PeerJ Computer Science

An adaptive method for determining the optimal number of topics in topic modeling

Yang Xu[1], Yueyi Zhang[1], Yefang Sun[2] and Hanting Zhou[3]

[1] College of Economics and Management, China Jiliang University, Hangzhou, Zhejiang, China
[2] College of Economics and Management, China Jiliang University College of Modern Science and Technology, Yiwu, Zhejiang, China
[3] School of Management, Hangzhou Dianzi University, Hangzhou, Zhejiang, China

ABSTRACT

Topic models have been successfully applied to information classification and retrieval. The difficulty in successfully applying these technologies is to select the appropriate number of topics for a given *corpus*. Selecting too few topics can result in information loss and topic omission, known as underfitting. Conversely, an excess of topics can introduce noise and complexity, resulting in overfitting. Therefore, this article considers the inter-class distance and proposes a new method to determine the number of topics based on clustering results, named average inter-class distance change rate (AICDR). AICDR employs the Ward's method to calculate inter-class distances, then calculates the average inter-class distance for different numbers of topics, and determines the optimal number of topics based on the average distance change rate. Experiments show that the number of topics determined by AICDR is more in line with the true classification of datasets, with high inter-class distance and low inter-class similarity, avoiding the phenomenon of topic overlap. AICDR is a technique predicated on clustering results to select the optimal number of topics and has strong adaptability to various topic models.

Subjects Algorithms and Analysis of Algorithms, Data Mining and Machine Learning, Text Mining, Neural Networks
Keywords Topic modeling, Inter-class distance, AICDR, Optimal number of topics

Submitted 16 September 2024
Accepted 30 January 2025
Published 28 February 2025

Corresponding author
Hanting Zhou,
zhouhanting@hdu.edu.cn

Academic editor
Giovanni Angiulli

Additional Information and Declarations can be found on page 21

DOI **10.7717/peerj-cs.2723**

INTRODUCTION

Modern information systems generate a huge number of texts such as policies, news, and comments. Analysis of big data is impossible without the construction of formalized mathematical models (*Ignatenko et al., 2018*). Topic models are gaining popularity in social sciences (*Peng & Zhang, 2024*). Models such as non-negative matrix factorization (*Kekere, Marivate & Hattingh, 2023*) and Latent Dirichlet Allocation (LDA) (*Ding, Kang & Ren, 2024*). These techniques rely on statistical modeling to mine the semantic information implied in large-scale text datasets and classify massive texts according to different topics, and each text cluster is interpreted by different topic words (*Zhao et al., 2018*).

In topic modeling, the determination of the number of topics (parameter K) is crucial for text analysis. Identification of the optimum number of topics is one of the main challenges for the existing methods for topic modeling. Choosing too few topics will result in overly broad topics, loss of information and omission of topics, while choosing too

many will result in the over-clustering of a *corpus* into many small, noisy and highly-similar topics (*Greene, O'Callaghan & Cunningham, 2014*), and the topics are prone to overlap and merge. Thus, the selection of K is particularly important.

Perplexity is a commonly used measurement in information theory to characterize topic quality, with lower perplexity denoting a better probabilistic model (*Huang, Ma & Chen, 2017*). To solve the problem of multiple elbows in perplexity, a new method for calculating the rate of change of perplexity (*Zhao et al., 2015*) is used to determine the appropriate number of topics. Similarly, coherence is an indicator used to evaluate the quality of a topic and is also applicable to the selection of the number of topics (*O'Callaghan et al., 2015*). It evaluates the frequency of word co-occurrence and measures the correlation between words in a topic. Perplexity and coherence are mostly used in probabilistic topic models, which have limited adaptability.

The classical probabilistic topic models (LDA) are widely used in topic classification tasks (*Chen et al., 2017*). Therefore, most studies use probabilistic topic models to research the selection of topic numbers. A density-based adaptive LDA model selection approach that integrates the concept of density clustering to adaptively determine an appropriate number of topics (*Cao et al., 2009*; *Lu et al., 2013*; *Wang et al., 2019*). Some improvements or combinations of metrics or evaluation indicators are equally applicable to this task. An improved online LDA (IOLDA) uses Jensen–Shannon (JS) scatter to calculate the association between topics and screen out similar topics, and the JS scatter is smaller when the number of topics is close to the optimal value (*He, Chen & Du, 2015*). The combination of JS divergence and perplexity is used to select the optimal number of topics, which improves the problem of using the perplexity formula alone (*Peng & Yuefen, 2016*). A comprehensive index of perplexity, isolation, stability, and consistency is constructed, which can effectively determine the optimal number of topics in the LDA model (*Gan & Qi, 2021*). The combination of stability and coherence is also applicable to neural topic models to select the number of topics (*Koltcov et al., 2024*).

To reduce the reliance on probability distributions or topic-term matrices, a stability analysis significantly enhances its versatility, rendering it applicable to topic models and classification methods (*Greene, O'Callaghan & Cunningham, 2014*). The elbow method is a commonly used technique for determining the optimal number of clusters in K-Means clustering. Its core idea is to identify the best number of clusters by observing the relationship between the sum of squared errors (SSE) of the clustering results and different K (*Liu & Deng, 2021*; *Shi et al., 2021*). The silhouette method is another well-known method with decent performance in estimating the potential optimal cluster number (*Arbelaitz et al., 2013*; *Rodriguez et al., 2019*). Similarly, it's not uncommon to use metrics to evaluate the optimal cluster number (*Ding, Tarokh & Yang, 2017*; *Zheng et al., 2023*), they may also be applied to topic models.

The optimal number of topics should produce a good clustering result, *i.e.*, high similarity of texts within topics and low similarity of texts within topics. In the literature, most methods for selecting the number of topics rely on probability distributions. For non-probabilistic topic models, these methods are difficult to apply. Motivated by this, this

article proposes a new method for determining the number of topics in topic models based on inter-class distance, named average inter-class distance change rate (AICDR). AICDR calculates the diameters for each class and the merged class, then derives the inter-class distance as the square root of the diameter difference between them. It subsequently computes the average inter-class distance across varying numbers of topics and identifies the optimal number of topics based on the average inter-class distance change rate. AICDR is not bound by the constraints of the topic models, and the selection of the optimal number of topics is done only through the clustering results. The inter-class distance between topics is calculated based on Ward's method (*Murtagh & Legendre, 2014*), the number of topics corresponding to the maximum AICDR is the optimal number of topics. To verify the feasibility and adaptability of the proposed method, this research compares stability analysis and elbow method to select the optimal K value. The contributions in this article are summarized as follows:

- Ward's method is a method of hierarchical clustering that aims to produce classes by minimizing the intra-class variance. It merges the two classes with the smallest sum of distances (sum of squared deviations) until the condition is satisfied. We define the inter-class distance as the square root of the difference between the diameter of the merged class and the diameter of the original classes. The sum of the Euclidean distances from all objects in a class to the class mean is defined as the diameter of the class (*i.e.*, the squared deviation). The diameter of class indicates the compactness of the sample, while the distance between classes indicates the degree of separation between classes.
- A new method for determining the number of topics is proposed, named AICDR. AICDR considers inter-class similarity and intra-class similarity comprehensively, the K corresponding to its maximum is the optimal number of topics, which avoids topic overlap. Mainly, it is not limited by topic models or clustering methods and has good adaptability to most methods.
- Through experiments on several public datasets, the feasibility of the proposed method is fully verified, which provides a useful reference for similar research.

The remainder of this article is organized as follows. "Related Work" introduces the previous research on topic models; "Methods" describes the principles of several topic models used in experiments, the formulaic definition of AICDR, and inter-class distance; "The Workflow of AICDR" introduces the process of selecting the optimum number of topics; "Experiments" talks about the experimental results and analysis; "Conclusion" concludes the article and proposes the future work.

RELATED WORK

Generally, topic modeling methods are mainly classified as probabilistic and non-probabilistic in the literature (*Kherwa & Bansal, 2018*). This section reviews three more detailed branches used for developing topic modeling algorithms: probabilistic topic models, matrix factorization-based topic models and neural topic models.

Probabilistic topic models

Introduced as an initial probabilistic approach to topic modeling, probabilistic latent semantic analysis (PLSA) has its limitations as it fixes the distributions of topics and words within a document. LDA (*Blei, Ng & Jordan, 2003*) addresses this by applying Dirichlet priors to the distributions, thus allowing for a probabilistic assignment of topics and words. LDA has shown remarkable efficacy, attracting persistent research attention (*Altarturi, Saadoon & Anuar, 2023*).

Due to the limited co-occurrence information of words in short texts, traditional long-text topic modeling algorithms (*e.g.*, PLSA and LDA) based on word co-occurrences cannot solve this problem very well (*Qiang et al., 2020*). A collapsed Gibbs Sampling algorithm for the Dirichlet multinomial mixture (GSDMM) (*Yin & Wang, 2014*) model for short text clustering defaults to all words in the document following a topic. WV +GSDMMK (*Agarwal, Sikka & Awasthi, 2024*) improves service-to-topic mapping by determining semantic similarity among features, and K-means clustering is applied on service to topic representation. Biterm topic model (BTM) (*Cheng et al., 2014*) addresses the challenge of short text topic modeling by directly modeling the generation of word co-occurrence patterns, or biterms, throughout the *corpus*. Therefore, a word co-occurrence network-based model (WNTM) (*Zuo, Zhao & Xu, 2016*) represents the word co-occurrence network back to a pseudo-document set, where a word forms a pseudo document with adjacent words and models the thematic distribution of each word. Similarly, pseudo-document-based topic model (PTM) (*Zuo et al., 2021*), also utilizes word co-occurrence information to construct pseudo-document and topic modeling.

Matrix factorization-based topic models

Latent semantic indexing (LSI) (*Papadimitriou et al., 1998*) uses singular value decomposition techniques to capture the latent semantic relationships between words, while non-negative matrix factorization (NMF) approximates the reconstruction of the original matrix by decomposing the data matrix into two non-negative matrices (*Xu, Liu & Gong, 2003*). Compared to LSI, the advantage of NMF lies in its non-negativity and sparsity constraints, making the decomposition results easier to interpret. NMF has been successfully applied to topic modeling and text clustering, due to its superior performance in clustering high-dimensional data (*Carbonetto et al., 2022*).

To continuously improve the performance of NMF, some scholars have considered constructing graphs, *i.e.* data graph and feature graph, to explore the geometric structure of data manifold and feature manifold (*Cai et al., 2011*; *Gu & Zhou, 2009*). Semantic information can also be embedded in NMF to adapt to short texts. Semantics-assisted NMF (SeaNMF) (*Shi et al., 2018*) effectively incorporates word-context semantic correlations into the model. Word co-occurrence regularized NMF (WC-NMTF) (*Salah et al., 2018*) maps frequently co-occurring words roughly to the same direction in the latent space to reflect the relationships between them. Both SeaNMF and WC-NMTF use point-wise mutual information (PMI) (*Levy & Goldberg, 2014*) Neighbourhood assistance-based NMF (NaNMF) (*Athukorala & Mohotti, 2022*) introduces text similarity as a regularization constraint to improve classification performance. Regularized asymmetric

NMF (RANMF) (*Aghdam & Zanjani, 2021*) is the same way. The incorporation of regularization constraints (*i.e.*, regularizers) in NMF is an effective approach. NMF-WR (*Li et al., 2024*) integrates the Wasserstein metric into the NMF framework to enhance semantic representation and improve the reliability and interpretability of text embeddings.

Neural topic models

Recent advances in neural variational inference have spawned a renaissance in deep latent variable models (*Miao, Yu & Blunsom, 2015*). Unlike traditional Bayesian topic models (*e.g.*, PLSA and LDA), neural topic models use deep neural networks to approximate the intractable marginal distribution and thus gain strong generalization ability.

Using neural methods to replace providing parameterizable distributions on topics, permits training by backpropagation in the framework of neural variational (*Miao, Grefenstette & Blunsom, 2017*). Autoencoding variational inference for topic model (AVITM) (*Srivastava & Sutton, 2017*) introduces an autoencoder to approximate the posterior distribution, improving efficiency and accuracy. Neural variational gaussian mixture topic model (NVGMTM) (*Tang et al., 2022*) uses Gaussian distribution to depict the semantic relevance between words in the topics, each topic is considered as a multivariate Gaussian distribution over words in the word-embedding space. To fully utilize the discreteness of the topic space, the discrete-variational-inference-based topic model (DVITM) (*Gupta & Zhang, 2023*), learns dense topic embeddings homomorphic to word embeddings *via* discrete variational inference. Self-attention mechanism can capture the dependency relationships within the sequence (*Vaswani et al., 2017*). Therefore, topic attention model (TAM) (*Wang & Yang, 2020*) utilizes document-specific topic proportions and global topic vectors learned from neural topic model in the attention mechanism.

METHODS

Topic models

LDA

LDA is a probabilistic statistical model for mining topic distributions and word distributions in document collections, identifying topic information hidden in document collections or *corpus*. The basic idea in LDA is that each document is represented as a probability distribution over hidden topics, while each topic is characterized as a probability distribution over some words. The generative process of the LDA model for each document is written as follows:

1) Draw each topic parameter $\beta_k \sim Dirichlet\ (\Phi)$, for each topic $k \in [1, \ldots K]$.

2) For each document $d \in D$:

a) Sample a topic distribution $\theta \sim\ Dirichlet\ (a)$

b) For each of the N words w_n:

Sample a topic $z_{m,n}\ \sim Multionmial\ (\theta_m)$

Sample a word $w\ \sim\ Multionmial\ (\beta_{\mathrm{k}}\ )$ from $p(w_n|z_{m,n}, \beta_k)$

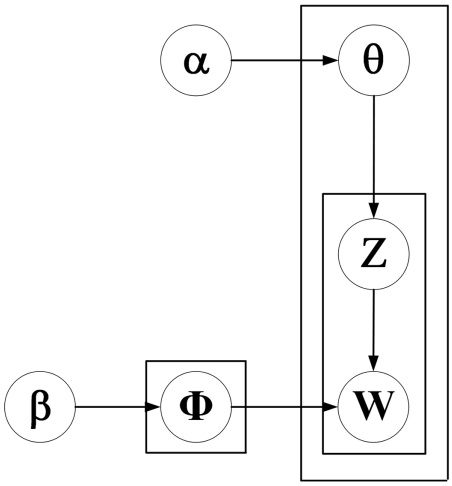

Figure 1 **Graphical model representation of LDA.** Full-size DOI: 10.7717/peerj-cs.2723/fig-1

In Fig. 1, Φ represents word distribution, θ represents topic distribution. α is a parameter of the prior distribution of topic distribution θ and β is a parameter of the prior distribution of word distribution. Z and W represent the distribution of document topic and word topic, respectively.

GSDMM

Dirichlet multinomial mixture (DMM) respectively chooses Dirichlet distribution for topic-word distribution Φ and document-topic distribution θ as prior distribution with parameter α and β. DMM samples a topic Z for the document by multinomial distribution θ, and then generates all words in the document from topic Z by multinomial distribution Φ. The generative process for DMM is described as follows:

1) Sample a topic distribution $\theta \sim Dirichlet\ (a)$.

2) For each topic $k \in [1, \ldots K]$:

3) For each document $d \in D$:

a) Sample a topic $z_d \sim Multionmial\ (\theta)$

b) For each word $w \in \left[w_{d,1}, \ldots,\ w_{d,n_d}\right]$:

Sample a word $w \sim\ Multionmial\ (\Phi_{z_d})$.

The graphical model representation of GSDMM is shown in Fig. 2. Gibbs sampling algorithm for DMM (GSDMM) assumes that each text is sampled by a single topic.

NMF

The NMF method has been successfully applied to topic modeling, due to its superior performance in clustering high-dimensional data. A text dataset can be represented by a matrix $X \in R_+^{m \times n}$. X approximates through two matrices $W \in R_+^{m \times k}$ and $H \in R_+^{n \times k}$, *i.e.*, $X \approx WH^T$. The NMF formula is as follows:

$$\begin{aligned} F &= min\ \left|\left|X - WH^T\right|\right|_F^2 \\ &s.t.\ W \geq 0,\ H \geq 0. \end{aligned} \tag{1}$$

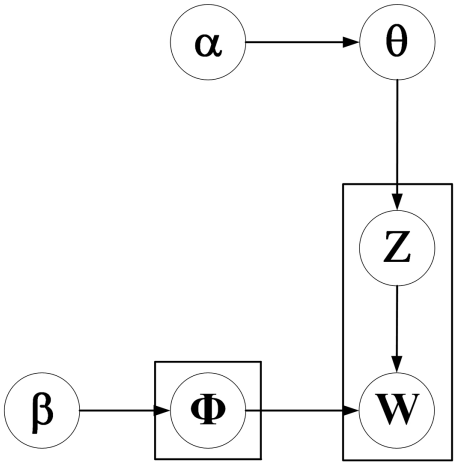

Figure 2 Graphical model representation of GSDMM. Full-size DOI: 10.7717/peerj-cs.2723/fig-2

The index of the maximum value for each row $argmax \sum_i^m W_{(i,:)}$ represents the topic of the *i-th* document. Similarly, $argmax \sum_j^n H_{(j,:)}$ represents the topic of the *j-th* word.

SeaNMF

SeaNMF effectively incorporates the word-context semantic correlations into the model, where the semantic relationships between the words and their contexts are learned from the skip-gram view of the *corpus*. These correlations can be viewed as an alternative form of the word co-occurrence. It can overcome the problem that arises due to the data sparsity. Therefore, its objective function is as follows:

$$\begin{aligned} F = min\ & ||X - WH^T||_F^2 + \alpha||P - HQ^T||_F^2 \\ & s.t. W \geq 0,\ H \geq 0,\ Q \geq 0, \end{aligned} \tag{2}$$

where P represents positive pointwise mutual information (PPMI) and $P \in R_+^{n \times n}$, and Q is a randomly initialized factor matrix and $Q \in R_+^{n \times k}$. SeaNMF incorporates word co-occurrence information as a regularization constraint to compensate for the sparsity of short text.

Table 1 outlines the specific characteristics of LDA, GSDMM, NMF, and SeaNMF.

The proposed AICDR

For unstructured document sets, both the document content and the number of relevant topics are unknown, and the best number of topics is unknown. An insufficient number of topics may lead to underfitting of the model, a higher number of topics could result in a model that is too complex, making topic overlap. It is necessary to select the appropriate number of topics. The best clustering result, with a specified number of topics, should exhibit high intra-cluster similarity within the same topic and lower inter-cluster similarity between different topics.

In this section, we introduce the proposed method for determining the number of topics, AICDR. When the value of AICDR is maximized, the corresponding number of topics is optimal. Using AICDR to determine the number of topics results in higher inter-

Table 1 Description of topic models.

Models	Type	Applicability
LDA	Probabilistic generative	Long text
GSDMM	Probabilistic generative	Short text
NMF	Non-probabilistic	Long text
SeaNMF	Non-probabilistic	Short text

class distance or lower inter-class similarity. Inter-class distance is used to describe the distance between different classes after categorization, which quantifies the degree of difference or similarity between different classes.

There are various methods for defining inter-class distance, including the single linkage method, complete linkage method, group average method, centroid method, and sum of squares method. The central method and sum of squares method consider the global features of the class. Ward's method (*Murtagh & Legendre, 2014*) is a hierarchical clustering approach based on the idea of sum of squared deviation, which defines the distance between two populations by merging the differences in intra-group variances before and after. This is precisely sum of squares method's definition of inter-class distance. We utilize Ward's method (*i.e.*, sum of squares method) to define the diameter of the class and the inter-class distance. Therefore, the steps of AICDR are as follows:

(1) Calculate the diameter of the class.

$$D_p = \sum_{i=1}^{m} \sqrt{(t_i - \bar{t})^T (t_i - \bar{t})}, \tag{3}$$

$$\bar{t} = \frac{1}{m} \sum_{i=1}^{m} t_i. \tag{4}$$

D_p represents the diameter of the *p-th* class, t_i represents the text vector, and $\bar{t}$ represents the mean text vector. Define the sum of the Euclidean distances from all objects in a class to the class mean as the diameter of the class (*i.e.*, the sum of squared deviations). The ideal clustering result should be a greater intra-class similarity, *i.e.*, a smaller intra-class distance, or a smaller diameter of the class.

(2) Calculate the inter-class distance.

$$D(p,q) = \sqrt{D_{p,q} - D_p - D_q}. \tag{5}$$

$D_{p,q}$ represents the diameter of the merged class, $D(p,q)$ represents the inter-class distance between the *p-th* class and the *q-th* class. Merge two classes into a single class and compute its diameter; the inter-class distance is then the square root of the difference between the new diameter and the sum of the diameters of the original classes. After classification, the inter-class similarity should be small, *i.e.* the inter-class is large.

(3) Calculate average inter-class distance.

$$ave\_dis(k) = \frac{1}{\frac{k(k-1)}{2}} \sum_{p=1}^{k} \sum_{q=p+1}^{k} D(p,q). \tag{6}$$

The $ave\_dis(k)$ denotes the average inter-class distance for a topic number of K. As the number of topics increases, the value of the $ave\_dis(k)$ decreases, indicating that the inter-class distance between topics decreases gradually. Although the average inter-class distance method may generate meaningful results in some cases, it is not stable and its value decreases as the number of topics increases.

(4) Calculate AICDR.

$$AICDR(k) = abs(ave\_dis(k+1) - ave\_dis(k)) \tag{7}$$

The larger the value of AICDR, the more the increase (decrease) in the number of topics affects the structure, and the corresponding parameter K is the optimal number of topics.

The complete AICDR algorithm is found in Algorithm 1.

THE WORKFLOW OF AICDR

AICDR is a method for selecting the number of topics based on clustering results. Firstly, it is necessary to apply the topic model to a complete dataset and obtain clustering results under different K values, $K\epsilon\ [K_{min}, K_{max}]$. Then, use AICDR to determine the optimal number of topics and re-model. The overall process of selecting an appropriate number of topics based on AICDR is presented, which includes text preprocessing, text vectorization, text pre-clustering, selecting the number of topics, and topic modeling. Figure 3 shows the complete process of selecting the number of topics.

Text preprocessing

Due to the large amount of data, noise, and other characteristics of text datasets, text preprocessing is required before topic modeling. It can filter out invalid information to improve the extraction accuracy of the core keywords. As shown in Fig. 4.

Text segmentation: In English, words are separated from each other by spaces, so English word separation is relatively simple. In Chinese, the text is usually separated using "Jieba" participles. "Jieba" separates the text precisely without redundant words, which better summarizes and expresses the topic and content of the text.

Delete stop words: There are lots of words in the text that have no practical meaning. These words are called stop words, such as "we", "of", "yes" and so on. Excessive stop words diminish the model's ability to generalize and increase computational costs. Therefore, some meaningless stop words should be deleted.

Load retained words: Contrary to stop words, retained words highlight key features of the text. These words should not be separated in the text. Such as "give up", "hold on" and

Algorithm 1 **Algorithm for the proposed AICDR.**

Data: Text vectorization matrix $X \in R_{+}^{m \times n}$, the number of topics (K)

define *ave_dis* list (L)
for each $k \in \{2, \ldots, K\}$ **do**
  text classification
  define diameter list ($L1$)
    for each $k_i \in \{1, \ldots, k\}$ **do**
      calculate D_{k_i} by Eq. (3)
      $L1$ append D_{k_i}
    end
  define diameter list ($L2$)
  for each $D_{k_p} \in \{L1_1, \ldots, L1_{m-1}\}$ **do**
    for each $D_{k_q} \in \{L1_2, \ldots, L1_{m-2}\}$ **do**
      calculate $D_{p,q}$ by Eq. (3)
      calculate $D(p, q)$ by Eq. (5)
      $L2$ append $D(p, q)$
    end
  end
  calculate $ave\_dis(k)$ by Eq. (6)
  L append $ave\_dis(k)$
end
calculate $AICDR(k)$ by Eq. (7)

"all in all". These words are usually phrases, and once separated, they lose their original meaning.

Delete low-frequency words: Based on data quality considerations, low-frequency words have a small amount of text and have a relatively small impact on classification results. Removing low-frequency words reduces the sparsity of data and reduces the consumption of computing resources. Therefore, delete words with a frequency below a certain threshold.

Text vectorization

Salton, Wong & Yang (1975) proposed vector space model (VSM), which converts each text into a certain vector, and the text dataset is transformed into a high-dimensional vector space. In the process of text vectorization, the text is broken down into smaller units, which can be words, phrases, and other semantic units. For a text dataset, we need to construct a vocabulary, where each word in the vocabulary has a unique index. Each text is represented with $t = (w_1, w_2, w_3, \ldots, w_m)$. Here t means a text, w_i means the weight of a word. The text dataset is viewed as a matrix, $D = (t_1, t_2, t_3, \ldots, t_n)$.

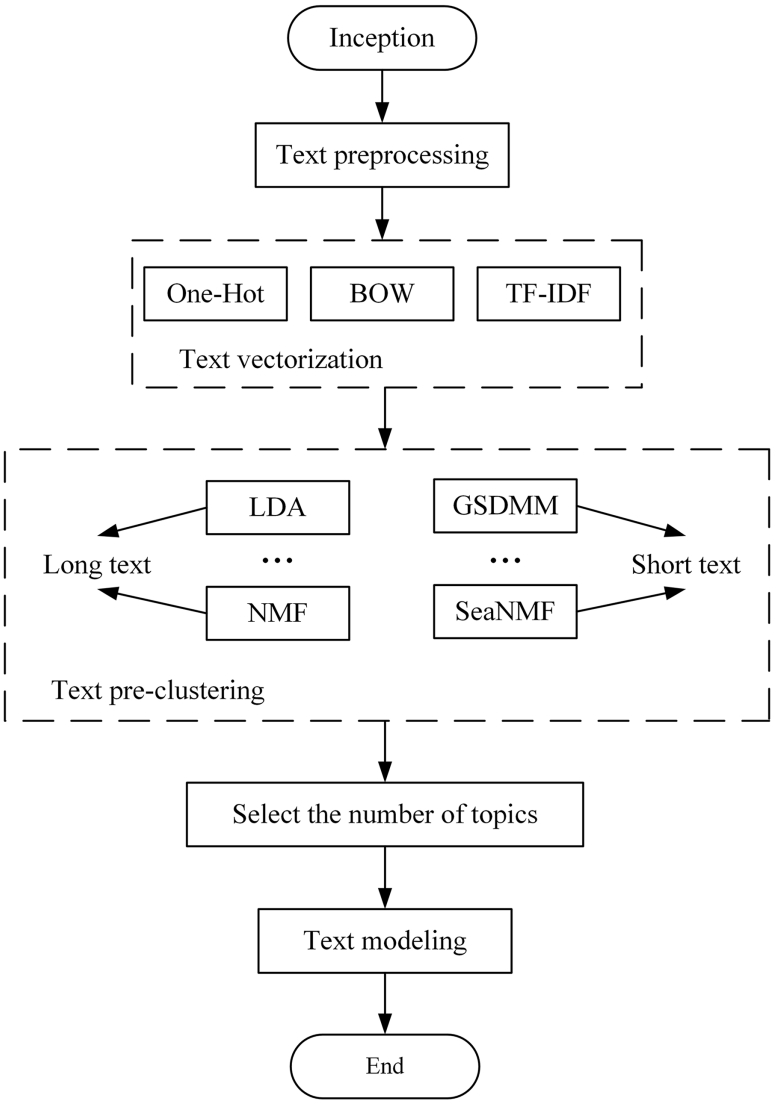

Figure 3 **Selection of the number of topics flow chart.** Full-size DOI: 10.7717/peerj-cs.2723/fig-3

One-Hot

One-Hot weighs a value of 0 or 1 to each word. If there is a word in the text, its weight is 1, otherwise it is 0. Despite the simplicity of the one-hot method, it is hard to capture the importance of words.

BOW

Bag-of-words (BOW) would similarly represent each text as a vector, where each element of the vector corresponds to a word in the vocabulary, and its value is the number of times the word occurs in the text. This method considers word frequency information but ignores word importance and fails to identify false keywords.

TF-IDF

TF-IDF approach evaluates the significance of terms in a document by integrating two metrics: term frequency (TF), which reflects how often a term appears in a document, and

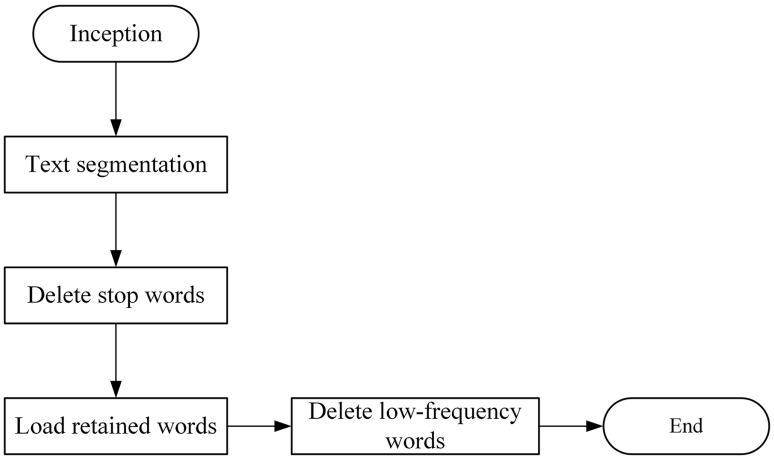

Figure 4 **Text preprocessing steps.** Full-size DOI: 10.7717/peerj-cs.2723/fig-4

inverse document frequency (IDF), which adjusts for the term's rarity across the *corpus*. Compared with BOW, TF-IDF has a strong ability to recognize false high-frequency words and redundant words, it is robust to sparse and unstructured data. Therefore, it is more advantageous in word weight assignment and noise suppression.

Text pre-clustering and selection of topic numbers

AICDR is calculated based on clustering results, and each text in the dataset needs to have a stable cluster label.

We have introduced four different topic models. LDA and NMF are suitable for short texts, while GSDMM and SeaNMF are good for long texts. The four models will be applied to multiple corpora for text clustering, with the resulting classifications utilized for subsequent AICDR calculation. When AICDR reaches its maximum value, the corresponding K is the optimal number of topics. Topic modeling is performed based on the optimal number of topics again.

EXPERIMENTS

In this section, we evaluate the proposed AICDR algorithm with relevant experiments and analyze it in comparison with other methods for selecting the number of topics, respectively. The whole experiment is executed on a CPU of Intel® Core™ i5-8300H.

Experiment datasets

We assess the effectiveness of our proposed approach using a variety of authentic textual data collections, which are detailed subsequently:

- BBCnews (*Greene & Cunningham, 2006*): This dataset contains 2,225 documents from the BBC news site, spanning five categories: business, entertainment, politics, sport, and tech, from 2004 to 2005.
- BBCsport (*Greene & Cunningham, 2006*): The dataset includes 737 sports articles from BBC, covering athletics, cricket, football, rugby, and tennis, from 2004 to 2005.

Table 2 Description of experimental datasets.

Datasets	Clusters	Doc	Word	Balance (%)
BBCnews	5	2,225	8,835	75.54
BBCsport	5	737	3,272	37.74
Reuters	3	1,399	3,440	91.50
AGNews	4	7,409	8,063	96.53
Snippets-1	4	7,870	6,294	56.39
Snippets-2	4	4,469	4,120	24.60

- Reuters (*Joachims, 1998*): The Reuters dataset, sourced from 1,987 financial news. We have selected 1,899 documents across three categories for our analysis.
- AGNews (*Lecun, 2015*): We have filtered 7,409 articles from AG news, categorized as: world, sport, business, and tech.
- Snippets: The dataset is from web search results, divided into Snippets-1 with 7,870 queries in business, computers, arts, and education, and Snippets-2 with 4,496 queries in engineering, health, politics, and sports.

The basic description of the dataset is shown in Table 2.

Comparison methods

Here are several topic models and three methods of selecting the number of topics introduced.

Methods of selecting the number of topics

- Elbow method. The elbow method is commonly used in cluster analysis to determine the optimal number of clusters, particularly suitable for the K-Means clustering algorithm.
- AQDEB (*Shi et al., 2021*). When the SSE curve is quite smooth, it is difficult to identify the inflection point. A quantitative discriminant method of elbow point (referred to as AQDEB) effectively has solved this problem. The index of the minimal inter-section angles between elbow points is used as the estimated potential optimal cluster number.
- Stability analysis (*Greene, O'Callaghan & Cunningham, 2014*). The term-centric stability analysis strategy can efficiently determine the appropriate number of topics while being more applicable to a wider range of topic models and classification methods. It evaluates the consistency between the ranking term lists of topic models generated under different data samples.

We have provided parameter explanations for some methods. To apply these methods, we first need to use various topic models to generate clustering results of text under different K values. Compared to the other two algorithms, elbow method is relatively simple. It evaluates the within-cluster sum of squares (WCSS) of intra-cluster errors under different numbers of clusters, and plots the relationship between K value and WCSS to find the position of the "elbow" in the curve, without the need for additional parameter settings.

For AQDEB, it is an improvement based on the elbow method that does not require additional parameters. For stability analysis, the depth is 10, we only focus on the top 10 ranked words. Then $\tau = 10$, extract only 10 times from each completed dataset. The ratio of the sample dataset to the complete dataset is 0.8. Due to the long computation time of stability analysis, the range of the number of topics is set to 2 to 10, $K \in [2, 10]$. The range of topic numbers for other methods is 2 to 20, $K \in [2, 20]$.

Although the methods based on the elbow method are more suitable for K-Means, they are all based on clustering results for selecting the number of topics. Therefore, we applied the three comparison methods and AICDR to different topic models and K-Means algorithms.

Brief description of topic models

In the previous section, we have introduced four topic models and three text vectorization methods. Therefore, we'll further provide simplified usage descriptions for various models.

LDA and NMF are suitable for short texts, while GSDMM and SeaNMF are good for long texts. Meanwhile, different topic models may use different vectorization methods. LDA and GSDMM adopt BOW, while NMF and SeaNMF adopt TF-IDF. The four models will be applied to multiple corpora for text clustering, with the resulting classifications utilized for subsequent AICDR calculation. Due to the elbow bending method being used as a comparison method, we will also apply the more suitable K-Means for long text classification. K-Means also uses TF-IDF.

Clustering accuracy

The performance of different models may vary on different datasets, such as limitations on text length. To select a more suitable number of topics, we need topic models with better performance. We only consider the clustering accuracy of the algorithm here. The evaluation indicators of topic models usually include topic coherence, topic stability, and topic interpretability (*Lau, Newman & Baldwin, 2014*). AICDR selects the number of topics based on the classification results of the dataset. Therefore, we only need to consider applying the topic model to document classification, and evaluate the clustering accuracy of the model using standardized mutual information (NMI), automatic readability index (ARI), and accuracy (ACC). Table 3 shows the performance of different models on different datasets.

Two methods are used to select the number of topics based on SSE, which can also be applied to centroid-based clustering algorithms, such as K-Means. Thus K-Means is introduced for text classification. Overall, K-Means performs well on long text datasets, especially on BBCsport and Reuters. NMF ranks second, with good performance on BBCNews. LDA performs relatively poorly. In short text datasets, SeaNMF has relatively high clustering accuracy, but the difference between GSDMM and it is not significant. Therefore, both can effectively perform short text clustering. Due to the sparsity inherent in short texts, the performance of K-Means is poor; hence, the clustering of short texts by K-Means is not presented here.

Table 3 The clustering accuracy of different models on the correct number of topics. The best results are highlighted in bold (The higher the better).

Datasets	Topic models	Metrics		
		NMI	ARI	ACC
BBCsport	LDA	0.709	0.662	0.851
	NMF	0.818	0.856	0.872
	K-Means	**0.894**	**0.896**	**0.963**
BBCNews	LDA	0.727	0.701	0.862
	NMF	**0.812**	**0.842**	**0.932**
	K-Means	0.751	0.726	0.868
Reuters	LDA	0.432	0.443	0.669
	NMF	0.550	0.597	0.776
	K-Means	**0.632**	**0.671**	**0.832**
AGNews	GSDMM	**0.585**	**0.622**	**0.833**
	SeaNMF	0.563	0.600	0.822
Snippets-1	GSDMM	0.565	0.592	0.830
	SeaNMF	**0.580**	**0.634**	**0.850**
Snippets-2	GSDMM	0.764	0.823	0.919
	SeaNMF	**0.787**	**0.850**	**0.939**

Table 4 The optimal number of topics is determined by different methods for different topic models. The best results are highlighted in bold.

Datasets	Topic models	Methods			
		Elbow method	AQDEB	Stability analysis	AICDR
BBCsport	LDA	13	13	**5**	**5**
	NMF	3	6	4	4
	K-Means	**5**	13	**5**	**5**
BBCNews	LDA	3	17	**5**	6
	NMF	3	19	2	**5**
	K-Means	3	16	6	6
Reuters	LDA	**3**	12	**3**	**3**
	NMF	**3**	7	4	**3**
	K-Means	**3**	13	**3**	**3**
AGNews	GSDMM	3	19	3	**4**
	SeaNMF	3	16	3	5
Snippets-1	GSDMM	**4**	12	**4**	**4**
	SeaNMF	18	**4**	**4**	**4**
Snippets-2	GSDMM	3	19	3	3
	SeaNMF	3	11	3	3

Results analysis

To assess the precision in determining the number of topics and the adaptability to diverse topic models, we compared AICDR with three other methods for establishing the optimal K across various datasets and topic models, as detailed in Table 4.

The experimental results highlight the effectiveness of determining the number of topics based on the AICDR method. Even on Snippets-1 and Reuters, AICDR based on different topic models can determine the optimal number of topics. For BBCsport, BBCNews, and AG-News, AICDR can effectively determine the number of topics in some topic models. Although it does not determine the optimal number of topics on other models, the difference from the optimal number of topics is not significant. Although the number of topics it determines on other models is not optimal, the difference from the optimal number of topics is generally minimal, the difference is usually 1. Referring to Table 3, BBCsport, BBCNews, and AG News showed better clustering accuracy when applied with K-Means, NMF, and GSDMM, respectively. Thus, the better the performance of the topic model (clustering accuracy), the higher the precision of the AICDR.

Elbow method only shows accurate performance on Reuters, with relatively poor performance on other datasets. However, the number of topics it determines is also close to the number of clusters in other datasets. Meanwhile, it primarily judges the elbow point of the SSE curve, this method can accurately determine the number of topics when applied with K-Means on both BBCsport and Reuters datasets. Elbow method is indeed more suitable for selecting K in K-Means.

AQDEB does not exhibit particularly outstanding performance across all datasets. It employs the arccosine theorem to compute the inter section angles between elbow points, the index of minimal inter section angles between elbow points is used as the estimated potential optimal cluster number. The mean distortion curves indicate that the index does not decrease completely with the increase of K, *i.e.*, the curves are not fairly smooth at lower K values. This also explains why the number of topics determined by AQDEB is more concentrated at the lower K values.

The number of topics determined based on Stability analysis is relatively accurate. This method does indeed show a certain level of adaptability and can effectively be used to determine the optimal number of topics for some models. When the determined number of topics is not optimal, it can still approximate the original number of clusters in the dataset, which can also be maintained at the same level as AICDR. But overall, its accuracy is slightly lower than AICDR.

As shown from Figs. 5 to 8, the variation of different indicators with K is presented. As the number of topics increases, AICDR exhibits a fluctuating decrease, akin to undulating peaks and troughs, this indicates that the inter-class distance also decreases as K increases The corresponding AICDR for The neighbor of the optimal K is also relatively high. For elbow method, the variation of SSE with K is presented and the elbow point (the optimal number of topics) is marked. AQDEB is an improved method for determining the elbow point. It is obvious that the SSE curve does not decrease completely smoothly at lower K values, so the optimal number of themes judged by AQDEB is mostly concentrated at

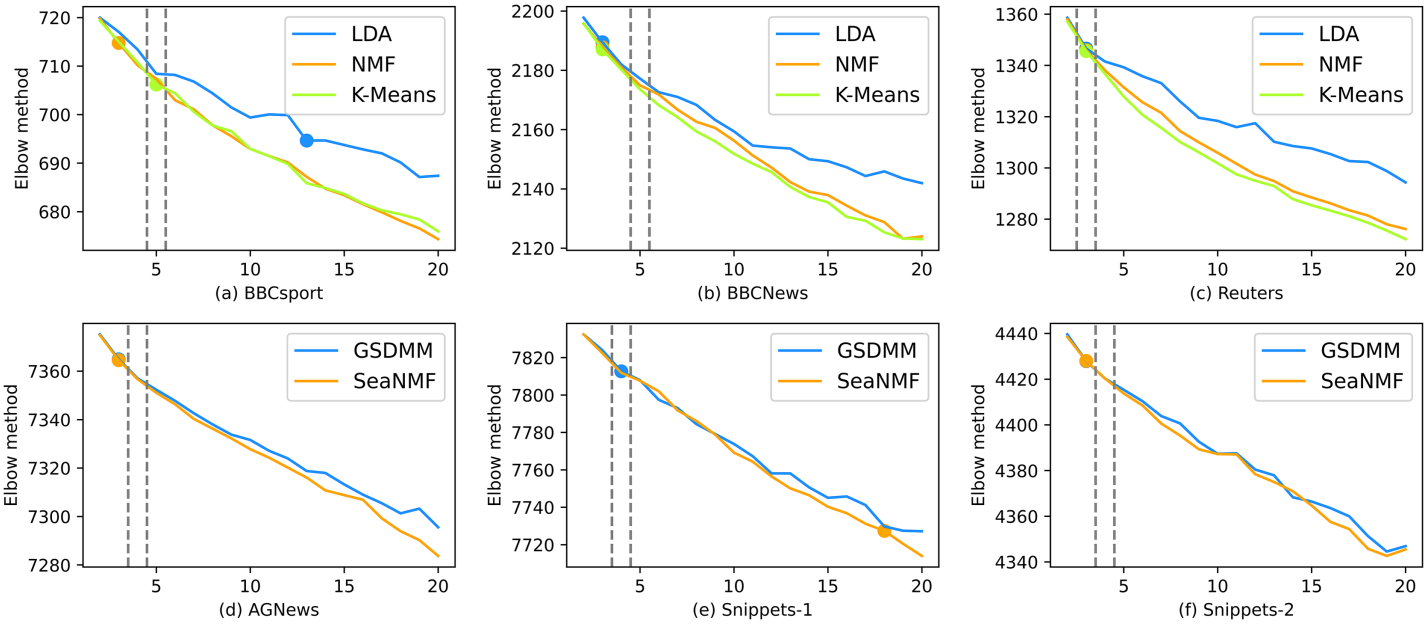

Figure 5 **The performance of the elbow method.** Full-size DOI: 10.7717/peerj-cs.2723/fig-5

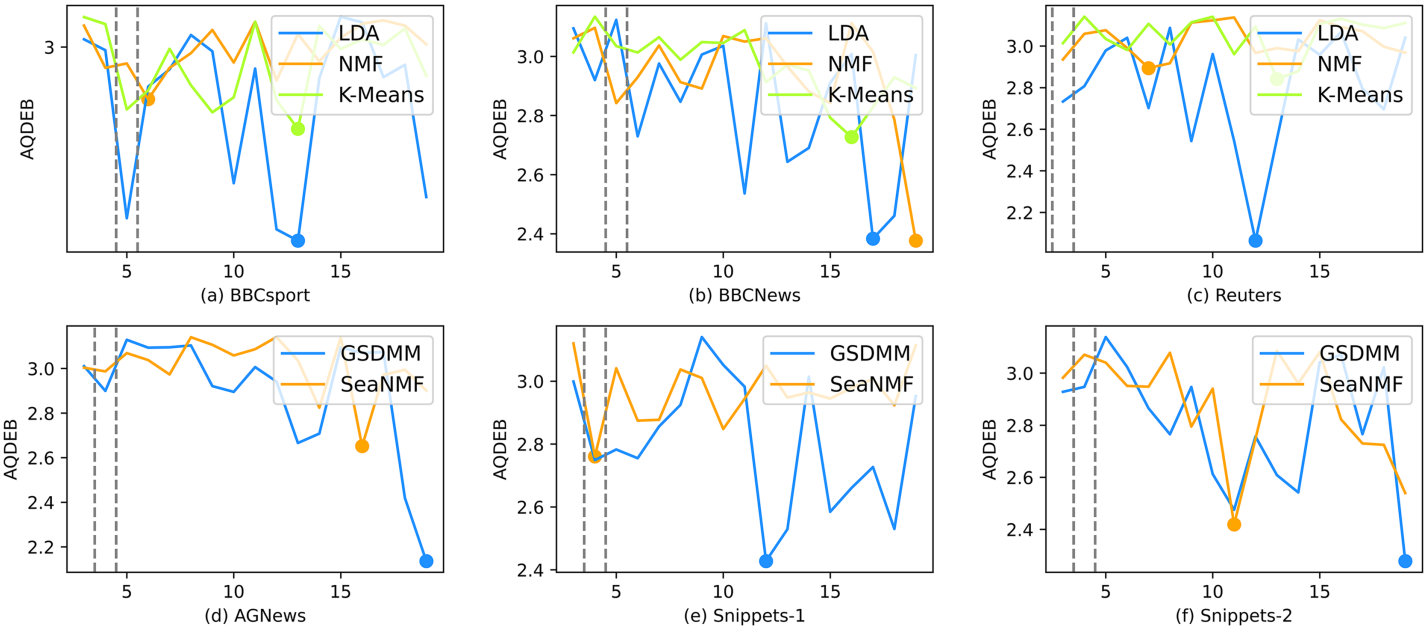

Figure 6 **The performance of the AQDEB.** Full-size DOI: 10.7717/peerj-cs.2723/fig-6

lower K values. Stability analysis is similar to AICDR in that its values are relatively high around the correct number of topics. This further demonstrates that stability analysis and AICDR indeed have excellent ability to identify the optimal number of topics, as well as strong adaptability to various models.

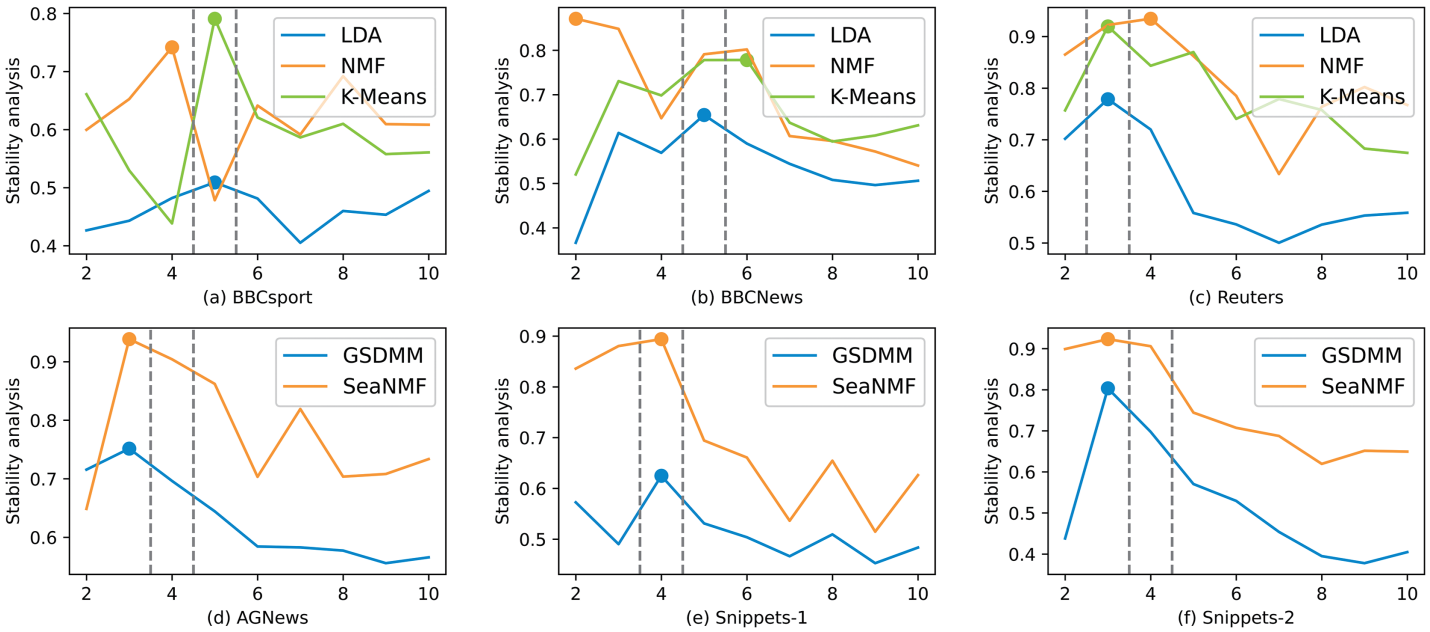

Figure 7 **The performance of the stability analysis.** Full-size DOI: 10.7717/peerj-cs.2723/fig-7

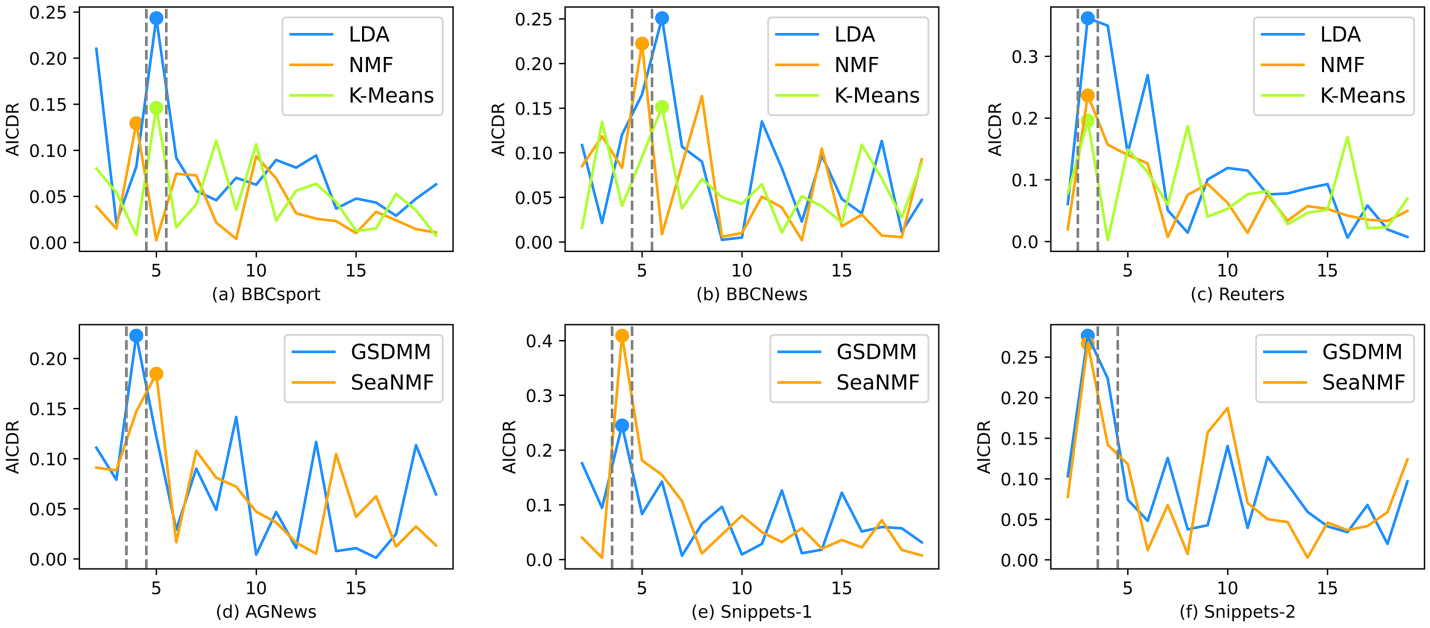

Figure 8 **The performance of the AICDR.** Full-size DOI: 10.7717/peerj-cs.2723/fig-8

Overall, AICDR can identify the optimal number of topics across all datasets by applying different topic models or clustering algorithms. Mainly, it is not limited by topic models or clustering algorithms and has good adaptability to most methods. Notably, the

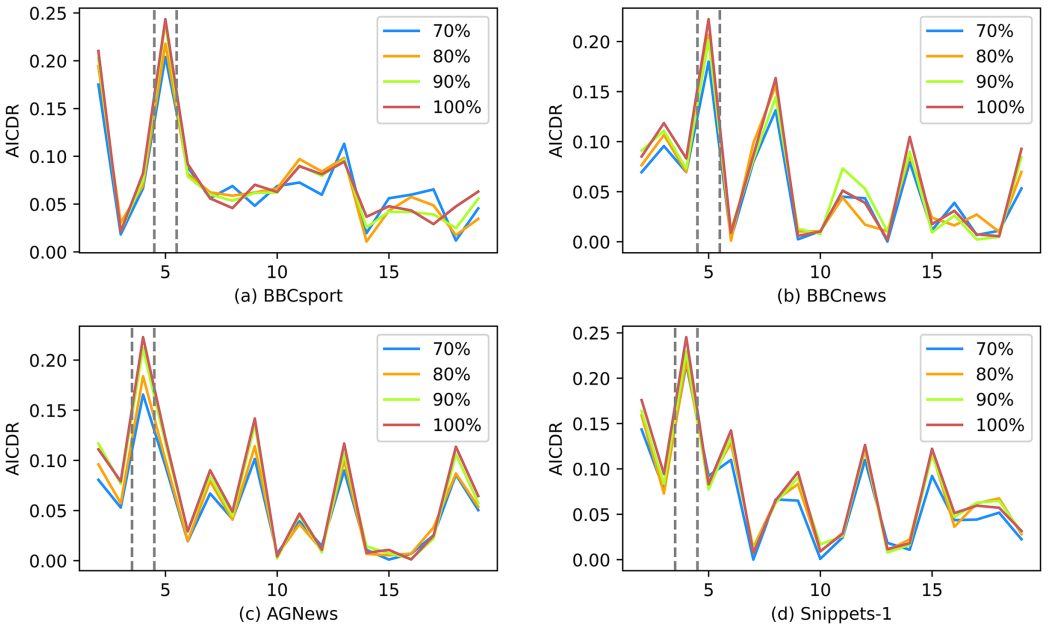

Figure 9 **The performance of AICDR on different subsets.**
Full-size DOI: 10.7717/peerj-cs.2723/fig-9

higher the clustering accuracy of the model, the stronger the identification capability of AICDR.

Stability analysis

Stability analysis is most commonly conducted by perturbing the data, which involves generating sub-samples through random sampling of the original objects. By doing so, we can assess the consistency and reliability of AICDR across different subsets of the data, ensuring its robustness and effectiveness in various scenarios. Here, subsets of the dataset are taken at 70%, 80%, and 90% respectively. During the sampling process, no random seeds are set to ensure the randomness of the subset.

Figure 9 shows the performance of AICDR on subsets of four datasets. From the figure, it can be observed that the AICDR curve exhibits high consistency across different subsets and the entire dataset. Although there are slight fluctuations within certain K value ranges, overall, the number of topics determined by AICDR on different subsets is completely accurate. This indicates that AICDR has excellent robustness and stability.

Preprocessing and vectorization in AICDR

In this section, we use different text vectorization methods and set different low-frequency words to explore the influencing factors of AICDR. The effects of low-frequency words and text vectorization on AICDR are shown in Tables 5 and 6.

In Table 5, different vectorization methods have a weak impact on AICDR, while One-Hot, BOW, and TF-IDF have no significant effect on AICDR overall. However, when the classification model is K-Means, One-Hot, and BOW reduces the performance of AICDR, and the number of determined topics deviates greatly from the original number of topics in

Table 5 The impact of low-frequency words on the performance of AICDR. The best results are highlighted in bold.

Datasets	Topic models	Methods		
		One-Hot	BOW	TF-IDF
BBCsport	LDA	**5**	**5**	**5**
	NMF	4	4	4
	K-Means	**5**	2	**5**
BBCNews	LDA	6	6	6
	NMF	**5**	**5**	**5**
	K-Means	3	3	6
Reuters	LDA	**3**	**3**	**3**
	NMF	**3**	**3**	**3**
	K-Means	**3**	2	**3**
AGNews	GSDMM	**4**	**4**	**4**
	SeaNMF	5	5	5
Snippets-1	GSDMM	**4**	**4**	**4**
	SeaNMF	**4**	**4**	**4**
Snippets-2	GSDMM	**3**	**3**	**3**
	SeaNMF	**3**	**3**	**3**

Table 6 The impact of text vectorization on the performance of AICDR. The best results are highlighted in bold.

Datasets	Topic models	Number of words		
		5	10	15
BBCsport	LDA	**5**	**5**	**5**
	NMF	4	4	4
	K-Means	**5**	**5**	**5**
BBCNews	LDA	6	6	6
	NMF	**5**	**5**	**5**
	K-Means	6	6	6
Reuters	LDA	**3**	**3**	**3**
	NMF	**3**	**3**	**3**
	K-Means	**3**	**3**	**3**
AGNews	GSDMM	**4**	**4**	**4**
	SeaNMF	5	5	5
Snippets-1	GSDMM	**4**	**4**	**4**
	SeaNMF	**4**	**4**	**4**
Snippets-2	GSDMM	**3**	**3**	**3**
	SeaNMF	**3**	**3**	**3**

the *corpus*. This influence appears in the long text *corpus*. When the vectorization method is TF-IDF, AICDR exhibits the best performance regardless of the topic model. Therefore, TF-IDF and AICDR are more compatible. In Table 6, low-frequency words have little

effect on AICDR. We set the thresholds for low-frequency words to 5, 10, and 15, while the number of topics determined by AICDR remained consistent. Therefore, low-frequency words have no significant impact on AICDR.

Overall, while preprocessing techniques and vectorization methods generally have limited impact on AICDR, the TF-IDF vectorization method demonstrates greater compatibility with AICDR.

CONCLUSION

A key challenge when applying topic modeling is the selection of an appropriate number of topics K. In this article, we propose a selection method for determining the number of topics based on inter-class distance, named average inter-class distance change rate (AICDR). By calculating the AICDR for consecutive K values, the optimal number of topics is selected as the previous K value when the difference between the two is maximized. The optimal clustering result, with a specified number of topics, should exhibit high intra-class similarity and low inter-class similarity. AICDR considers the inter-class distance comprehensively, which is computed by the diameter of the class, and avoids topic overlap, improves intra-class similarity, and reduces inter-class similarity. Meanwhile, it is not limited by topic models or clustering algorithms and can effectively determine the number of topics in most methods. Evaluations on several real text datasets have suggested that AICDR can provide a useful guide for selecting the optimal number of topics.

In upcoming research endeavors, we mainly focus on improving the robustness and stability of AICDR to noise. Meanwhile, we will explore other potential limitations of AICDR and alleviate these limitations to improve the performance of AICDR.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The work was supported by National Natural Science Foundation of China (NSFC) No. 72401080. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
National Natural Science Foundation of China (NSFC): 72401080.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Yang Xu conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

- Yueyi Zhang analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Yefang Sun analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Hanting Zhou conceived and designed the experiments, performed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The code is available at GitHub and Zenodo:

- https://github.com/cjluxy/AICDR-Selecting-the-number-of-topics.

- cjluxy. (2025). cjluxy/AICDR-Selecting-the-number-of-topics: Python code and data for AICDR (AICDR-v1). Zenodo. https://doi.org/10.5281/zenodo.14784034.

The BBCnews dataset and BBCsport dataset are available at http://mlg.ucd.ie/datasets/bbc.html.

The Reuters dataset is available at https://martin-thoma.com/nlp-reuters.

The AGNews dataset is available at Kaggle: https://www.kaggle.com/datasets/amananandrai/ag-news-classification-dataset.

The Snippets1 dataset is available at Zenodo: Xu, Y. (2025). Snippets1 [Data set]. Zenodo. https://doi.org/10.5281/zenodo.14822881.

The Snippets2 dataset is available at Zenodo: Xu, Y. (2025). Snippets2 [Data set]. Zenodo. https://doi.org/10.5281/zenodo.14822895.

The data is available in the Supplemental File.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.2723#supplemental-information.

REFERENCES

Agarwal N, Sikka G, Awasthi LK. 2024. Integrating semantic similarity with Dirichlet multinomial mixture model for enhanced web service clustering. *Knowledge and Information Systems* **66(4)**:2327–2353 DOI 10.1007/s10115-023-02034-x.

Aghdam MH, Zanjani MD. 2021. A novel regularized asymmetric non-negative matrix factorization for text clustering. *Information Processing & Management* **58(6)**:102694 DOI 10.1016/j.ipm.2021.102694.

Altarturi HHM, Saadoon M, Anuar NB. 2023. Web content topic modeling using LDA and HTML tags. *Peerj Computer Science* **9(6)**:e1459 DOI 10.7717/peerj-cs.1459.

Arbelaitz O, Gurrutxaga I, Muguerza J, Pérez JM, Perona I. 2013. An extensive comparative study of cluster validity indices. *Pattern Recognition: The Journal of the Pattern Recognition Society* **46(1)**:243–256 DOI 10.1016/j.patcog.2012.07.021.

Athukorala S, Mohotti W. 2022. An effective short-text topic modelling with neighbourhood assistance-driven NMF in Twitter. *Social Network Analysis and Mining* **12(1)**:89 DOI 10.1007/s13278-022-00898-5.

Blei DM, Ng A, Jordan MI. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research* 993–1022 DOI 10.1162/jmlr.2003.3.4-5.993.

Cai D, He X, Han J, Huang TS. 2011. Graph regularized nonnegative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33(8)**:1548–1560 DOI 10.1109/TPAMI.2010.231.

Cao J, Xia T, Li J, Zhang Y, Tang S. 2009. A density-based method for adaptive LDA model selection. *Neurocomputing* **72(7–9)**:1775–1781 DOI 10.1016/j.neucom.2008.06.011.

Carbonetto P, Sarkar A, Wang Z, Stephens M. 2022. Non-negative matrix factorization algorithms greatly improve topic model fits. ArXiv preprint DOI 10.48550/arXiv.2105.13440.

Chen P, Guo W, Wang Q, Song Y. 2017. Topic classification based on distributed document representation and latent topic information. In: *9th Annual Summit and Conference of the Asia-Pacific-Signal-and-Information-Processing-Association (APSIPA ASC)*. Piscataway: IEEE, 614–617.

Cheng X, Yan X, Lan Y, Guo J. 2014. BTM: topic modeling over short texts. *IEEE Transactions on Knowledge & Data Engineering* **26(12)**:2928–2941 DOI 10.1109/TKDE.2014.2313872.

Ding F, Kang X, Ren F. 2024. Neuro or symbolic? fine-tuned transformer with unsupervised LDA topic clustering for text sentiment analysis. *IEEE Transactions on Affective Computing* **15(2)**:493–507 DOI 10.1109/TAFFC.2023.3279318.

Ding J, Tarokh V, Yang Y. 2017. Bridging AIC and BIC: a new criterion for autoregression. *IEEE Transactions on Information Theory* **64(6)**:4024–4043 DOI 10.1109/TIT.2017.2717599.

Gan J, Qi Y. 2021. Selection of the optimal number of topics for LDA topic model-taking patent policy analysis as an example. *Entropy* **23(10)**:1301 DOI 10.3390/e23101301.

Greene D, Cunningham P. 2006. Practical solutions to the problem of diagonal dominance in kernel document clustering. In: *International Conference on Machine Learning* DOI 10.1145/1143844.1143892.

Greene D, O'Callaghan D, Cunningham P. 2014. *How many topics? stability analysis for topic models*. Berlin, Heidelberg: Springer.

Gu Q, Zhou J. 2009. Co-clustering on manifolds. In: *15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. USA: ACM, 359–367.

Gupta A, Zhang Z. 2023. Neural topic modeling via discrete variational inference. *ACM Transactions on Intelligent Systems and Technology* **14(2)**:1–33 DOI 10.1145/3570509.

He J, Chen X, Du M. 2015. Topic evolution analysis based on improved online LDA model. *Journal of Central South University (Science and Technology)* **46**:547–553 DOI 10.11817/j.issn.1672-7207.2015.02.024.

Huang L, Ma J, Chen C. 2017. Topic detection from microblogs using T-LDA and perplexity. In: *24th Asia-Pacific Software Engineering Conference (APSEC)*. Piscataway: IEEE, 71–77.

Ignatenko V, Koltcov S, Staab S, Boukhers Z. 2018. Fractal approach for determining the optimal number of topics in the field of topic modeling. In: *3rd International Conference on Computer Simulation in Physics and Beyond (CSP) 1163*.

Joachims T. 1998. Text categorization with support vector machines: learning with many relevant features. In: *Conference on Machine Learning* DOI 10.1007/BFb0026683.

Kekere T, Marivate V, Hattingh M. 2023. Exploring COVID-19 public perceptions in South Africa through sentiment analysis and topic modelling of Twitter posts. *The African Journal of Information and Communication* **31(31)**:1–27 DOI 10.23962/ajic.i31.14834.

Kherwa P, Bansal P. 2018. Topic modeling: a comprehensive review. *ICST Transactions on Scalable Information Systems* **7**:159623 DOI 10.4108/eai.13-7-2018.159623.

Koltcov S, Surkov A, Filippov V, Ignatenko V. 2024. Topic models with elements of neural networks: investigation of stability, coherence, and determining the optimal number of topics. *Peerj Computer Science* **10(4)**:e1758 DOI 10.7717/peerj-cs.1758.

Lau JH, Newman D, Baldwin T. 2014. Machine reading tea leaves: automatically evaluating topic coherence and topic model quality. In: *Conference of the European Chapter of the Association for Computational Linguistics*.

Lecun XZJZY. 2015. Character-level convolutional networks for text classification. In: *International Conference on Neural Information Processing Systems*.

Levy O, Goldberg Y. 2014. Neural word embedding as implicit matrix factorization. In: *28th Conference on Neural Information Processing Systems (NIPS)*.

Li M, Wang X, Li C, Zeng A. 2024. Nonnegative matrix factorization with Wasserstein metric-based regularization for enhanced text embedding. *PLOS ONE* **19(12)**:e0314762 DOI 10.1371/journal.pone.0314762.

Liu F, Deng Y. 2021. Determine the number of unknown targets in open world based on elbow method. *IEEE Transactions on Fuzzy Systems* **29(5)**:986–995 DOI 10.1109/tfuzz.2020.2966182.

Lu F, Shen B, Lin J, Zhang H. 2013. A method of SNS topic models extraction based on self-adaptively LDA modeling. In: *3rd International Conference on Intelligent System Design and Engineering Applications (ISDEA)*. Piscataway: IEEE, 112–115.

Miao Y, Grefenstette E, Blunsom P. 2017. Discovering discrete latent topics with neural variational inference. *International Conference on Machine Learning* 2410–2419 DOI 10.48550/arXiv.1706.00359.

Miao Y, Yu L, Blunsom P. 2015. Neural variational inference for text processing. *Computer Science* 1791–1799 DOI 10.48550/arXiv.1511.06038.

Murtagh F, Legendre P. 2014. Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *Journal of Classification* **31**:274–295 DOI 10.1007/s00357-014-9161-z.

O'Callaghan D, Greene D, Carthy J, Cunningham P. 2015. An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications* **42**:5645–5657 DOI 10.1016/j.eswa.2015.02.055.

Papadimitriou CH, Raghavan P, Tamaki H, Vempala S. 1998. Latent semantic indexing: a probabilistic analysis. *Journal of Computer and System Sciences* **61**:217–235 DOI 10.1006/jcss.2000.1711.

Peng G, Yuefen W. 2016. Identifying optimal topic numbers from sci-tech information with LDA model. *Data Analysis and Knowledge Discovery* **32**:42–50.

Peng M, Zhang Z. 2024. Research on the hot spots and theme evolution of artificial intelligence education policy in China. In: *International Conference on Informatics Education and Computer Technology Applications (IECA)*, 160–164.

Qiang J, Qian Z, Li Y, Yuan Y, Wu X. 2020. Short text topic modeling techniques, applications, and performance: a survey. *IEEE Transactions on Knowledge and Data Engineering* **34**:1427–1445 DOI 10.1109/TKDE.2020.2992485.

Rodriguez MZ, Comin CH, Casanova D, Bruno OM, Amancio DR, Costa LdF, Rodrigues FA. 2019. Clustering algorithms: a comparative approach. *PLOS ONE* **14**:e0210236 DOI 10.1371/journal.pone.0210236.

Salah A, Ailem M, Nadif M, Aaai. 2018. Word co-occurrence regularized non-negative matrix tri-factorization for text data co-clustering. In: *32nd AAAI Conference on Artificial Intelligence/30th Innovative Applications of Artificial Intelligence Conference/8th AAAI Symposium on Educational Advances in Artificial Intelligence*, 3992–3999.

Salton G, Wong A, Yang CS. 1975. A vector space model for automatic indexing. *Communications of the ACM* **18(11)**:613–620 DOI 10.1145/361219.361220.

Shi T, Kang K, Choo J, Reddy CK. 2018. Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In: *27th World Wide Web (WWW) Conference 1105-1114*. ACM, 1105–1114.

Shi C, Wei B, Wei S, Wang W, Liu H, Liu J. 2021. A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *EURASIP Journal on Wireless Communications and Networking* 1–16 DOI 10.1186/s13638-021-01910-w.

Srivastava A, Sutton C. 2017. Autoencoding variational inference for topic models. ArXiv preprint DOI 10.48550/arXiv.1703.01488.

Tang YK, Huang H, Shi X, Mao XL. 2022. Neural variational gaussian mixture topic model. *ACM Transactions on Asian and Low-Resource Language Information Processing* **22(11)**:1–18 DOI 10.1145/3629518.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. 2017. Attention is all you need. *NIPS* DOI 10.48550/arXiv.1706.03762.

Wang H, Wang J, Zhang Y, Wang M, Mao C. 2019. Optimization of topic recognition model for news texts based on LDA. *Journal of Digital Information Management* **17(5)**:257–269.

Wang X, Yang Y. 2020. Neural topic model with attention for supervised learning. In: *23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Xu W, Liu X, Gong Y. 2003. Document clustering based on non-negative matrix factorization. In: *ACM SIGIR FORUM*, 267–273.

Yin J, Wang J. 2014. A dirichlet multinomial mixture model-based approach for short text clustering. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 233–242 DOI 10.1145/2623330.2623715.

Zhao W, Chen JJ, Perkins R, Liu Z, Ge W, Ding Y, Zou W. 2015. A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC Bioinformatics* **16(S13)**:391 DOI 10.1186/1471-2105-16-S13-S8.

Zhao H, Du L, Buntine W, Zhou M. 2018. Dirichlet belief networks for topic structure learning. In: *32nd Conference on Neural Information Processing Systems (NIPS)*.

Zheng M, Jiang K, Xu R, Qi L. 2023. An adaptive LDA optimal topic number selection method in news topic identification. *IEEE Access* **11**:92273–92284 DOI 10.1109/ACCESS.2023.3308520.

Zuo Y, Li C, Lin H, Wu J. 2021. Topic modeling of short texts: a pseudo-document view with word embedding enhancement. *IEEE Transactions on Knowledge and Data Engineering* **35**:972–985 DOI 10.1109/TKDE.2021.3073195.

Zuo Y, Zhao J, Xu K. 2016. Word network topic model: a simple but general solution for short and imbalanced texts. *Knowledge & Information Systems* **48(2)**:379–398 DOI 10.1007/s10115-015-0882-z.