

# A data-driven analysis to determine the optimal number of topics 'K' for latent Dirichlet allocation model

Astha Goyal, Indu Kashyap

Department of Computer Science and Engineering (CSE), SET, MRIIRS, Faridabad, India

---

## Article Info

### Article history:

Received Nov 11, 2023

Revised Feb 29, 2024

Accepted Mar 17, 2024

---

### Keywords:

Grid search approach  
Latent dirichlet allocation  
LDA hyperparameter tuning  
Objective functions  
Optimal number of topics  
Topic modeling

---

## ABSTRACT

Topic modeling is an unsupervised machine learning technique successfully used to classify and retrieve textual data. However, the performance of topic models is sensitive to selecting optimal hyperparameters, the number of topics 'K' and Dirichlet priors ' $\alpha$ ' and ' $\beta$ .' This data-driven analysis aims to determine the optimum number of topics, 'K,' within the latent Dirichlet allocation (LDA) model. This work utilizes three datasets, namely 20-Newsgroups news articles, Wikipedia articles, and Web of Science containing science articles, to assess and compare various 'K' values through the grid search approach. The grid search approach finds the best combination of hyperparameter values by trying all possible combinations to see which performs best. This research seeks to identify the 'K' that optimizes topic relevance, coherence, and model performance by leveraging statistical metrics, such as coherence scores, perplexity, and topic distribution quality. Through empirical analysis and rigorous evaluation, this work provides valuable insights for determining the ideal 'K' for LDA models.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



---

## Corresponding Author:

Astha Goyal

Department of Computer Science and Engineering (CSE), SET, MRIIRS

Faridabad, India

Email: [astha.gmca84@gmail.com](mailto:astha.gmca84@gmail.com)

---

## 1. INTRODUCTION

The digitization trend is ongoing, and more data is being collected digitally every day. The transformation of this enormous amount of unstructured data into a structured form to extract relevant information is a significant issue in text mining. Text-mining techniques, such as topic modeling, are applied to determine the themes or topics of unstructured data. It is an unsupervised machine learning method that aids in categorizing documents according to predetermined subjects [1]. Several versions of topic modeling algorithms called topic models have evolved. Latent semantic analysis (LSA) [2], latent Dirichlet allocation (LDA) [3], probabilistic latent semantic analysis (PLSA) [4], and non-negative matrix factorization (NMF) [5] are some of the most commonly used topic models. Of all topic models, LDA is the most popular topic modeling algorithm.

LDA is a generative algorithm in which a document is viewed as a distribution over topics, whereas a topic is considered as a distribution over words. These distributions reveal the underlying themes of the document collection [6]. LDA iteratively optimizes the Dirichlet priors ' $\alpha$ ' and ' $\beta$ ' over the topics' multinomial distributions and the document's words to identify the topic. The model discovers the topic by determining the correlation between words but cannot capture the correlation between topics. In the real world, topic correlations are expected, which limits LDA's ability to analyze large-scale real-world data. LDA has been further enhanced to achieve better topic discovery, and different types of correlation models such as the correlated topic model (CTM) [7], pachinko allocation model (PAM) [8], hierarchical LDA (HLDA) [9],

dynamic topic model (DTM) [10], and author topic model (ATM) [11] have been developed. These models can all describe the topic correlations, but none can decide on the optimal number of topics, 'K.' When employing any of the topic models, it is crucial to determine the optimal number of topics extracted by the model because this significantly affects the efficacy of the results. There is no easy method to choose the optimal number of topics, and no standard procedures have been established [12].

One approach [13] for selecting the number of topics is iterative. The approach involves starting with a small number of topics and then increasing the number of topics until the model's performance no longer improves. However, this approach is time-consuming and computationally expensive. Another popular approach [14] across the literature to choose the number of topics is to use a variety of statistical metrics, also known as objective functions, to evaluate the model's performance. These metrics can include the perplexity of the model, the coherence of the topics, and the interpretability of the topics analyzed in terms of stability or divergence. However, there is no consensus on which metrics are most effective for choosing the number of topics. The choice of the number of topics in a topic model is subjective. There is no single "correct" number of topics, and the best number of topics varies depending on the specific application. Therefore, the main contributions of this research to the LDA topic model are:

- Investigate the significance of various evaluation metrics in discerning the optimal value of 'K.'
- Utilization of a composite of evaluation metrics instead of a singular metric for the optimal 'K.'
- Validation of the effectiveness of derived 'K' across datasets where the number of topics is known.
- Determination of the optimal value of 'K' for a dataset in which the number of topics is unknown.

## 2. METHOD

The performance of LDA depends not only on the quality and representative nature of the selected dataset but also on the values of specific parameters chosen during initialization to control the learning process. Such manually calibrated parameters are termed as hyperparameters [15], and their optimal values influence the learning capability of any model, which directly impacts the performance. For the LDA algorithm, the number of topics (K), the Dirichlet priors for the document-topic distribution ( $\alpha$ ), and the word-topic distribution ( $\beta$ ) are important hyperparameters. The LDA algorithm is sensitive to the choice of K [16]. A small value of K may produce overly generic topics that are highly overlapping in terms of concepts. For instance, in a large dataset of articles related to networking, information security, and operating systems, using small values of K may yield over-general topics about themes.

On the other hand, a large value of K can produce sparse non-interpretable topics. For instance, using extremely large values of K in the dataset referred to before may yield topics where the word distribution is sparsely populated, and the density assigned per word becomes extremely small to distribute the density across many topics. Therefore, identifying the optimal 'K' is crucial. This research work proposes a grid-search approach [17], [18], over the topic 'K'. A set of values for the number of topics 'K' are taken equidistant from one another within a proposed range. The process for determining the optimal 'K' amongst the proposed values is executed as shown in Algorithm 1.

Algorithm 1. Algorithm for proposed methodology

**Defining Variables:**

- K** = number of topics
  - n<sub>K</sub>** = number of varying values of K chosen for training the models
  - metricDict** = {**Perplexity**: Minimize, **Average Cosine Distance**: Minimize, **Symmetric KL-Divergence**: Minimize, **C<sub>UMass</sub>** : Maximize, **C<sub>UCI</sub>** : Maximize, **C<sub>NPMI</sub>** : Maximize, **C<sub>V</sub>**: Maximize, **C<sub>v</sub>**: Maximize, **Topic Superiority**: Maximize}
1. For each metric **metricDict**.
    - a. for **K** in the range from 1 to **n<sub>K</sub>**
      - i. Compute the metric value for **K**.
  2. Plot a graph between different **K** values and their metric values for each metric in **metricDict**.
  3. For each metric in **metricDict**
    - a. Using a grid-search strategy, select a single value of K for which most metrics agree based on whether the metric should be maximized or minimized based on corresponding values to the metric in **metricDict**.
  4. The K obtained from the above method gives the **optimal K**.

### 2.1. Generate LDA models

Using online variational inference over the selected corpus, generate LDA models with varying 'K' values. The different LDA models are trained over the same dataset. This step produces  $n_K$  LDA models, where  $n_K$  is the number of varying K values chosen for training the models. For better results, K values may

be selected from a range proportional to the number of documents in the corpus, with initially a larger difference between K values for two models out of the set of trained models. Gradually, this difference can be reduced to derive a finer value for the hyperparameter. The LDA is an unsupervised topic model based on generative probabilistic modeling [19], [20]. It uses the term frequency to determine the probability of document-topic association [21]. LDA is based on the intuition that a document is composed of multiple topics, and each topic is effectively a distribution of words from a fixed vocabulary. The structural diagram of LDA is shown in Figure 1, known as the plate model of LDA.

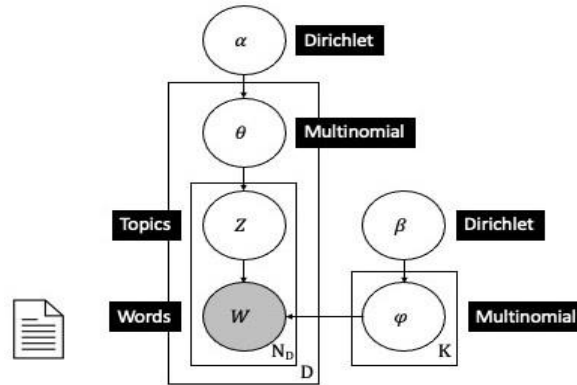


Figure 1. Plate model of LDA [6]

The variable names in the Figure 1 can be defined as:

$D$ : Number of documents

$N$ : Number of words in the document (document  $x$  has  $N_x$  words)

$K$ : Number of latent topics

$\alpha$ : Dirichlet prior for the per-document topic distribution

$\beta$ : Dirichlet prior for the per topic word distribution

$\theta_x$ : Vector of topic distribution over document  $x$

$\phi_k$ : Word distribution for topic  $k$

$z_{xy}$ : Topic for the  $y$  word in document  $x$

$w_{xy}$ : The specific word

As the topics are not known a priori, LDA utilizes a latent variable model to deduce the distribution parameters of these latent variables by employing a posterior probability inference based on the observed terms and documents [22]. LDA evaluates the joint probability distribution between words and topics in the given corpus [23]. The probability of a word belonging to one of the  $K$  topics can be computed using (1).

$$P(W, Z, \theta, \phi, \alpha, \beta) = \prod_{j=1}^M P(\theta_j; \alpha) \prod_{i=1}^K P(\phi_i; \beta) \prod_{t=1}^N P(Z_{j,t} | \theta_j) P(W_{j,t} | \phi_{Z_{j,t}}) \quad (1)$$

The inference of probabilities follows a generative process. For each word  $N_x$  in document  $x$ :

- Choose a topic  $z_{xy} \sim \text{Multinomial}(\theta_x)$
- Choose a word  $w_{xy} \sim \text{Multinomial}(\phi_{z_{xy}})$

where the multinomial parameters for topics in a document  $\theta_x$  and words in a topic  $\phi_k$  have Dirichlet priors  $\alpha$  and  $\beta$ . Posterior probabilities for these distributions are learned by the expectation-maximization (EM) algorithm, which finds maximum posteriori (MAP) estimates of parameters along with variational inference (VI) methods that allow for online LDA learning [24].

LDA excels in simultaneous inference and handling documents of arbitrary sizes [25], making it a leading choice in diverse fields [26], including topic classification, clustering, short-text analysis, summarization, literature review, and sentiment analysis. A few of its applications include event extraction from Twitter [27], summarising Twitter posts based on selected topics [28], query intent recognition [29],

searching and classifying topics in a text corpus [30], improving document classification using domain-specific vocabulary [31], and customer opinion mining using Twitter topic modeling and logistic regression [32]. While applicable to a large corpus of documents, LDA makes some rigid assumptions regarding a corpus, suggesting areas for improvisation. Like its predecessors [33], LDA assumes a bag-of-words model for a document, which may not be applicable in all situations. Further, no correlation between topics is made, and the order of documents is also not considered. Also, the constraint over the number of topics, which must be specified a priori, may be unsuitable for different corpora, as the obtained results are sensitive to the choice of  $K$ , and non-optimal values hinder model performance. Other considerations include a dynamic inference [34] from documents or other accompanying information. These constraints are examined and partially relaxed in various variants of LDA developed in the last two decades; however, these improvised models still depend on the optimal choices for involved hyperparameters.

**2.2. Evaluate generated LDA models on metric set**

Choosing a value of  $K$  that generates the best results for a selected corpus is crucial for the optimal performance of any topic model [35]. The quality of topics generated by topic models can be assessed using evaluation metrics defined to quantify various topics, such as correlation, similarity, saliency, and relevance. Based on extensive research, various metrics and techniques have been devised to evaluate the topics produced by topic models and determine the optimal number of topics over the last two decades [36], [37].

A framework was devised [38] to select the best LDA model based on topic density and to integrate this with model parameter estimation. This work is based on finding the model's best 'K' using an iterative process over the correlations in the document collection independent of the size of the data set. As described in (2), the proposed metric assesses the quality of topic structure using the average dissimilarity between topics based on cosine distance as described in (3). The metric's value was expected to decrease as the quality of topics/topic structure stability increased.

$$\frac{\sum_{i=1}^K \sum_{j=i+1}^K corr(T_i, T_j)}{K \times \frac{K-1}{2}} \tag{2}$$

$$corr(T_i, T_j) = \frac{\sum_{v=0}^V T_{iv} T_{jv}}{\sqrt{\sum_{v=0}^V (T_{iv})^2} \sqrt{\sum_{v=0}^V (T_{jv})^2}} \tag{3}$$

Where  $T_i, T_j$ : topics.

A metric for determining the optimal number of topics was proposed [39], demonstrating that by interpreting LDA as a matrix factorization method, a lower value of a proposed divergence metric between the factorized document-topic and topic-word distributions corresponds to a richer topic structure. As described in (4), the metric focuses on assessing topic similarity using the symmetric Kullback–Leibler (KL) Scatter among document-topic and topic-word matrices. KL Scatter is smaller when the number of topics is close to the optimal value. The metric is based on KL divergence, as defined in (5).

$$KL(C_{M1} || C_{M2}) + KL(C_{M2} || C_{M1}) \tag{4}$$

$$KL(R_{I1} || R_{I2}) = \sum_{i=1}^T R_{I1}(i) * \log\left(\frac{R_{I1}(i)}{R_{I2}(i)}\right) \tag{5}$$

Where,

$C_{M1}$ : distribution of normalized topic-word matrix  $M1$  over corpus  $C$

$C_{M2}$ : distribution of normalized product of a document-topic matrix and length of documents  $L * M2$  over corpus  $C$ .

A general mechanism was developed for building metrics to assess topic coincidence [40] and applying the mechanism to define the now de facto standard metrics for topic coherence. The idea was to assess the effectiveness of topics for classification by measuring the degree of semantic co-occurrence or coherence between high-probability words in the topic. The proposed metrics were defined using a general pipeline of four tasks: segmentation, probability estimation, confirmation, and aggregation. Metrics in (6) to (11) were proposed by combining the pipeline stages.

$$C_{UCI} = \frac{\sum_{i=1}^{N-1} \sum_{j=i+1}^N \log\left(\frac{P(w_i, w_j) + \epsilon}{P(w_j)P(w_i)}\right)}{\frac{N(N-1)}{2}} \tag{6}$$

$$C_{UMass} = \frac{\sum_{i=2}^N \sum_{j=1}^{i-1} \log\left(\frac{P(w_i, w_j) + \epsilon}{P(w_j)}\right)}{\frac{N(N-1)}{2}} \quad (7)$$

$$C_{NPMI} = \frac{\sum_{i=1}^{N-1} \sum_{j=i+1}^N NPMI(w_i, w_j)}{\frac{N(N-1)}{2}} \quad (8)$$

$$\vec{v}_{m,\gamma}(W') = \{\sum_{w_i \in W'} m(w_i, w_j)^\gamma\}_{j=1 \dots |W|} \quad (9)$$

$$C_V = \frac{\sum_{i=1}^{|W|} \vec{v}_{NPMI,1}(W')_i \cdot \vec{v}_{NPMI,1}(W^*)_i}{\|\vec{v}_{NPMI,1}(W')\|_2 \|\vec{v}_{NPMI,1}(W^*)\|_2} \quad (10)$$

$$NPMI(w_i, w_j) = \frac{\log\left(\frac{P(w_i, w_j) + \epsilon}{P(w_j)P(w_i)}\right)}{-\log(P(w_i, w_j) + \epsilon)} \quad (11)$$

Where,

$C_{UCI}$	UCI Coherence, based on pointwise mutual information (PMI)
$C_{UMass}$	UMass Coherence
$C_V$	Newly-proposed coherence measure
$NPMI$	Normalized PMI
$W', W^*$	Word subsets generated by segmentation
$N$	Number of most probable words per topic
$w_i, w_j$	Words (specific to a topic)
$P(w_i), P(w_j)$	Word probabilities
$P(w_i, w_j)$	Joint probability of observing words $w_i, w_j$
$\vec{v}_{m,\gamma}(W')$	Context vector for words in $W'$ , using direct confirmation measure $m$ and power $\gamma$
$\epsilon$ :	Epsilon for avoiding indeterminate log (0)

Further enhancements were proposed [41] over the previously used metrics of perplexity and coherence. This study focused on an extensive evaluation of the influence of perplexity and coherence over the topics produced by the LDA model to improve the F-measure for assessing the quality of topics for topic classification. The work determines the optimal K based on values of the normalized absolute perplexity (NAP) and normalized absolute coherence (NAC) (UMass), described in (12) and (13), respectively. The P stands for Perplexity for a topic model given in (14), and C for coherence of the topic model.

$$NAP = \frac{|P|}{\max(|P|)} \quad (12)$$

$$NAC = \frac{|C|}{\max(|C|)} \quad (13)$$

$$perplexity(D_{test}) = \exp\left(\frac{-\sum_{d=1}^M \log(p(w_d))}{\sum_{d=1}^M N_d}\right) \quad (14)$$

Another comprehensive metric [42], described in (15), focuses on judging the quality based on stability, predictive ability, and topic isolation of topic models. This combination metric is dependent on perplexity in (14), Jensen-Shannon divergence in (16) for isolation, stability (18), and coincidence based on coherence in (21).

$$\frac{Perplexity \times Coincidence}{\sigma_{JS} \times Stability} \quad (15)$$

Jensen-Shannon divergence-based isolation metric  $\sigma_{JS}$  is defined as (16).

$$\sigma_{JS} = \sqrt{\frac{\sum_{i=1}^K JS(S_i || S_{avg})^2}{k}} \quad (16)$$

Where,

$S_j$ :  $j$ th topic's word distribution  
 $S_{avg}$ : average word distribution for topics  
 $JS$ : Jensen-Shannon divergence, defined in (17).

$$JS(P||Q) = \frac{1}{2}KL(P||\frac{P+Q}{2}) + \frac{1}{2}KL(Q||\frac{P+Q}{2}) \tag{17}$$

To evaluate the stability, a set of topics ( $T_x, T_y$ ) were computed over the same dataset using as (19).

$$Stability(T_x, T_y) = \frac{1}{K} \sum_{i=1}^K AJ(R_{xi}, \pi(R_{xi})), \tag{18}$$

$$T_x = \{R_{x1}, R_{x2}, \dots, R_{xK}\} \tag{19}$$

Where,

$R_{xi}$ : Topic, as a set of vocabulary words  
 $\pi(R_{xi})$ : aligned topic from  $T_y$  with the highest AJ score (similarity) for  $R_{xi}$   
 $AJ$ : Average Jaccard index, expressing a word similarity across topic pairs as given in (20).

$$AJ(R_i, R_j) = \frac{1}{t} \sum_{d=1}^t \frac{R_{id} \cap R_{jd}}{R_{id} \cup R_{jd}} \tag{20}$$

Where,

$R_i, R_j$ : Topics, as a set of vocabulary words.  
 $R_{id}, R_{jd}$ : Sets of  $d$  most probable words of the topics.  
 $t$ : Maximum number of probable words to compare.  
 Coincidence based on coherence is evaluated as (21):

$$Coincidence = \frac{1}{c} \sum_{i=1}^K count(\pi(R_{xi})) \tag{21}$$

where,  $c$ : Number of coincident topics between  $T_x$  and  $T_y$

Apart from determining the optimal value of  $K$  using proposed metrics that assess topic structure, alternative techniques focusing on the automatic inference of the optimal value using assumed priors have also been proposed [43], [44]. Research work [45] analyzed and developed objective functions for evaluating topic models to determine optimal hyperparameters. The objective functions used in the approach improvise over the topic coherence metrics and are combined with external topic evaluation functions for analyzing topic similarity across pairs of generated topics. The combination allows for a holistic assessment with an extended focus on interpretability. Based on the observation that increasing the value of  $K$  leads to the factorization of coherent topics into constituents that still resemble the parent topic, a mapping  $\vec{m}$  of every topic to its most similar topic, the result is generated using as (22) and (23).

$$\vec{m} = \vec{m}_{ove} = \begin{pmatrix} \max \left\{ \frac{|t_1 \cap t|}{2N} \mid \forall t \in T \setminus \{t_1\} \right\} \\ \vdots \\ \max \left\{ \frac{|t_K \cap t|}{2N} \mid \forall t \in T \setminus \{t_K\} \right\} \end{pmatrix} \tag{22}$$

$$\vec{m} = \vec{m}_{cos} = \begin{pmatrix} \max \left\{ \cos_{t_1, t}(\theta) \mid \forall t \in T \setminus \{t_1\} \right\} \\ \vdots \\ \max \left\{ \cos_{t_K, t}(\theta) \mid \forall t \in T \setminus \{t_K\} \right\} \end{pmatrix} \tag{23}$$

Where,

$t, t_i$ : topic as word distribution,  $i$ th topic word distribution  
 $T$ : a set of word distributions for all topics  
 $N$ : number of most probable words per topic taken into consideration

The mapping is combined with the topic coherence of individual topics to obtain a measure  $\vec{\phi}_{Cv}$ . This measure assesses topics over both coherence and the presence of unique features. The measure is then

averaged to obtain the combined coherence value  $C_{C_v}$  described in (24) and (25). The topic model with the maximum  $C_{C_v}$  value corresponds to the best value of K.

$$\vec{\phi}_{C_{C_v}} = \vec{C}_v \odot (\vec{1} - \vec{m}) \quad (24)$$

$$C_{C_v} = \text{avg}(\vec{\phi}_{C_{C_v}}) \quad (25)$$

For this study, a set of evaluation metrics is chosen, as the observation of trends across multiple metrics allows for higher confidence and robustness in selecting the optimal value of 'K.' The LDA models generated in sub section 2.1 are evaluated on these metrics. Optimization direction varies across different evaluation metrics, as discussed in Table 1.

Table 1. Evaluation metrics used to estimate the optimal value of 'K'

Paper	Evaluation metric	Optimization direction
Chen <i>et. al.</i> 2008 [46]	Perplexity: likelihood of held-out data	Minimize
Cao <i>et. al.</i> 2009 [38]	Average cosine distance between topics	Minimize
Arun <i>et. al.</i> , 2010 [39]	Symmetric KL-divergence between normalized topic-word and document-topic distributions	Minimize
Röder <i>et. al.</i> , 2015 [40]	Coherence measures over topics: $C_{UMass}$ , $C_{UCI}$ , $C_{NPMI}$ & $C_V$	Maximize
Peikert <i>et al.</i> , 2021 [45]	Combined coherence value: $C_{C_v}$	Maximize
Gan and Qi, 2021 [42]	Topic superiority	Maximize

### 2.3. Determine optimal 'K' using grid search approach

Select the value of K for which all or the majority of the metrics agree. The rationale for the same is that various evaluation metrics evaluate the topic structure over different aspects, such as perplexity, isolation (Kullback-Leibler or Jensen-Shannon divergence), stability (average Jaccard index or cosine similarity), and coincidence (coherence measures). A common agreement of these varying metrics will likely ensure a higher quality topic structure, which an optimal value of hyperparameters will likely produce. The method is effectively a grid search [47] procedure due to the nature of exploration of the hyperparameter space to determine the best value based on an evaluation metric.

### 2.4. Datasets and experimental details

The proposed methodology is evaluated over three datasets: The 20-Newsgroups dataset [48] with over 15,000 news articles distributed across 20 categories; the web of science dataset [49] containing scientific documents from the web of science journal divided across multiple categories; and a recently curated dataset of Wikipedia articles [50], plain text Wikipedia 2020-21, comprising of over 1 million plain text Wikipedia articles across 605 bundles. Due to resource limitations, a randomly sampled subset of 2463 articles from the Wikipedia dataset and only the training partition of the 20-Newsgroups dataset with about 11,000 articles were used for evaluation. Also, the WOS11967 partition of the Web of Science dataset, comprising 11,967 documents across 35 categories, is utilized. In the experimental suite, LDA models were trained over the training subset for the different datasets, as mentioned in Table 2. Symmetric Dirichlet priors were used for both distributions, with the value set to 1/K. The models' training and the computation of evaluation metrics were performed in Python using the tmtoolkit library [51]. The models were evaluated for the chosen metric set with the training configuration of 10 passes over the corpus, 300 iterations per pass.

Table 2. Experimental details

Dataset used for the experiment	Experiment details
20-Newsgroups	For this experiment, LDA models were trained on the 20-Newsgroups dataset, and the values of K were varied within a range of 5 to 40, with an initial step of 5 and a finer step of 2 for the range of 16 to 24.
Web of Science	LDA models were trained on the Web of Science dataset in this experiment. For the experiment, the values of K were varied within a range of 5 to 40, with an initial step of 5 and a finer step of 2 for the range of 31 to 39.
Plain text Wikipedia	In the third experiment, LDA models were trained on a corpus consisting of a subset of 2,463 article documents from a recently-curated Wikipedia dataset. For the experiment, the values of K were varied within a range of 5 to 40, with an initial step of 5 and a finer step of 2 for the range of 10 to 20.

### 3. RESULTS AND DISCUSSION

The LDA models are evaluated on the proposed metric set to determine the optimal number of topic 'K'. The evaluation results are presented using the graphs. The graph is compared with the individual metrics to show the insufficiency of a single metric in determining the optimal 'K'.

#### 3.1. 20-Newsgroups

The graphical results of various evaluation metrics for varying values of K in the range of 5 to 40 with a step of 5 is shown in Figure 2 in appendix. From the graph, high agreement across metrics is observed in the range of 16 to 24, based on the respective optimization directions of the metrics. In this range, the value of perplexity is low, metrics with minimization as the optimization objective, such as `cao_juan_2009`, are low, and other metrics, such as coherence, combined coherence values (ccv), and topic score, demonstrate the highest values, with the peak value at 20, which is also the actual number of topics for the dataset. The metric `arun_2010` is also favorable in this range as the values are moderately low compared to the values across larger values of K, which indicates an increase in divergence and is unfavorable for the optimization direction of the metric (minimize). Following the results from the chosen range, Figure 3 in appendix demonstrates the results of evaluation metrics across a refined range of 16 to 24 with a step of 2. Similar to the trends observed before, the metrics are in high agreement in this range, with the highest agreement at the value of K=20, indicating the effectiveness of the method in determining the optimal value of K. Metrics such as coherence, combined coherence scores and topic scores demonstrate peak values at the actual value of K for the dataset, while other metrics such as perplexity, `cao_juan_2009`, and `arun_2010` are also suitably low in favor of the optimization objective. It is to note that while for this dataset, metrics like coherence or topic score alone may have been sufficient to determine the optimal K, using the agreement point across multiple metrics increases the confidence in choosing the value of K. Further, the benefit of use of multiple metrics and the agreement across them is observed for datasets where the actual value of K is unknown. All metrics do not fully agree at the same value of K. For this dataset, the method accurately deduces the optimal value of K as 20.

#### 3.2. Web of Science

The outcomes of diverse evaluation metrics for different values of K ranging from 5 to 40, with increments of 5, are shown in Figure 4 in appendix. The graph indicates a substantial concurrence among the metrics in the interval of 30 to 40, aligning with their respective optimization objectives. Within this range, the perplexity value is minimal, and metrics oriented towards minimization, such as `cao_juan_2009`, achieve their lowest values. On the other hand, metrics like coherence, combined coherence values (ccv), and topic score exhibit the highest values, reaching a peak at K=35, which coincides with the actual number of topics in the dataset. The metric `arun_2010` also performs favorably in this range, displaying moderately low values compared to larger K values, indicating an undesirable increase in divergence for the metric's optimization objective (minimize). Based on the findings from the selected range, Figure 5 in appendix illustrates the evaluation metrics' results within a more refined range, from 31 to 39, with increments of 2. Similar to the previous observations, the metrics demonstrate strong agreement within this range, with the highest concurrence occurring at K=35, indicating the method's effectiveness in determining the optimal K value. Metrics such as coherence, combined coherence scores, and topic scores achieve their peak values at the actual K value of the dataset. Furthermore, other metrics like perplexity, `cao_juan_2009`, and `arun_2010` remain suitably low, aligning with their respective optimization objectives. Although all metrics do not ideally agree on the same K value, the method accurately deduces this dataset's optimal K value as 35.

#### 3.3. Plain text Wikipedia

The values of the various evaluation metrics for values of K in the range of 5 to 40 with a step of 5 are shown in the graphs in Figure 6 in appendix. The graphs show that the values of the metrics are in agreement (based on the desired optimization direction) in the range of 10 to 20. For this range, the perplexity value is low, and the value does not decrease significantly as the number of topics increases, indicating a saturation region. The other metrics to be minimized, `cao_juan_2009` and `arun_2010`, are also relatively small. While the values of these metrics continue to decrease steadily as the number of topics increases, the change in perplexity is minimal, which supports the choice of the specified range. The decision for the range is further supported by the coherence measures, which have the highest recorded values from 10 to 20. As the number of topics increases, the values of these measures decrease rapidly, representing a declining topic structure. The combined coherence values and topic score are also relatively high in this range, with the increase diminishing for higher K values, indicating more stable and coherent topics from the selected range. The metrics values are computed for topics in the range of 10 to 20 with a step of 3 for finer tuning, as shown in Figure 7 in appendix. The graphs show that the metrics are in high agreement in the range of 14 to 16, where the values of KL-divergence and cosine dissimilarity are considerably low, and the



coherence values of produced topics are high, as indicated by all the coherence metrics. Topic scores and combined coherence values are also the highest in this local range. Further, the perplexity also agrees with the other metrics and is relatively minimal in the constrained range. The benefit of the proposed method is visible here, as metrics are not in complete agreement at any point. However, considering the highest agreement point, the most optimal value for K for the dataset can be deduced. For the selected corpus, the value of K=15 is the most suitable.

### 3.4. Comparison with existing work

The proposed work uses a combination of metrics rather than a single metric. This approach provides a comprehensive evaluation as different metrics capture different aspects of the model's performance. Perplexity focuses on the model's predictive power, coherence measures the interpretability of topics, and topic superiority evaluates the extent to which a topic is distinguishable. The exclusive reliance on Cao\_Juan\_2009 and Arun\_2010 methodologies preclude the determination of optimal values for K in our topic modeling endeavors. This limitation becomes evident when scrutinizing an expansive array of metrics, vividly illustrated in Figures 2 and 3 for experiment 1. A similar trend has been observed in experiment 2, as depicted in Figures 4 and 5, thereby underscoring the necessity for combining these metrics to provide a more holistic view of the model's effectiveness. Factors such as dataset characteristics, modeling assumptions, and algorithmic choices can influence the model's evaluation. Using a diverse set of metrics helps to ensure the robustness of the evaluation, as inconsistencies in one metric may be compensated by others. Based on this, Table 3 discusses a detailed comparison of the individual metrics with the proposed work.

Table 3. Comparison of proposed work with the baseline metrics

	For K in the range	Perplexity (minimize)	Cao_Juan_2009 (minimize)	Arun_2010 (minimize)	Coherence measures (maximize)	Combined coherence (maximize)	Topic superiority (maximize)	Result from proposed framework
Experiment 1: 20-Newsgroups (K=20)	10 to 30 (Figure 2)	As seen in all the graphs, this metric shows a decreasing trend with increasing topics. Hence, this metric alone cannot depict the optimal 'K.'	25	25	20	20	18	20
	16 to 24 (Figure 3)		22	20	20	20	20	
Experiment 2: Web of Science (K= 35)	25 to 40 (Figure 4)		40	30	33	34	34	35
	31 to 39 (Figure 5)		39	33	35	35	35	
Experiment 3: plain text Wikipedia (unknown K)	5 to 20 (Figure 6)		20	20	16	16	18	15
	10 to 20 (Figure 7)		18	15	14	16	15	

## 4. CONCLUSION AND FUTURE WORK

This research aims to determine the optimal value of 'K' in training an LDA model while maintaining symmetric  $\alpha$  and  $\beta$  priors using a grid search approach over the metric set. The methodology effectively identifies the optimal metric values by leveraging the consensus among established evaluation metrics. Experimental results from three real-world text corpora, two with a known optimal value of 'K' and the third with an unknown optimal value, demonstrate the method's ability to accurately match the known value and predict a suitable value for the unknown dataset. As a result, the models generated using this method exhibit high-quality topic structure. The proposed approach enhances the robustness and reliability of LDA models by providing a systematic and objective approach to determine the optimal value of 'K,' ensuring improved accuracy and interpretability across many applications for the LDA models. This work contributes to advancing the state-of-the-art in LDA model training, providing a valuable resource for researchers seeking a principled approach to enhance the performance of their models in diverse real-world scenarios. Further, this work can be extended to determine the optimal values of other hyperparameters ' $\alpha$ ' and ' $\beta$ .' These hyperparameters can be optimized by evaluating them one at a time or with different values of the other hyperparameters. Another prospect is to study the method's effectiveness over non-textual corpora comprising the image or audio-visual data or word embeddings extracted from text to devise topic models. The results benefit the recent applications of topic models over image data, where topic models are incorporated for discovering image topics to aid in classification and image captioning tasks.

APPENDIX

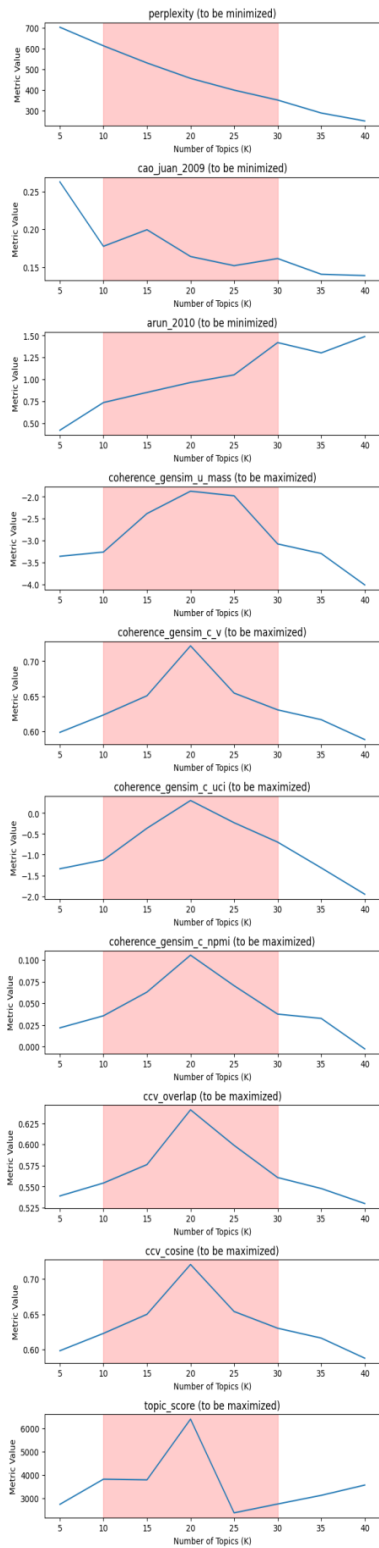


Figure 2. Plots of evaluation metrics against the number of topics 'K' for K in the range of 5 to 40 (inclusive) for the 20-Newsgroups dataset. The shaded region denotes a suitable region from which to select K

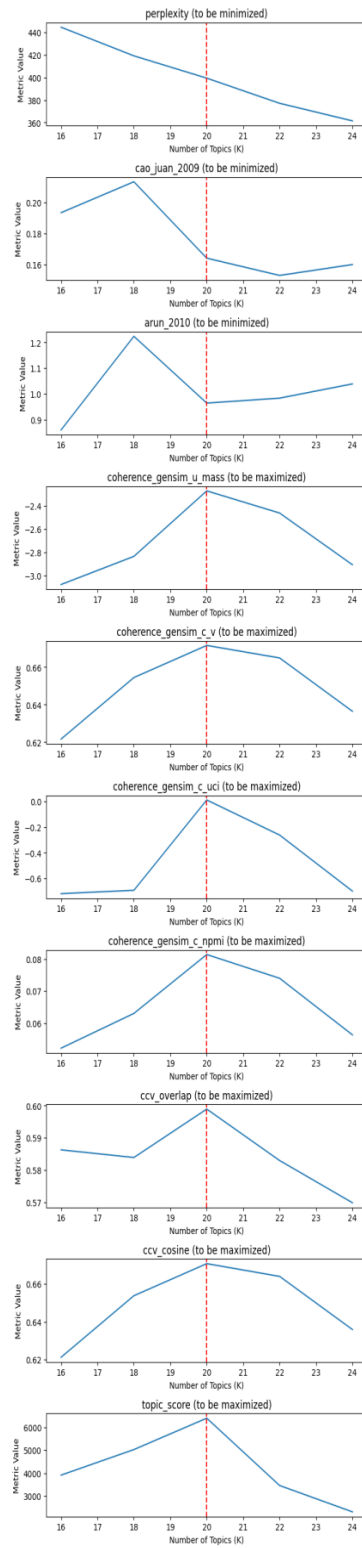


Figure 3. Plots of evaluation metrics against the number of topics 'K' for K in the range of 16 to 24 (inclusive) for the 20-Newsgroups dataset. The suitable region for selecting K is between K=19 and K=21, with the most reasonable value (K=20) denoted by the red dashed line

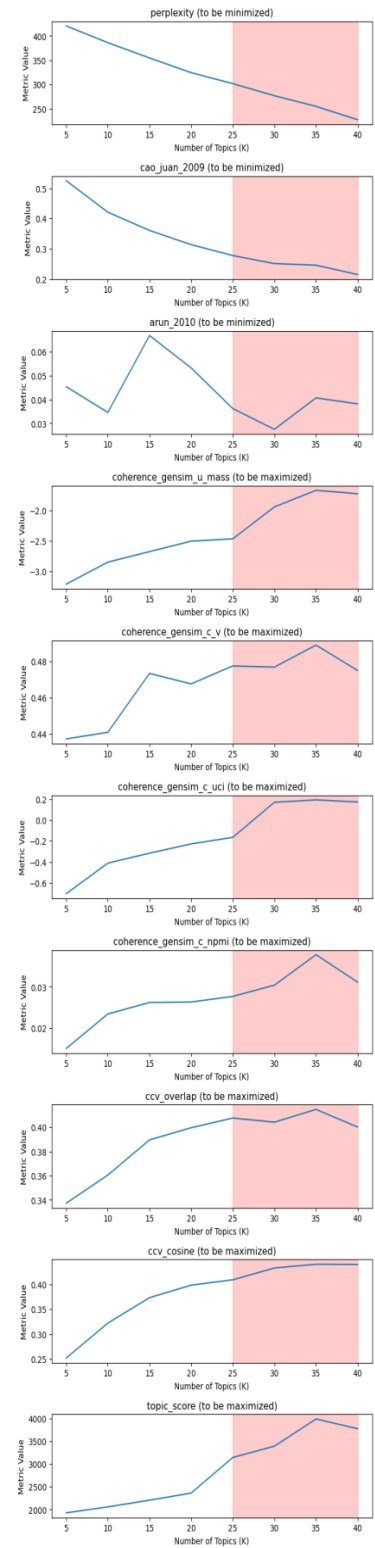


Figure 4. Plots of evaluation metrics against the number of topics 'K' for K in the range of 5 to 40 (inclusive) for the Web of Science dataset. The shaded region denotes a suitable region from which to select K

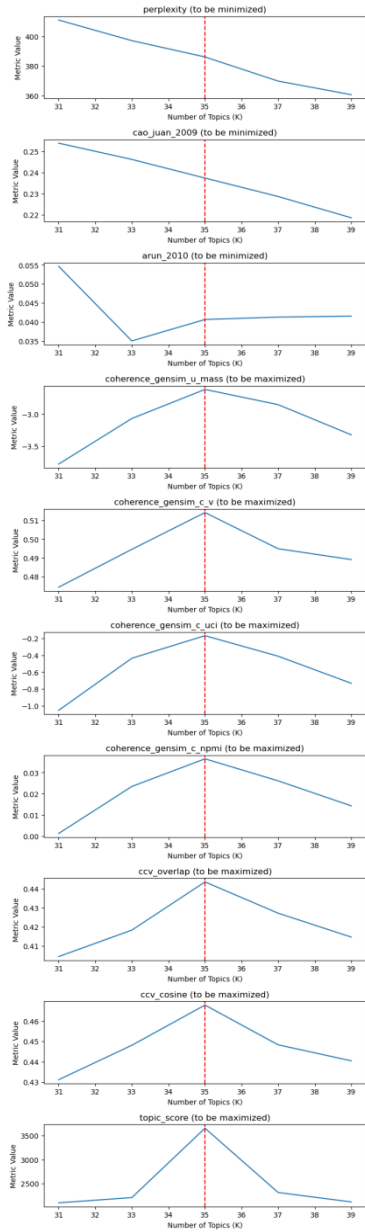


Figure 5. Plots of evaluation metrics against the number of topics 'K' for K in the range of 31 to 39 (inclusive) for the Web of Science dataset. The red dashed line denotes the most reasonable value (K=35)

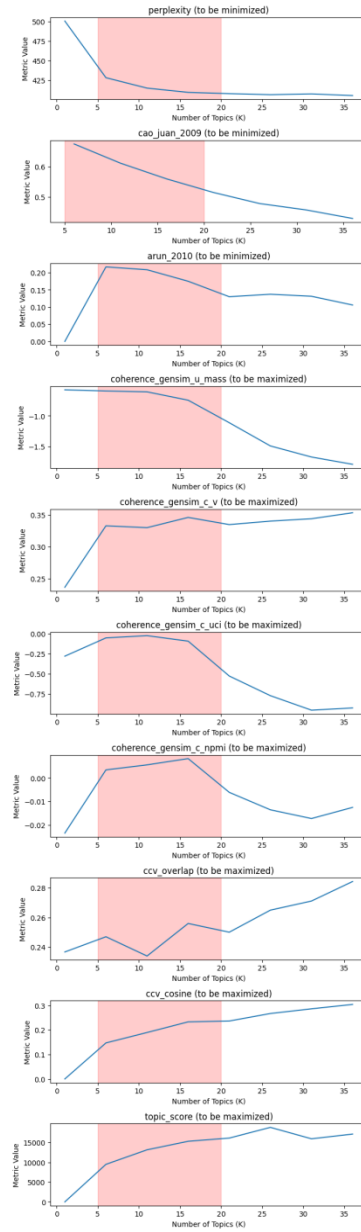


Figure 6. Plots of evaluation metrics against the number of topics 'K' for K in the range of 5 to 40 (inclusive) for the Wikipedia dataset. The shaded region denotes a suitable region from which to select K

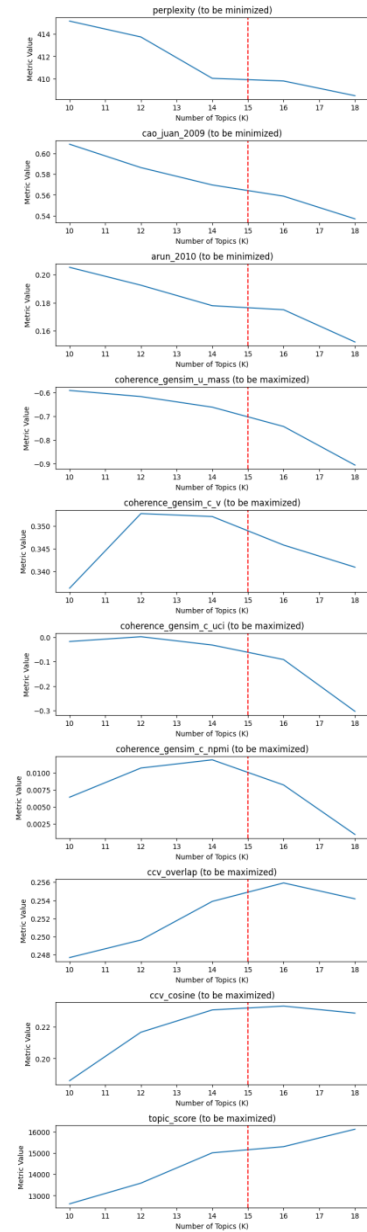


Figure 7. Plots of evaluation metrics against the number of topics 'K' for K in the range of 10 to 20 (inclusive) for the Wikipedia dataset. The suitable region for selecting K is between K=14 and K=16, with the most reasonable value (K=15) denoted by the red dashed line

**REFERENCES**




- [1] P. Kherwa and P. Bansal, "Topic modeling: a comprehensive review," *EAI Endorsed Transactions on Scalable Information Systems*, vol. 7, no. 24, pp. 1–16, Jul. 2020, doi: 10.4108/eai.13-7-2018.159623.
- [2] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, Sep. 1990, doi: 10.1002/(SICI)1097-4571(199009)41:6<391:AID-ASII>3.0.CO;2-9.
- [3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003, doi: 10.1162/jmlr.2003.3.4-5.993.
- [4] T. Hofmann, "Probabilistic latent semantic analysis," *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, 1999, doi: 10.48550/arXiv.1301.6705.
- [5] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999, doi: 10.1038/44565.

- [6] A. Goyal and I. Kashyap, "Latent Dirichlet allocation - an approach for topic discovery," in *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing, COM-IT-CON 2022*, May 2022, pp. 97–102. doi: 10.1109/COM-IT-CON54601.2022.9850912.
- [7] D. M. Blei and J. D. Lafferty, "A correlated topic model of Science," *The Annals of Applied Statistics*, vol. 1, no. 1, Jun. 2007, doi: 10.1214/07-aos114.
- [8] L. Wei and A. McCallum, "Pachinko allocation: DAG-structured mixture models of topic correlations," in *ACM International Conference Proceeding Series*, 2006, vol. 148, pp. 577–584. doi: 10.1145/1143844.1143917.
- [9] P. Liu, L. Li, W. Heng, and B. Wang, "HLDA based text clustering," in *2012 IEEE 2nd International Conference on Cloud Computing and Intelligence Systems*, Oct. 2012, pp. 1465–1469. doi: 10.1109/CCIS.2012.6664628.
- [10] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd international conference on Machine learning - ICML '06*, 2006, pp. 113–120. doi: 10.1145/1143844.1143859.
- [11] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," 2012, doi: 10.48550/arXiv.1207.4169.
- [12] C. Zou, "Analyzing research trends on drug safety using topic modeling," *Expert Opinion on Drug Safety*, vol. 17, no. 6, pp. 629–636, Jun. 2018, doi: 10.1080/14740338.2018.1458838.
- [13] M. Rüdiger, D. Antons, A. M. Joshi, and T.-O. Salge, "Topic modeling revisited: New evidence on algorithm performance and quality metrics," *PLOS ONE*, vol. 17, no. 4, p. e0266325, Apr. 2022, doi: 10.1371/journal.pone.0266325.
- [14] N. Aletras and M. Stevenson, "Evaluating topic coherence using distributional semantics," *Proceedings of the 10th International Conference on Computational Semantics, IWCS 2013 - Long Papers*, 2013.
- [15] S. Terragni, A. Candelieri, and E. Fersini, "The role of hyper-parameters in relational topic models: Prediction capabilities vs topic quality," *Information Sciences*, vol. 632, pp. 252–268, Jun. 2023, doi: 10.1016/j.ins.2023.02.076.
- [16] W. Zhao *et al.*, "A heuristic approach to determine an appropriate number of topics in topic modeling," *BMC Bioinformatics*, vol. 16, no. 13, p. S8, Dec. 2015, doi: 10.1186/1471-2105-16-S13-S8.
- [17] A. Ianina and K. Vorontsov, "Regularized multimodal hierarchical topic model for document-by-document exploratory search," in *Conference of Open Innovation Association, FRUCT*, Nov. 2019, pp. 131–138. doi: 10.23919/FRUCT48121.2019.8981493.
- [18] R. Hossain and D. D. Timmer, "Machine learning model optimization with hyper parameter tuning approach," *Global Journal of Computer Science and Technology*, vol. 21, no. 2, pp. 7–13, 2021.
- [19] R. Alghamdi and K. Alfalqi, "A survey of topic modeling in text mining," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 1, 2015, doi: 10.14569/ijacsa.2015.060121.
- [20] U. Chauhan and A. Shah, "Topic modeling using latent Dirichlet allocation: a Survey," *ACM Computing Surveys*, vol. 54, no. 7, pp. 1–35, Sep. 2022, doi: 10.1145/3462478.
- [21] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, Apr. 2012, doi: 10.1145/2133806.2133826.
- [22] G. Ignatow and R. Mihalcea, "Topic models," *Text Mining: A Guidebook for the Social Sciences*, pp. 156–162, 2018, doi: 10.4135/9781483399782.n15.
- [23] R. Albalawi, T. H. Yeap, and M. Benyoucef, "Using topic modeling methods for short-text data: a comparative analysis," *Frontiers in Artificial Intelligence*, vol. 3, Jul. 2020, doi: 10.3389/frai.2020.00042.
- [24] M. D. Hoffman, D. M. Blei, and F. Bach, "Online learning for latent Dirichlet allocation," *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010, NIPS 2010*, 2010, doi: 10.5555/2997189.2997285.
- [25] A. Goyal and I. Kashyap, "Comprehensive analysis of topic models for short and long text data," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 12, pp. 249–259, 2023, doi: 10.14569/IJACSA.2023.0141226.
- [26] H. Jelodar *et al.*, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 15169–15211, Jun. 2019, doi: 10.1007/s11042-018-6894-4.
- [27] M. Gupta and P. Gupta, "Research and implementation of event extraction from twitter using LDA and scoring function," *International Journal of Information Technology (Singapore)*, vol. 11, no. 2, pp. 365–371, Jun. 2019, doi: 10.1007/s41870-018-0206-0.
- [28] N. N. Alabid and Z. Naseer, "Summarizing twitter posts regarding COVID-19 based on n-grams," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 31, no. 2, pp. 1008–1015, Aug. 2023, doi: 10.11591/ijeecs.v31.i2.pp1008-1015.
- [29] N. Shafi and M. A. Chachoo, "Query intent recognition by integrating latent Dirichlet allocation in conditional random field," *International Journal of Information Technology (Singapore)*, vol. 15, no. 1, pp. 183–191, Jan. 2023, doi: 10.1007/s41870-022-01108-3.
- [30] O. Iparraguirre-Villanueva *et al.*, "Search and classify topics in a corpus of text using the latent Dirichlet allocation model," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 30, no. 1, pp. 246–256, Apr. 2023, doi: 10.11591/ijeecs.v30.i1.pp246-256.
- [31] V. Kalra, I. Kashyap, and H. Kaur, "Classification based topic extraction using domain-specific vocabulary: a supervised approach," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 26, no. 1, p. 442, Apr. 2022, doi: 10.11591/ijeecs.v26.i1.pp442-449.
- [32] O. Ugochi, R. Prasad, N. Odu, E. Ogidiaka, and B. H. Ibrahim, "Customer opinion mining in electricity distribution company using twitter topic modeling and logistic regression," *International Journal of Information Technology (Singapore)*, vol. 14, no. 4, pp. 2005–2012, Jun. 2022, doi: 10.1007/s41870-022-00890-4.
- [33] A. Abdelrazek, Y. Eid, E. Gawish, W. Medhat, and A. Hassan, "Topic modeling algorithms and applications: a survey," *Information Systems*, vol. 112, p. 102131, Feb. 2023, doi: 10.1016/j.is.2022.102131.
- [34] A. Xu and P. Guan, "Bayesian learning for dynamic inference," 2022, doi: 10.48550/arXiv.2301.00032.
- [35] D. Greene, D. O'Callaghan, and P. Cunningham, "How many topics? Stability analysis for topic models," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8724 LNAI, no. PART 1, 2014, pp. 498–513. doi: 10.1007/978-3-662-44848-9\_32.
- [36] D. Korencic, S. Ristov, J. Repar, and J. Snajder, "A topic coverage approach to evaluation of topic models," *IEEE Access*, vol. 9, pp. 123280–123312, 2021, doi: 10.1109/ACCESS.2021.3109425.
- [37] A. Pereira, F. Viegas, M. A. Gonçalves, and L. Rocha, "Evaluating the limits of the current evaluation metrics for topic modeling," in *ACM International Conference Proceeding Series*, Oct. 2023, pp. 119–127. doi: 10.1145/3617023.3617040.
- [38] J. Cao, T. Xia, J. Li, Y. Zhang, and S. Tang, "A density-based method for adaptive LDA model selection," *Neurocomputing*, vol. 72, no. 7–9, pp. 1775–1781, Mar. 2009, doi: 10.1016/j.neucom.2008.06.011.




- [39] R. Arun, V. Suresh, C. E. V. Madhavan, and M. N. Murty, "On finding the natural number of topics with latent Dirichlet allocation: some observations," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6118 LNAI, no. PART 1, 2010, pp. 391–402. doi: 10.1007/978-3-642-13657-3\_43.
- [40] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, Feb. 2015, pp. 399–408. doi: 10.1145/2684822.2685324.
- [41] M. Hasan, A. Rahman, M. R. Karim, M. S. I. Khan, and M. J. Islam, "Normalized approach to find optimal number of topics in latent dirichlet allocation (lda)," in *Advances in Intelligent Systems and Computing*, vol. 1309, 2021, pp. 341–354. doi: 10.1007/978-981-33-4673-4\_27.
- [42] J. Gan and Y. Qi, "Selection of the optimal number of topics for LDA topic model—taking patent policy analysis as an example," *Entropy*, vol. 23, no. 10, p. 1301, Oct. 2021, doi: 10.3390/e23101301.
- [43] Z. Chen and H. Doss, "Inference for the number of topics in the latent Dirichlet allocation model via Bayesian mixture modeling," *Journal of Computational and Graphical Statistics*, vol. 28, no. 3, pp. 567–585, Jul. 2019, doi: 10.1080/10618600.2018.1558063.
- [44] W. Xia and H. Doss, "Scalable hyperparameter selection for latent Dirichlet allocation," *Journal of Computational and Graphical Statistics*, vol. 29, no. 4, pp. 875–895, Oct. 2020, doi: 10.1080/10618600.2020.1741378.
- [45] S. Peikert, C. Kubach, J. Al Qundus, L. D. S. Vu, and A. Paschke, "Objective functions to determine the number of topics for topic modeling," *ACM International Conference Proceeding Series*, pp. 328–332, 2021, doi: 10.1145/3487664.3487710.
- [46] S. Chen, D. Beeferman, and R. Rosenfeld, "Evaluation metrics for language models," *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pp. 275–280, 1998, doi: 10.1184/R1/6605324.v1.
- [47] R. R. Sarra, A. M. Dinar, and M. A. Mohammed, "Enhanced accuracy for heart disease prediction using artificial neural network," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 29, no. 1, pp. 375–383, Jan. 2023, doi: 10.11591/ijeecs.v29.i1.pp375-383.
- [48] K. Lang, "NewsWeeder: learning to filter netnews," *Proceedings of the 12th International Conference on Machine Learning, ICML 1995*, pp. 331–339, 1995, doi: 10.1016/b978-1-55860-377-6.50048-7.
- [49] K. Kowsari, *Web of Science Dataset*, Mendeley, 2018
- [50] D. Shapiro, "Plain Text Wikipedia 2020-21: De-marked up Wikipedia for offline use," 21 September 2021. [Online]. Available: <https://www.kaggle.com/datasets/lcmdrdata/plain-text-wikipedia-202011>.
- [51] M. Konrad, "tmtoolkit: Text mining and topic modeling toolkit, version 0.11.2," 2022. [Online]. Available: <https://github.com/WZBSocialScienceCenter/tmtoolkit/>.

## BIOGRAPHIES OF AUTHORS



**Astha Goyal**    is an assistant professor at Keshav Mahavidyalaya, University of Delhi, India. She has been teaching for the last 12 years and has been active in the field of research and innovation, as well as other administrative tasks. She is a gold medalist in MCA from Guru Gobind Singh Indraprastha University, Delhi. She is currently pursuing her Ph.D. (Computer Science and Engineering) from Manav Rachna International Institute of Research and Science. Her research areas include natural language processing, text mining, machine learning, and image processing. She can be contacted at email: [astha.gmca84@gmail.com](mailto:astha.gmca84@gmail.com).



**Indu Kashyap**    has 15 years of teaching, administration, and research experience. She is a Professor at Manav Rachna International Institute of Research and Studies. She has a Ph.D. from Chaudhary Charan Singh University, Meerut. She has over 60 publications in reputed journals and conferences, including Elsevier, Springer, and Taylor & Francis. Her research areas include wireless networks, machine learning, data analytics, and recommender systems. She can be contacted at email: [indu.kashyap82@gmail.com](mailto:indu.kashyap82@gmail.com).