (9) gives:

$$p\left(\vec{z},\vec{w}\middle|\vec{\alpha},\vec{\beta}\right)=\prod_{k=1}^{K}\frac{\Delta\left(\overrightarrow{n_k}+\vec{\beta}\right)}{\Delta\left(\vec{\beta}\right)}\prod_{m-1}^{M}\frac{\Delta\left(\overrightarrow{n_m}+\vec{\alpha}\right)}{\Delta\left(\vec{\alpha}\right)} \tag{12}$$

From the above, it can be learnt that $\vec{\alpha}$ is the parameter used to control $\overrightarrow{\theta_m}$, which is independent of the obtained $z_{m,n}$, and $\vec{\beta}$ is the parameter used to control $\vec{\varphi}_k$, which is independent of the obtained $w_{m,n}$. Therefore, the a posteriori probabilities are calculated for parameters $\overrightarrow{\theta_m}$ and $\vec{\varphi}_k$, respectively:

$$p\left(\overrightarrow{\vartheta_m}\middle|\overrightarrow{z_m},\vec{\alpha}\right)=Dir\left(\overrightarrow{\theta_m}\middle|\overrightarrow{n_m}+\vec{\alpha}\right) \tag{13}$$

$$p\left(\overrightarrow{\varphi_k}\middle|\vec{z},\vec{w},\vec{\beta}\right)=Dir\left(\overrightarrow{\varphi_k}\middle|\overrightarrow{n_k}+\vec{\beta}\right) \tag{14}$$

Thus, $\overrightarrow{\theta_m}\sim Dir\left(\overrightarrow{\theta_m}\middle|\overrightarrow{n_m}+\vec{\alpha}\right)$, $\overrightarrow{\varphi_k}\sim Dir\left(\overrightarrow{\varphi_k}\middle|\overrightarrow{n_k}+\vec{\beta}\right)$. And can be obtained from the expectation of the Dirichlet distribution:

$$\vartheta_{m,k}=\frac{n_m^{(k)}+\alpha_k}{\sum_k\left(n_m^{(k)}+\alpha_k\right)} \tag{15}$$

$$\varphi_{k,t}=\frac{n_k^{(t)}+\beta_t}{\sum_t\left(n_k^{(t)}+\beta_t\right)} \tag{16}$$

Establishing an LDA model requires setting the number of topics, and the number of topics will be directly related to the topic quality and clustering effect of the LDA model, if too many topics are set it may lead to problems such as overfitting of the model, overlapping between topics, unclear extraction of topics and increased model complexity. On the contrary, setting too few topics may result in the LDA model not being able to adequately capture the rich topic diversity in the text data, resulting in information loss, topic confusion and poor model performance.

2.1.2. LDA model improvement. The base LDA model does not take into account the degree of differentiation of the word items, if a word item occurs with high frequency in most of the topics, then this word item cannot be used as a feature word to represent the topic. Text is sparse and direct removal of lexical items that are not discriminative to the topic will have some impact on the model, for this reason, a new model is proposed to introduce binomial distribution into the base LDA model to increase the discriminative ability of the lexical items and to parallelise it [23].

Words are sampled by treating the affiliation of the same word to different topics as independent of each other, and the discriminative power of the lexical items is denoted by γ. If topic z contains lexical item w, then $\gamma_{w,z}=1$, and if it does not contain w, then $\gamma_{w,z}=0$. It can be shown that γ is a parameter obeying a binomial distribution, i.e.:

$$\gamma\sim B\left(K,\lambda_w\right) \tag{17}$$

where λ_α represents the distribution of word w over all topics. In this paper, we assume that $\lambda_w\sim Beta\left(a,b\right)$, because α and β parameters have been used, so in this paper, a and b parameters are used to represent the parameters in the beta distribution. Namely:

$$p\left(\lambda_w\middle|a,b\right)=\frac{1}{B(a,b)}\lambda_w^{a-1}\left(1-\lambda_w\right)^{b-1} \tag{18}$$

Assuming that lexical item w is counted as $n_w^{(1)}$ in K topics and not occurring $n_w^{(0)}=K-n_w^{(1)}$, the conditional probability corresponding to the occurrence of this lexical item w in the next topic, which is noted as $p\left(\gamma=1\middle|w\right)$, can be derived from the Bayesian formula:

$$p\left(\lambda_w \middle| Y,a,b\right) = Beta\left(\lambda_w \middle| a+\sum_{t=1}^{K} Y_{w,z}=1, b+\sum_{t=1}^{K} Y_{w,z}=0\right) \tag{19}$$

Of these, $Y=\left\{Y_{w,1},Y_{w,2},\cdots,Y_{w,K}\right\}$.

$$\lambda_w = p\left(\gamma_{w,z}=1\middle|\gamma\right)=\frac{\sum_{z=1}^{K}\gamma_{w,z}=1+a}{K+a+b} \tag{20}$$

If a lexical item w occurs uniformly across topics, a larger value of λ_w does not discriminate much between topics; on the contrary, if a lexical item w concentrates on only one or a small number of topics, the lexical item's value of λ_w is smaller. Therefore, $\omega=1/\lambda_w$ is used to represent the discriminative power of the lexical items, and greater weights are assigned to lexical items with high discriminative power to reduce the influence of words with low discriminative power on the clustering. Regularisation of ω using $\omega\in[0,1]$ is expressed as:

$$\omega=\frac{\left(K-n_w^{(1)}\right)(a+1)}{(K-1)\left(a+n_w^{(1)}\right)} \tag{21}$$

Gibbs sampling can be used to solve high-dimensional probabilistic models by iterating, each iteration sampling only the value of the current dimension and fixing the values of the other dimensions. The algorithm keeps iterating until it converges and finally outputs the parameters to be estimated. In the LDA algorithm that is the output φ and θ.

After a series of derivation transformations, the updated Gibbs sampling of the new model can be obtained as Eq:

$$p\left(z_i=k\middle|\overrightarrow{z_{\neg i}},\vec{w}\right)\propto\frac{\left(K-n_{w_i}^{(1)}\right)(a+1)}{(K-1)\left(a+n_w^{(1)}\right)}\cdot\frac{n_{d,\neg i}^{z}+\alpha}{\sum_{z=1}^{K}n_{d,\neg i}^{z}+K\alpha}\cdot\frac{n_{w,i,\neg i}+\beta_t}{\sum_{i=1}^{V}n_{w,z,\neg i}+\beta_t} \tag{22}$$

At the initial moment, a random assignment of topic z_i is performed for each word item of the document. Then the number of word items w under z_i is counted, as well as the number of word items under document i containing words in topic z_i. Each round estimates the probability that each word is assigned to each topic according to Eq. $p\left(z_i=k\middle|\overrightarrow{z_{\neg i}},\vec{w}\right)$, and then samples new topics using this probability distribution; the same method is also used for each word to update the topic of the word. If the distribution of the word items in each topic and the update of the distribution of the training documents under each topic converge, the model parameters are output.

The expressions for the θ-parameter and φ-parameter are obtained after the Gibbs sampling process as:

$$\theta_{d,z}=\frac{n_{z,d}+\alpha}{\sum_{Z=1}^{K}n_{z,d}+K\alpha} \tag{23}$$

$$\varphi_{i,w,w}=\frac{n_{w,z}+\beta}{\sum_{i=1}^{V}n_{w,z}+V\beta}\cdot\frac{\left(K-n_m^{(1)}\right)(a+1)}{(K-1)\left(a+n_w^{(i)}\right)} \tag{24}$$

2.2. Model for analysing news topics

2.2.1. Obtaining Topic Feature Vector. In this paper, the improved LDA model is applied to the construction of the news topic analysis model for cultural communication, and the news topic feature vector is obtained by improving the LDA model [24]. For the news topic applied to cultural communication can be understood as the words and the weights of the words that are semantically similar to the meaning of the topic, which can also be regarded as the vector representation composed of the conditional probabilities of all the words in the topic. Then the theme vector representation is:

$$Z = \left(\left(w_1, P\left(w_1 | z\right)\right), \left(w_2, P\left(w_2 | z\right)\right), \cdots\cdots \left(w_n, P\left(w_n | z\right)\right)\right) \tag{25}$$

In the topic vector formula w_n represents the vocabulary, $P\left(w_n | z\right)$ is the probability of the vocabulary appearing on the topic z. The greater the conditional probability of its vocabulary, the closer the vocabulary is to the topic z, which also indicates that the vocabulary is more expressive of the topic's meaning, and vice versa. Combining the document topic vector formula and the model sampling principle, the total LDA topic model formula is derived as:

$$P\left(\theta, Z, W | \alpha, \beta\right) = P\left(\theta | \alpha\right) \prod_{n=1}^{N} P\left(Z_i | \theta\right) P\left(w_n | z_n, \beta\right) \tag{26}$$

Where θ is the probability of each topic appearing in the document. N represents the number of words in the document, d is the name of the document, w_n represents the nth word in the document d. z_n represents the probability that the word selects the main present z, $P\left(z | \theta\right)$ is the probability distribution of the topic z, and $P\left(w | z\right)$ is the probability distribution of the vocabulary belonging to the topic. By controlling the parameters α and β in the above equation, the model is trained and the cosine distance between the words in each topic and the document, i.e., the topic feature vector of the corpus text, is calculated.

2.2.2. Theme word generation process. Suppose there are d documents in corpus D with V words, and $w_{dn} \in \{1, \ldots, V\}$ represents the nth word in the dth document. ρ is the word embedding matrix of $L \times V$, column ρ_v is the word vector representation of w_v, which is obtained by the BERT model, $\rho_v \in R^L$. K is the number of topics, α is the topic matrix of $L \times K$, column α_k is the vector of the kth topic, $\alpha_k \in R^t$, θ are the parameters of the document-topic distribution, β is the parameter of the topic-word distribution, and z is the topic, and the topic words are generated as follows:

1) Generate the document-topic distribution θ In order to facilitate the use of variational inference for solving the document-topic distribution, θ obeys a logistic-normal distribution, which is a normalised standard Gaussian distribution. That is:

$$\theta_d \sim LN\left(0, I\right) \tag{27}$$

$$\delta_d \sim N\left(0, I\right); \theta_d = soft\max\left(\delta_d\right) \tag{28}$$

2) for each word in document d. The model assigns the topic z_{dn} of the nrd word in the dnd document to each observed word sample to get the topic of the nth word. Then:

$$z_{dn} \sim Cat\left(\theta_d\right) \tag{29}$$

The model obtains word embeddings by BERT and gets the word vector matrix ρ.

The model obtains topic vector matrix α in the same semantic space of word vector matrix ρ. After obtaining topic z_{dn}, word vector matrix ρ is multiplied by topic vector matrix α, and the matrix product is taken as topic-word distribution, and then the vocabulary is generated to obtain the nth word. Then:

$$w_{dn} \sim soft\max\left(\rho^T \alpha_{z_{dn}}\right) \tag{30}$$

Relying on the improved LDA model to establish a news theme analysis model applied to cultural communication, the purpose is to understand the evolution trend of news themes of different cultural communication, analyse the innovative development trend of news forms, and provide support for further optimising the development of news forms of cultural communication.

3. Experimental design

News communication and cultural communication are conceptually different, but in practice they are closely linked and integrated. The deep integration of media development requires the news media to effectively enhance cultural self-awareness, adhere to objectivity and impartiality, to ensure the authenticity of cultural communication. Adhere to the people-oriented, enhance the momentum of cultural communication, adhere to reform and innovation, enhance the aesthetic nature of cultural communication. The news media will be able to promote the high-quality development of advanced culture in the new era by transforming people and communicating with the world, and will show the mission and responsibility of the news media in the process of constructing cultural self-confidence. Based on this, this chapter establishes news communication related data sets and sets corresponding parameters and evaluation indexes for the improved LDA model, which provides help for the analysis of news topics applied to cultural communication.

3.1. News dataset and pre-processing

3.1.1. Description of the data set. In this paper, we use the news dataset THUCNews, which is generated by screening and filtering the historical data of Sina News RS subscription channel from 2010 to 2023, and contains a variety of different cultural categories and 700,000 news documents, all in UTF-8 plain text format. 10,000 news texts are randomly selected for each category, and the news title, news body and category information are extracted from the dataset. The training, validation and test sets are divided in the ratio of 8:1:1, where the training set for a single category consists of 8000 news texts for updating the model parameters, the validation set consists of 1000 news texts for selecting and tuning the parameters of the model, and the test set consists of 1000 news texts for evaluating the generalisation ability of the final selected model.

3.1.2. Data pre-processing. The preprocessing of the selected THUCNews news dataset can be divided into the following four steps:
1) Read the news dataset, the first two characters are the category label of the news, followed by the specific content of the news, and each news in the UTF-8 file occupies one line.
2) Read the predefined deactivated thesaurus, and perform the slicing and return list operation.
3) Perform the segmentation operation on the news text to convert the complete sentences in the news into words. The word segmentation step is achieved by the Jieba library in Python, which is an open source library dedicated to word segmentation for Chinese text processing, through which each text sequence can be sliced into a word or phrase. For news data, good word separation can accurately separate out the keywords in the news to help the computer better understand. Since this paper is to analyse the news text for themes, the precise mode in Jieba is more suitable for text preprocessing before text classification, so this paper chooses the precise mode to separate the text.
4) Traverse the news data with good words, compare each word in the news with the words in the deactivated word list one by one, and delete the word in the news data set with good words when it is the same as the word in the deactivated word list, and complete the operation of removing deactivated words after traversing and screening each word. The main purpose of removing deactivated words is to filter out some words, phrases, letters and numbers that have little or no effect on the meaning of the text itself. The deactivation word list is manually set, but there is no deactivation word list that can accurately fit all kinds of text datasets, because the inaccuracy of the deactivation word list will lead to many noisy words not being deleted or some key words being wrongly deleted, therefore, the inaccuracy of the deactivation word list in the text categorisation is also a source of noise. In this paper, we choose Baidu deactivation word list,

which is commonly used in text data mining, to deactivate the text.

3.2. *Parameter setting and evaluation indicators*

3.2.1. Parameter setting. During model training, the number of iterations of the model is set to 1000 in this paper.The other preset hyperparameters of this model are mainly $\alpha,\beta,\gamma,\delta,\varepsilon,\zeta,\eta,\omega$. Parameter δ,ε, which controls the contextual information, and parameter α,γ, which controls the contextual information in the improved LDA model, are mainly determined by the two indicator variables, namely, the number of feature words and the complexity. The so-called complexity is a parameter to assess the performance of the probabilistic model, which can be expressed as:

$$P(p) = 2^{H(p)} = 2^{-\sum_{x} p(x)\log_2^{p(x)}} \tag{31}$$

The smaller the complexity value, the easier it is to produce good results. Also, the smaller the number of feature words, the better the control of the model. So setting a combination value to make the two influencing factors as small as possible can make the value of the parameter more desirable. In the actual experiment, it is found that α can be set to 0.05, γ can be set to 0.001, δ can be set to 0.25, and ε can be set to 0.65 which is more reasonable. In the model training, Adam is chosen as the optimiser, the learning rate is set to 0.001, and in order to prevent overfitting, the random loss rate is added after the linear change in the experiment, and it is set to 0.05.

3.2.2. Evaluation indicators. In order to effectively carry out the analysis of the evolutionary trend of news theme patterns, this paper introduces the following three evaluation indicators:

1) Theme consistency. The number of topics is one of the most important parameters when constructing LDA models, and the number of topics has a direct impact on the potential topic identification effect, this paper chooses topic consistency to measure the semantic similarity between high scoring words in a topic to score a single topic [25]. Words with higher scores indicate higher topic coherence and better interpretability and topic representation. The formula for thematic coherence is as follows:

$$Cv = \mu\left(\left\{s_{\cos}\left(\underset{u}{\longrightarrow},\underset{w}{\longrightarrow}\right)\middle|\underset{u}{\longrightarrow},\underset{w}{\longrightarrow}\in W\right\}\right) \tag{32}$$

According to the above equation, it can be seen that the cosine similarity of a particular topic word is first calculated and then its arithmetic mean is calculated. According to the result of theme consistency, the optimal parameters are set for LDA theme recognition, and after visualising the results, the theme probability distribution graph is obtained.

2) Average Point Mutual Information (APMI). In order to make a more accurate assessment of news topics, this paper uses the average point mutual information between any two words contained in a topic to measure on the basis of topic consistency.APMI can quantify the probability of different words co-occurring in the representation of the topic, and if the greater the probability of words co-occurring in a topic, it means that the consistency of the topic is stronger and the topic is more interpretable. Then:

$$TC = 1\Big/ k\sum_{k-1}^{k} 1\Big/ 45\sum_{i=1}^{10}\sum_{j=i+1}^{10} f\left(w_i^{(k)}, w_j^{(k)}\right) \tag{33}$$

where $\left\{w_1^{(k)},\ldots\ldots w_{10}^{(k)}\right\}$ denotes the most likely top 10 to get the nrd word in Theme k, and $f\left(w_i, w_j\right)$ denotes normalised pointwise mutual information. Then:

$$f\left(w_i, w_j\right) = \frac{\log\left(p\left(w_i, w_j\right)\big/ p\left(w_i\right)p\left(w_j\right)\right)}{-\log p\left(w_i\right)p\left(w_j\right)} \tag{34}$$

where $p\left(w_i, w_j\right)$ is the probability that words w_i and w_j co-occur in the document and $p\left(w_i\right)$ is the marginal probability of word w_i.

3) Topic Strength. The calculation of theme intensity helps to assess the importance of each theme and theme words in the document, so as to select the theme concepts that are more suitable for the research topic as indicators. In practice, the theme strength is generally reflected by the number of documents containing the theme, i.e., the greater the number of documents containing the relevant theme, the greater the strength of the theme. The formula for calculating theme strength is:

$$S_k = \frac{\sum_N^i \theta_{ki}}{N} \tag{35}$$

where N denotes the number of documents in the document set, θ_{ki} denotes the probability of topic k appearing in the i th document, and S_k is the average probability of topic k in the document set, which is the intensity value of topic K.

4. Analysis of results

Standing on a new historical starting point, cultural communication work shoulders a new mission of the times. To enhance the national cultural soft power and external communication power, the new mainstream media, as the front-runner of cultural communication and the 'pilot' of external communication of the think tank and media matrix, need to clarify the evolution trend of the news form and understand the intensity of the news theme at different stages. Only by better integrating cultural connotation into the news theme can we lay a solid foundation for enhancing cultural self-confidence and cultural soft power.

4.1. Improving LDA model performance

4.1.1. Determination of optimal number of topics. Before performing LDA topic modelling, topic consistency under different topics needs to be calculated to ensure the accuracy of the LDA topic model training results. Too small a number of topics may make the number of keywords after the LDA model's word-splitting insufficient to depict all the current status of research in the research area, and too many topics may make too many keywords, resulting in the inability to condense the topics for analysing and researching. The number of topics traversed by the LDA model was set from 2 to 12 depending on the number of texts selected. Figure 1 shows the trend of theme consistency under different number of themes. With the gradual increase in the number of themes, the theme consistency value of the model shows irregular changes. When the number of topics is 4 the maximum theme consistency is -12.551, so in this experimental data, when the number of topics is 4, the optimal news topic analysis model can be constructed.

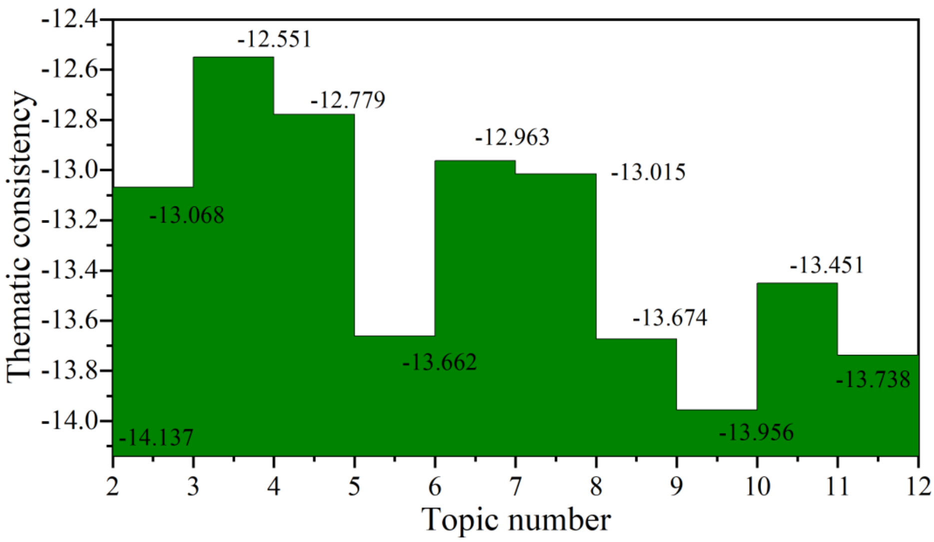


Fig. 1. Optimal topic number determination

4.1.2. Model performance tests. In order to verify the performance of the improved LDA model, this paper compares and evaluates the improved LDA with the traditional LDA model and BERT-LDA from two perspectives: topic consistency and topic clustering. The mean and standard deviation of the ten results of different model experiments are recorded during the experiment. First, thematic consistency calculation is performed. Topic consistency assessment refers to the extent to which the high-probability words in each topic are topic-consistent, and is used here as a way to verify the quality of the topic-word mapping of the improved model. Based on the theme consistency calculation method given in the previous section, the values are calculated based on the top 20 most probable words for each theme, and the theme consistency score on the dataset is obtained as shown in Figure 2.

As can be seen from the figure, under the experimental environment with the same objective conditions, the NPMI scores of the traditional LDA model show a constant trend with -16.74 as the starting value, and the overall fluctuation range is between [-16.74,-16.48], and the fluctuation range of the BERT-LDA model is between -16.63 and -14.84. While the improved LDA model takes -16.52 as the starting value, the overall trend is upward, and the topic consistency score obtained when the word generation probability is 1 can be up to -13.39, which is 19.14% and 9.77% higher than the traditional LDA model and the BERT-LDA model, respectively. Therefore, the improved LDA model is significantly better than the traditional LDA model and BERT-LDA model in topic consistency calculation, and can obtain more accurate consistency results in news topic analysis.

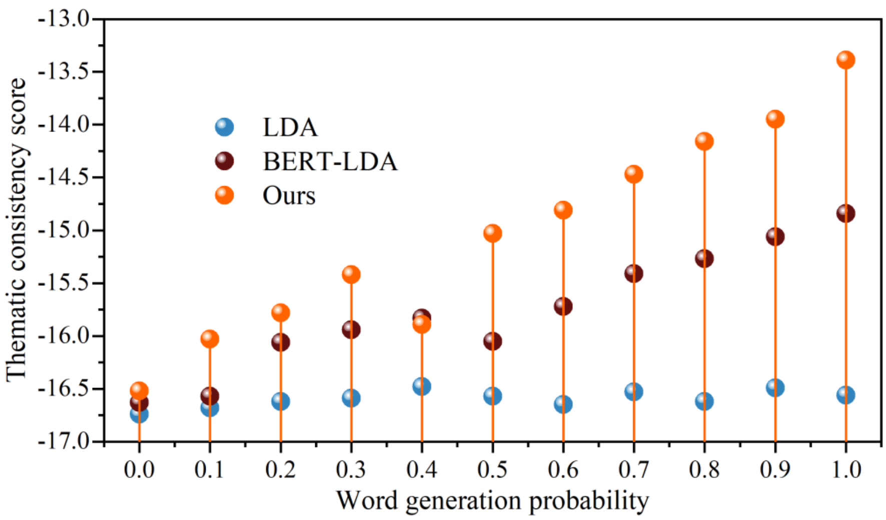


Fig. 2. Thematic consistency score

Secondly, the improved LDA model was compared with the LDA model and the BERT-LDA model in the topic clustering task. The performance of topic clustering was evaluated using the APMI metrics given in the previous section, where the APMI scores are always between 0.0 and 1.0, and the higher the score, the better the clustering performance it reflects. Figure 3 shows the NMI results obtained for different models on the dataset respectively when the number of topics is set to 4. According to the comparison of APMI score results, the MMI scores of the traditional LDA model are stable between 0.41 and 0.45 in the same experimental environment, and the improved LDA model is significantly better than the traditional LDA model after the word generation probability is more than 0.5, and the topic clustering performance of the improved LDA model is optimal when the word generation probability is 0.6, and the APMI score is as high as 0.51. Therefore, the improved LDA model has the best performance in topic clustering, and the APMI score is as high as 0.51. In topic clustering, the improved LDA model performs better than the traditional LDA model and can better cluster news topic patterns related to cultural communication.

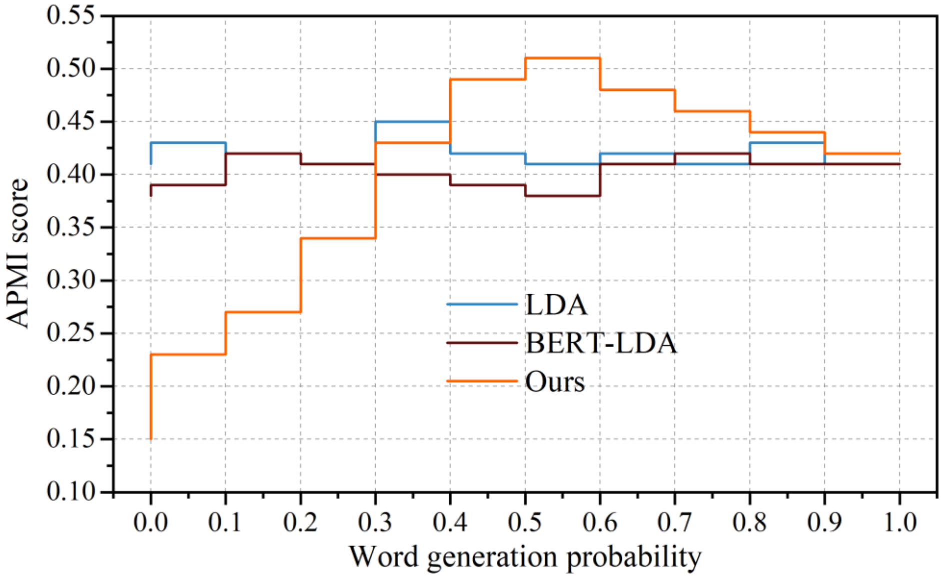
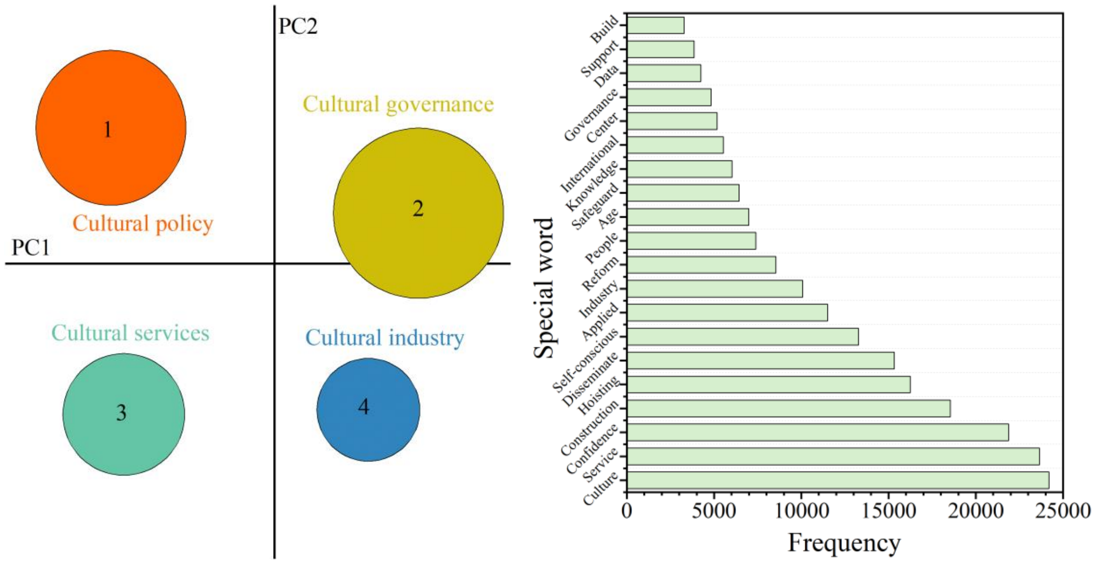


Fig. 3. Comparison of thematic clustering results

4.2. *Example data analysis*

4.2.1. Visualisation of news topics. The pyLDAvis module was used to study the visualisation of the news topic model. When sorting the topic feature words, the feature effect of words under the corresponding topic can be obtained by adjusting the λ value. If the word generation probability is close to 1, then the words that occur more frequently under the topic are more relevant to the topic. If the word generation probability is close to 0, then words that are more special and exclusive under that topic are more relevant to the topic. In contrast, in this paper, after balancing the weights of the two, we give priority to the uniqueness of the words under the topic, and regulate the word generation probability to 0.5.Figure 4 shows the visualisation results of the news topic model, where Figures 4(a)~(b) show the clustering results and the corresponding list of feature words, respectively.

For the extraction of news themes related to cultural communication in the dataset, the obtained results contain four themes, which are interpreted as four news frames, and themes 1~4 are the data of cultural policy, cultural governance, cultural services and cultural industry, respectively. The larger the circle in the graph indicates that the theme appears more frequently in all documents, while the distance between the circles represents the distance between themes. In the news frame of Theme 1 Cultural Policies, the main contents include policy implementation opinions, draft laws and regulations, regulations related to cultural change, and reports on cultural reform work. In the news framework of Theme 2 Cultural Governance, the main content includes technological transformation and upgrading of cultural communication in the context of technological support and change. In the news framework of Theme 3 Cultural Services, the main contents cover cultural protection, optimisation of cultural services, cultural life, social and cultural construction, etc. In Theme 4, Cultural Industries, the main content emphasises the level of development and coverage of cultural industries based on cultural communication, as well as the international development of cultural industries. Relatively speaking, the characteristic word ‘culture’ appears as frequently as 24,185 times, which reflects the importance of the news media in the process of cultural communication. Relying on the news media to expand the depth and breadth of cultural communication provides a reliable channel for people to obtain cultural information in order to enhance their cultural confidence and self-awareness.



(a) Theme clustering (b) Special word

Fig. 4. Visual results of the news theme model

4.2.2. Evolution of news topic patterns. Theme intensity is analysed from the perspective of theme, and the theme intensity of the k th theme in a certain time period is analysed according to the word frequency statistics, and the theme intensity of the theme in each year is obtained to analyse the evolution trend of each hot theme, including rising, declining, smooth, and so on. Based on the theme intensity calculation method given in the previous section, the news theme pattern evolution trend applied to cultural communication is obtained as shown in Figure 5.

As can be seen from the figure, the four culture-related news theme patterns from Topic1 to Topic4 all show an upward trend, and their corresponding themes are cultural policy, cultural communication, cultural services and cultural industry. Among them, the theme intensity of 'cultural policy' has been on the rise from 2010 to 2023, with its theme intensity increasing from 0.009 in 2010 to 0.148 in 2023, an increase of 14.44 times. This shows on the one hand that the state has paid more and more attention to the policies related to cultural communication during the thirteen years, and on the other hand, it also shows that the national demand for the policies related to cultural communication is increasing, although its theme intensity slightly declined in 2019, but then re-grew in 2020, which shows that the theme is in the period of rapid grinding, and on the whole, it seems that the cultural policy has been in the state of rising from 2010 to 2023. Overall, cultural policy appears to be on the rise from 2010 to 2023. The overall trend of the theme 'cultural governance' shows an upward trend, in which the theme intensity was 0.051 in 2010, and reached 0.088 in 2013, and then showed a certain decay to 0.069 in the following two years, and then began to increase in 2016, and finally in 2023 the theme intensity reached 0.263. This shows that the continuous growth of technology provides more and more possibilities for cultural communication, and the radical change of news forms in the era of integrated media also provides a reliable carrier for the rapid development of cultural communication. The intensity of the theme 'cultural services' has been increasing since 2010, with a slight decrease in 2015, and reaches around 0.169 in 2023. This indicates that research on the theme of cultural services in the field of cultural communication has been continuously paid attention to, and that the news media have been used to help the public clarify the contents of cultural communication and provide more diverse cultural services for the public. The intensity of the theme of 'cultural industry' continues to rise, from 0.079 in 2010 to 0.177 in 2023, an increase of 124.05%. The prosperous development of the cultural industry cannot be separated from the high-speed development of news media forms in the new era, and the technologically-supported news forms have been transformed from traditional paper media to digital media, giving the cultural industry more development possibilities, promoting the breadth and depth of cultural dissemination, and becoming the main support form for opening up the cultural sharing data islands and realising the interconnection and interoperability of cultural data.

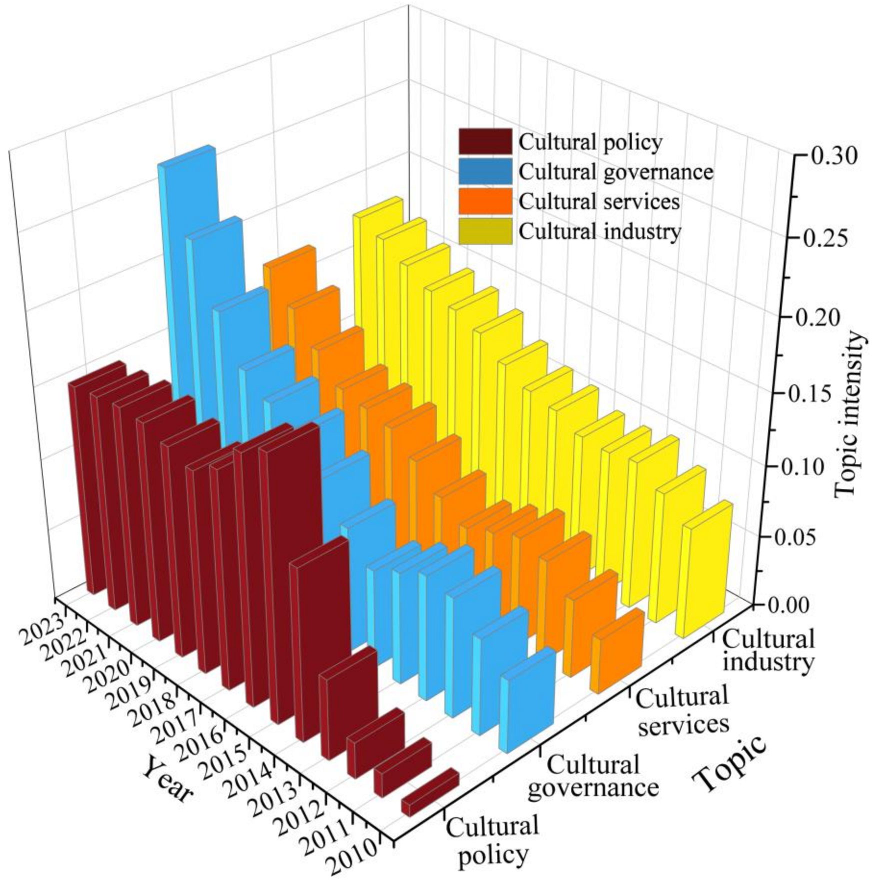


Fig. 5. The trend of news theme evolution

4.2.3. Degree of contribution to news topics. In order to further analyse the extent of the contribution of news theme pattern innovations applied to cultural communication, this paper applies a factor model to analyse the extent of the contribution of different news themes as a means of analysing the distribution of the contribution of each news theme pattern to cultural communication in different time periods. Figure 6 shows the results of the degree of contribution of news theme patterns to cultural communication in different time periods.

Overall, the news theme of 'cultural governance' has a relatively high contribution to cultural communication from 2010 to 2023, with a mean value of 0.091, while the news theme of 'cultural services' has a relatively low contribution to cultural communication, with a mean value of only 0.091. In terms of fluctuation, the range of fluctuation of the news thematic pattern of 'cultural policy' is larger, with a maximum of 0.17, while the range of fluctuation of the news thematic pattern of 'cultural governance' is more stable, mainly focusing on the range of 0.09 to 0.05, and the range of fluctuation of the news thematic pattern of 'cultural governance' is more stable, mainly focusing on the range of 0.09 to 0.05. The fluctuation range of the news theme pattern of 'cultural governance' is more stable, mainly focusing on 0.09~0.12. In order to carry out cultural communication under the news theme pattern, it is necessary to optimise the expression form of the news media and match it with cultural diversity, so as to give more possibilities to cultural communication, and help satisfy the increasing cultural needs of the masses. Comparatively speaking, 'cultural services' and 'cultural industries' are better contributors to the dissemination of culture, through the news media to realise the diversity of cultural services, promote the organic integration of cultural industries, and realise the widespread dissemination of the spirit of culture. It provides a reliable vehicle for improving national cultural literacy and the country's cultural soft power.

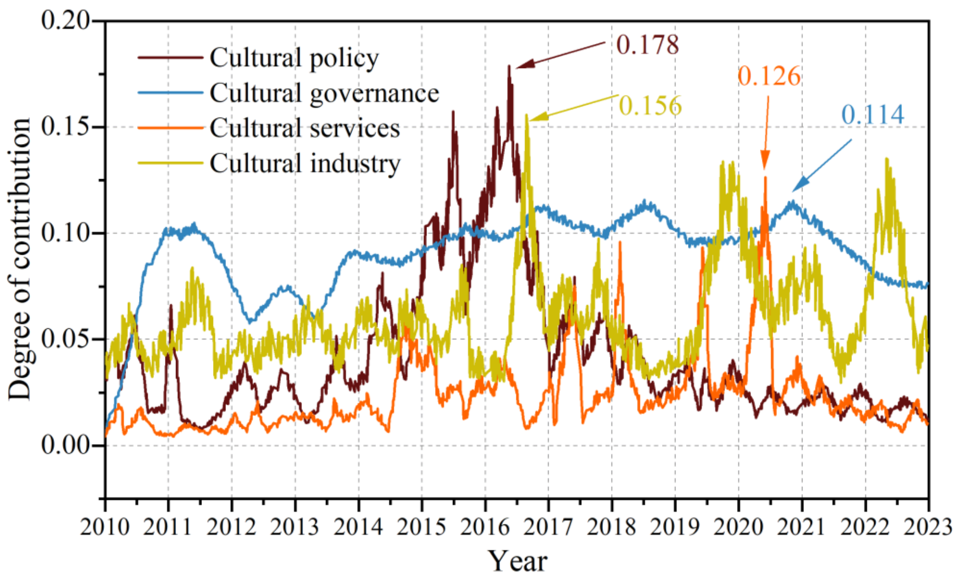


Fig. 6. The degree of contribution of news topics

5. Conclusion

In this paper, the improved LDA model is used to analyse the evolution of news topic patterns applying cultural communication. When the improved LDA model is applied to the analysis of news topic pattern evolution, the optimal number of topics is 4 to get a topic consistency value of -12.551, at this time, the news topic pattern obtained can completely represent the situation of cultural communication. When the word generation probability is 1, the obtained theme consistency score can reach -13.39, which is 19.14% and 9.77% higher than the traditional LDA model and BERT-LDA model respectively. Using the improved LDA model, a total of four themes of news forms applied to cultural communication were obtained, namely, cultural policy, cultural communication, cultural service and cultural industry. Among them, the theme intensity of 'cultural policy' increases from 0.009 in 2010 to 0.148 in 2023, which is a 14.44-fold increase. The 'cultural governance' news theme pattern has a relatively high contribution to cultural communication, with a mean value of 0.091. Using the improved LDA model, we can analyse the evolution trend of the news theme pattern in different periods, and also explore the contribution of different theme patterns to cultural communication, which can provide a reference for the application and optimisation of the news media in cultural communication. It can also explore the degree of contribution of different themes to cultural communication, which can be used by news media in cultural communication.

The extraction and evolution analysis of news form themes for cultural communication is a relatively popular, dynamic, challenging, practical and valuable research field. Although many scholars are currently conducting related research, this paper has also achieved better results in theme extraction, theme quality evaluation and theme evolution. However, there are still some deficiencies in the research field of theme extraction and evolution, which need to be further improved and refined in order to obtain better results in the future in the field of theme extraction and evolution of news and cultural forms, so as to improve the scope of cultural communication and realise the innovative development of news media forms.

Project Fund

The study was supported by "Research on Film Cultural Landscape Based on Intelligent Monitoring Project of Beijing Institute of Petro-chemical Technology, China" (Project Number: 22032205003/025).

References

[1] Cheng, Y. (2022). Analysis on the Communication Path of Major News Reports Through All Media. Scientific

and Social Research, 4(8), 56-62.

[2] Zhang, N., Huang, H., Su, B., Zhao, J., & Zhang, B. (2014). Information dissemination analysis of different media towards the application for disaster pre-warning. PloS one, 9(5), e98649.

[3] Cai, S., & Xie, F. (2020). Internet Thinking for Omnimedia Reconstruction of News Production and Dissemination Mechanism of Chinese Traditional Mass Media . Probe-Media and Communication Studies, 2020, 2 (1).

[4] Luo, J. (2021, December). Intelligent Technology in the All-Media Era Promotes the Spread of Chinese Culture. In International Conference On Signal And Information Processing, Networking And Computers (pp. 573-579). Singapore: Springer Nature Singapore.

[5] da Silva, G. C., & Sanseverino, G. G. (2020). Business model innovation in news media. Media and Communication, 8(2), 28-39.

[6] Hess, K., & Waller, L. (2020). Charting the media innovations landscape for regional and rural newspapers. Australian journalism review, 42(1), 59-75.

[7] Lowrey, W., Deavours, D., & Singleton, W. (2023). Agents of meta: Institutional actors in the journalism space and the innovation of local news. Journalism, 24(6), 1155-1173.

[8] Maheshwari, M. (2015). Innovation in creation and delivery of news content: A Case study of India Today news channel. International Journal in Management & Social Science, 3(8), 352-366.

[9] Belair-Gagnon, V., & Steinke, A. J. (2020). Capturing digital news innovation research in organizations, 1990–2018. Journalism studies, 21(12), 1724-1743.

[10] Wang, C. (2024, August). Innovation of Radio and Television News Production in the Era of Media Convergence. In 2024 International Conference on Humanities, Arts, and Cultural Industry Development (HACID 2024) (pp. 106-115). Atlantis Press.

[11] García-Avilés, J. A., Carvajal-Prieto, M., Arias, F., & De Lara-González, A. (2019). How journalists innovate in the newsroom. Proposing a model of the diffusion of innovations in media outlets. The journal of media innovations, 5(1), 1-16.

[12] Useille, P. (2017, July). Innovations in journalism and new information experiences. In IAMCR 2017.

[13] Palacios, M., Barbosa, S., da Silva, F. F., & da Cunha, R. (2016). Mobile journalism and innovation: A study on content formats of autochthonous news apps for tablets. In Emerging perspectives on the mobile content evolution (pp. 239-262). IGI Global.

[14] Luengo, M., & Herrera-Damas, S. (Eds.). (2021). News Media Innovation Reconsidered: Ethics and Values in a Creative Reconstruction of Journalism. John Wiley & Sons.

[15] Sixto-García, J., Silva-Rodríguez, A., Rodríguez-Vázquez, A. I., & López-García, X. (2023). Redefining journalism narratives, distribution strategies, and user involvement based on innovation in digital native media. Journalism, 24(6), 1322-1341.

[16] Breese, E., & Luengo, M. (2016). News innovations and enduring commitments. Cambridge University Press.

[17] Gynnild, A. (2014). Journalism innovation leads to innovation journalism: The impact of computational exploration on changing mindsets. Journalism, 15(6), 713-730.

[18] García-Avilés, J. A. (2020). Reinventing television news: Innovative formats in a social media environment. Journalistic Metamorphosis: Media Transformation in the Digital Age, 143-155.

[19] Nunes, A. C. B., & Canavilhas, J. (2020). Journalism innovation and its influences in the future of news: a European perspective around Google DNI Fund initiatives. Journalistic metamorphosis: Media transformation in the digital age, 41-56.

[20] Imre, I., & Wenger, D. (2020). Where newsroom leaders see technology facilitating innovation in local TV news. Electronic news, 14(4), 151-167.

[21] Lopezosa, C., Codina, L., Fernández-Planells, A., & Freixa, P. (2023). Journalistic innovation: How new formats of digital journalism are perceived in the academic literature. Journalism, 24(4), 821-838.

[22] Shun Ying. (2024). Guests' Aesthetic experience with lifestyle hotels: An application of LDA topic modelling analysis. Heliyon(16),e35894-e35894.

[23] Xuan WANG,Bofeng ZHANG,Mingqing HUANG,Furong CHANG & Zhuocheng ZHOU. (2019). Improved LDA Model for Credibility Evaluation of Online Product Reviews. IEICE Transactions on Information and Systems(11),2148-2158.

[24] Mayor Eric & Miani Alessandro. (2023). A topic models analysis of the news coverage of the Omicron variant in the United Kingdom press. BMC Public Health(1),1509-1509.

[25] Wang Yalin,Li Yaokuang & Wu Juan. (2022). Digital identities of female founders and crowdfunding performance: an exploration based on the LDA topic model. Gender in Management: An International Journal(5),659-678.