Comprehensive Analysis of Topic Models for Short and Long Text Data

Astha Goyal[1], Indu Kashyap[2]
Research Scholar, Department of CSE, MRIIRS, Faridabad, India[1]
Professor, Department of CSE, MRIIRS, Faridabad, India[2]

***Abstract*—The digital age has brought significant information to the Internet through long text articles, webpages, and short text messages on social media platforms. As the information sources continue to grow, Machine Learning and Natural Language Processing techniques, including topic modeling, are employed to analyze and demystify this data. The performance of topic modeling algorithms varies significantly depending on the text data's characteristics, such as text length. This comprehensive analysis aims to compare the performance of the state-of-the-art topic models: Nonnegative Matrix Factorization (NMF), Latent Dirichlet Allocation using Variational Bayes modeling (LDA-VB), and Latent Dirichlet Allocation using Collapsed Gibbs-Sampling (LDA-CGS), over short and long text datasets. This work utilizes four datasets: Conceptual Captions and Wider Captions, image captions for short text data, and 20 Newsgroups news articles and Web of Science containing science articles for long text data. The topic models are evaluated for each dataset using internal and external evaluation metrics and are compared against a known value of topic 'K.' The internal and external evaluation metrics are the statistical metrics that assess the model's performance on classification, significance, coherence, diversity, similarity, and clustering aspects. Through comprehensive analysis and rigorous evaluation, this work illustrates the impact of text length on the choice of topic model and suggests a topic model that works for varied text length data. The experiment shows that LDA-CGS performed better than other topic models over the internal and external evaluation metrics for short and long text data.**

Keywords*—Topic modeling; Nonnegative Matrix Factorization (NMF); Latent Dirichlet Allocation (LDA); evaluation metrics; short text mining; long text mining

I. Introduction

Topic modeling has emerged as a powerful technique for uncovering hidden thematic structures within large volumes of textual data. It provides valuable insights by automatically identifying and extracting topics from unstructured text. It is a powerful tool for text analysis, and it has been used for various applications, including document classification, summarization, and recommendation systems. The basic idea behind topic modeling is to assume that each document in a corpus can be represented as a mixture of topics. The topics are latent variables that are not directly observed in the data. However, the topics can be inferred from the words that appear in the documents.

The performance of topic modeling algorithms is influenced by the characteristics of the data, including its length. Short text data poses challenges in finding the co-occurrence of topic patterns due to data sparsity, noise, topic imbalance, and lack of context [1]. Conversely, long text data presents computational complexity, overfitting, and interpretability issues when employing topic modeling algorithms. It is also hardly feasible to infer a unique and coherent topic from a long text because it usually contains various topics [2]. In this research domain, the foremost obstacle lies in selecting an optimal topic model that remains effective irrespective of the length of the text or that works for the specific application domain. The secondary challenge entails the fine-tuning and enhancement of the chosen topic model.

The present study aims to compare various topic modeling algorithms in the context of short and long text data, investigating their effectiveness and efficiency under different conditions. After an extensive review of existing literature in this field, LDA emerges as a top-performing model for extensive text data. At the same time, NMF excels in shorter text contexts [3], [4]. Recent attention has also turned to LDA's adaptability. One key feature that makes LDA adaptable is the use of the Bayesian framework. This means that LDA incorporates the prior knowledge about the data into the model, improving the model's performance, especially for small or noisy datasets. Therefore, this study evaluates the NMF topic model against two distinct variations of the LDA topic model: LDA employing Variational Bayes (LDA-VB) and LDA using Collapsed Gibbs Sampling (LDA-CGS). The goal is to introduce a topic modeling approach that remains effective regardless of the text length.

These models are subjected to rigorous evaluation using internal and external evaluation metrics, referencing the known number of topics 'K' to ensure unbiased results [5], [6]. Internal evaluation metrics measure the quality of the topic model itself without reference to any external data [7]. External evaluation metrics measure the quality of the topic model by evaluating its performance on a downstream task, such as text classification or document clustering [6]. The findings of these experiments illuminate how text length influences topic modeling outcomes, offering insights into selecting the most suitable topic model regardless of text length. Additionally, to ensure the robustness of the evaluation, a diverse selection of datasets encompassing both long and short text data, thereby mirroring the heterogeneity of real-world textual information sources, has been thoughtfully chosen to ensure the robustness of the evaluation. Namely, the Conceptual Captions [8] and WIDER Captions [9] datasets are selected for short text, and the 20

Newsgroups [10] and Web of Science [11] datasets are selected for long text.

The paper is structured as follows: Section II discusses the related work for topic models NMF and LDA. The experimental exploration for comparing different topic models on both short and long text data using evaluation metrics is presented in Section III. Section IV presents a comparative analysis of the obtained results. The paper is concluded in Section V with suggestions for future research directions.

II. Background

Topic modeling represents a fundamental technique in unsupervised text analysis, crucial in unveiling concealed thematic structures within a collection of texts. These algorithms are developed to detect patterns of words appearing together and capture the underlying semantic themes that define a group of documents. Topic models identify these themes and allocate them to individual documents. A document encompasses multiple topics, as reflected by its weighted coefficient. The topic with the highest weight dictates the document's primary association, disregarding all other assignments. The subsequent sections discuss the topic models and their assessment using evaluation metrics.

A. *Topic Models*

Topic models produce a list of outputs termed "topic descriptors" using words strongly linked to each topic [12]. From an algorithmic perspective, topics can be envisioned as patterns that emerge from the co-occurrence of words within a given corpus. Numerous algorithms for topic modeling have been developed to address the intricacies inherent in capturing and representing topics within textual data. These algorithms use various mathematical and probabilistic techniques to accomplish their intended objectives.

The development of topic modeling methodologies commenced with Latent Semantic Indexing (LSI), which is referred to as Latent Semantic Analysis (LSA) within the context of topic modeling [13]. LSA employs Singular Value Decomposition (SVD) on the term-document matrix, effectively reducing its dimensionality and capturing latent topics. Although not strictly a probabilistic model, LSA played a pivotal role in the evolution of topic modeling. Another variant of SVD, NMF, was subsequently devised to handle sparse data. NMF dissects the term-document matrix into nonnegative matrices representing topics and their corresponding word distributions [14]. Following this, a probabilistic variant of LSA emerged, known as Probabilistic Latent Semantic Analysis (PLSA), serving as a precursor to the LDA topic model [15]. Across the literature, it has been found that the NMF topic model works better for short text data, whereas LDA is famous for long text data. LDA, the most widely utilized algorithm in topic modeling, has been explored with different sampling methods, and it has been observed that LDA-VB and LDA-CGS are two top-performing variants for topic modeling.

B. *Nonnegative Matrix Factorization (NMF)*

NMF is a non-probabilistic topic model based on the factorization method [16]. In this method, the encoded TF-IDF term-document matrix (sized by M × N) for a given text corpus is decomposed into two matrices: term-topic matrix U and topic-document matrix V, corresponding to K coordinate axes and N points (each point represents one document) in a new semantic space, respectively as shown in Fig. 1.

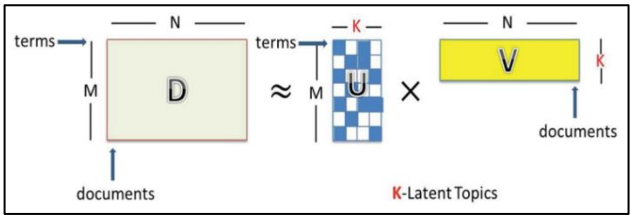


Fig. 1. NMF topic model using factorization method: D ≈ UV, with U and V elementwise nonnegative. [16].

A myriad of diverse applications in various fields use NMF. It is being applied in computational biology to analyze gene expression, unveiling metagenes and their expression patterns [17]. Image processing identifies hidden structures by extracting basis and coding matrices, revealing distinct image components. NMF has been extended to data clustering and pattern discovery across domains through automatic cluster extraction [18].

Concurrently, within the domain of linear algebraic models, there is a consensus among scholars regarding the effectiveness of NMF in handling brief textual content, exemplified by tweets. Notably, NMF's capacity for topic extraction requires no prerequisite knowledge, making it particularly advantageous for research endeavors rooted in social media data [4]. However, being a non-probabilistic topic model, it has limited utility.

C. *Latent Dirichlet Allocation (LDA)*

LDA is a probabilistic topic model postulates that each document consists of a combination of a limited number of topics, each characterized by a word distribution. The primary objective of LDA is to ascertain the optimal topic distribution for each document and the word distribution associated with each topic.

In recent times, the growing demand for topic modeling has stimulated research efforts to enhance the precision and efficiency of inference methods. LDA with Variational Bayes (LDA-VB) assumes a central role within this framework by approximating the posterior distribution governing latent topics and topic proportions. This approximation not only enhances the computational efficiency of LDA but also renders it amenable to the analysis of substantial text datasets [6], in contrast to traditional LDA, which exhibits limitations in computational efficiency and scalability. Despite this distinction, both approaches yield comparable levels of accuracy. Traditional LDA holds an advantage in terms of flexibility, as it accommodates the modeling of hierarchical topic structures, a feature lacking in LDA-VB [19]. LDA's Bayesian framework enhances performance on small or noisy datasets by incorporating prior knowledge, such as known vocabulary, to improve topic identification accuracy. Its flexibility in determining the number of topics makes it suitable for short and long text data, adapting to the data's characteristics. The algorithm for LDA-VB is shown in Fig. 2.

```
Initialize $\lambda^{(0)}$ randomly.
Set the step-size schedule $\rho_t$ appropriately.
**repeat**
  Sample a document $w_d$ uniformly from the data set.
  Initialize $\gamma_{dk} = 1$, for $k \in \{1, \dots, K\}$.
  **repeat**
    For $n \in \{1, \dots, N\}$ set

        $\phi_{dn}^{k} \propto exp\{\mathbb{E}[log\, \theta_{dk}] + \mathbb{E}[log\, \beta_{k,w_{dn}}]\}, k \in \{1, \dots, K\}.$

    Set $\gamma_d = \alpha + \sum_n \phi_{dn}.$
  **until** local parameters $\phi_{dn}$ and $\gamma_d$ converge.
  For $k \in \{1, \dots, K\}$ set intermediate topics

        $\hat{\lambda}_k = \eta + D\sum_{n=1}^{N} \phi_{dn}^{k} w_{dn}.$

  Set $\lambda^{(t)} = (1 - \rho_t)\lambda^{(t-1)} + \rho_t \hat{\lambda}.$
**until** forever
```

Fig. 2. Algorithm for LDA variational bayes [20].

Among the favored techniques based on sampling for inference, Collapsed Gibbs Sampling (CGS) in conjunction with LDA stands out as a prominent choice. CGS has proven effective and is commonly employed to infer latent topic models [21]. LDA-CGS optimizes LDA for short text data by collapsing latent variables, reducing computational complexity. Empirical evidence suggests LDA-CGS outperforms traditional LDA in various short text applications [22], [23]. The algorithm for collapsed Gibbs sampling is shown in Fig. 3.

```
**for** $a \leftarrow 1$ to $N$:
        $\boldsymbol{u} \leftarrow$ draw from Uniform $[0,1]$
        **for** $k \leftarrow 1$ to $K$:
            $P[k] \leftarrow P[k-1] + \frac{\left(N_{kj}^{\neg aj} + \alpha\right)\left(N_{x_{aj}k}^{\neg aj} + \beta\right)}{\left(N_{k}^{\neg ij} + W\beta\right)}$
        **for** $k \leftarrow 1$ to $K$:
            **if** $u < P[k]/P[K]$
                **then** $z_{aj} = k$, stop
```

Fig. 3. Algorithm for Collapsed Gibbs Sampling [24]

LDA finds applications in text classification, document clustering, recommendation systems, topic tracking, and sentiment analysis, enabling categorization, grouping, recommendations, temporal analysis, and sentiment extraction [12]. The following section reviews previous research work for evaluating and comparing topic models.

D. Related Work

Several studies highlight the crucial aspects of assessing topic models using statistical metrics like coherence, stability, diversity, and topic score, comparing them under different conditions using varied datasets.

The comprehensive evaluation performed by Albalawi et al. [7], using the 20-newsgroup dataset and short conversation data from Facebook, sheds light on the efficacy of various Topic Modeling methods. Standard metrics, including recall, precision, F-score, and coherence, were employed to ascertain these methods' ability to generate well-organized and meaningful topics. Notably, this rigorous assessment revealed that two TM methods, LDA and NMF, outperformed the others. Lim et al. [25] explored coherence metrics, introducing a new evaluation approach. They measured differences between topics in different sets of documents and examined how well automated metrics aligned with human judgment, particularly for common topics. Marani et al. [26] focused on enhancing topic model stability by reviewing various methods to measure and improve it. They stressed the importance of considering stability and quality when evaluating topic models. Harrando et al. [27] conducted a comparative analysis of nine popular topic modeling techniques, shedding light on issues in standard evaluation methods [35].

In particular, LDA and NMF exhibited exceptional capabilities in producing diverse and meaningful topic outputs. These findings underscore the prominence of NMF, especially in handling short text data. NMF's superiority suggests its efficacy in addressing the inherent challenges of brevity and noise often found in short texts.

One notable aspect of this research is the focused use of a limited set of standard metrics for evaluation. Rather than delving into an extensive list of metrics, this work highlights the effectiveness of using these fundamental measures to assess topic models.

The suite of experiments in this research work evaluates NMF [14] and variants of LDA in terms of implementation: online Variational Bayes inference (LDA-VB) [28] and inference using Collapsed Gibbs Sampling (LDA-CGS) [16]. These models are selected owing to their popularity and effectiveness across various works in the literature. However, this work aims to identify a topic model that performs effectively and efficiently, irrespective of the text length. The following section focuses on conducting comprehensive experiments to evaluate and compare these prominent topic modeling methods.

III. Experiments

The experimental design provides valuable insights and empirical evidence to draw informed conclusions about the effectiveness and efficiency of chosen topic models. This work uses public datasets to comprehensively compare NMF, LDA-VB, and LDA-CGS-based schemes for short and long text topic mining. All the experiments are conducted on a workstation using Python 3 Google compute engine backend.

A. Datasets

Many datasets exist to assess topic models, showcasing notable variations in corpus size, document length, topic intricacy, and noise levels. The choice of the dataset can profoundly impact the results. Hence, selecting a suitable topic model that aligns with the dataset's attributes is paramount. In the experiment's suite, short and long text datasets are utilized, with additional known associations of labels/topics to assess the performance of model-inferred topics for actual ground truth labels. Short text datasets involve caption datasets, Conceptual Captions, and Wider Captions datasets, comprising captions extracted from the web across various images. The Conceptual Captions [8] dataset is a recently proposed dataset for image captioning comprising instances across multiple categories. Owing to resource constraints, a subset of 10,000 captions with 40 diverse labels were selected. The Wider Captions [9] dataset is another captioning dataset comprising over 50,000 images of events and diverse actions; a sample of 10,000 instances is evaluated. The seminal datasets commonly utilized for topic model evaluation are employed for long text datasets, namely the 20 Newsgroups [10] and the Web of Science [11] datasets. Table I shows that choosing these datasets with varying values of optimal topics and characteristics in terms of document instance sizes and topics allows for a more general evaluation of the models. It can aid in determining the most appropriate model for the task.

TABLE I. DATASETS USED IN EXPERIMENT

Datasets	Documents	Categories/ Topics' K'
Conceptual Captions	10,000/3M	40
Wider Captions	10,000/50,000	61
20-Newsgroups	18,000	20
Web of Science	11697	35

B. Evaluation Metrics for Assessing Topic Models

Implementing topic models necessitates making several critical design choices, such as selecting an appropriate algorithm, inference method, model parametrization, and determining the optimal number of topics to uncover. To effectively streamline the process of making these design choices, it becomes imperative to establish a singular, overarching criterion for assessing quality, with accuracy being the foremost consideration. While other factors, such as computational complexity and processing speed, are certainly pertinent, accuracy is paramount to achieving clustering that most faithfully mirrors real-world patterns. Furthermore, the selection of the topic model can vary significantly based on the specific application, and it may yield different outcomes across multiple runs on the same dataset.

Topic modeling evaluation can follow two paths. One involves assessing the internal properties of the clustering result [6] and examining elements like topic-document assignments or topic descriptors corresponding to internal evaluation metrics. These internal metrics scrutinize structural aspects of clusters, such as their separation, without relying on additional input data. However, their quality assessment may not align with human perception, making them the primary choice when a definitive knowledge structure for text clustering is absent. The alternative approach [29] entails comparing clustering results with external knowledge sources, often termed ground truth, which typically takes the form of a pre-defined classification. This classification is often manually assigned and is rooted in human perceptions and the expertise of raters. This approach is known as external evaluation metrics. This research work utilizes internal and external evaluation metrics.

1) Internal evaluation metrics: The internal evaluation metrics can be broadly categorized as Topic Classification, Topic Significance, Topic Coherence, Topic Diversity, and Topic Similarity. Notably, Topic Stability, although an internal measure, is not included in this research as it does not serve as a quality criterion. Instead, it is considered a desirable property for algorithms incorporating stochastic elements [15]. These metrics assess inferred topics' quality, similarity, coherence, divergence, and perplexity. All experiments use the same set of evaluation metrics: internal metrics comprising of diversity [30] and KL-divergence [31] for topic diversity metrics, C_{UMass}, C_V, C_{NPMI}, C_{UCI}, WE-Pairwise and WE-Centroid [30] for topic coherence metrics [32], Jaccard similarity for topic similarity, KL-Uniform, KL-Vacuous, and KL-Background [33] for topic significance, and precision, recall, F1-score, and accuracy [34] for topic-based classification.

Topic Classification Metrics [34] assess document classifier performance using the learned document-topic distribution. This distribution is a K-dimensional representation for training the classifier to predict document classes. Subsequently, the classifier's performance is evaluated using key metrics, including precision, recall, and the F1-Score, which is the harmonic mean of precision, recall, and accuracy. These metrics are calculated using the Formulas in (1) to (4).

$$Precision_i = \frac{c_{ii}}{\sum_{j=1}^{n_c} c_{ji}} \quad (1)$$

$$Recall_i = \frac{c_{ii}}{\sum_{j=1}^{n_c} c_{ij}} \quad (2)$$

$$F1 - Score_i = \frac{2\ X\ Precision_i\ X\ Recall_i}{Precision_i + Recall_i} \quad (3)$$

$$\text{Accuracy} = \frac{\sum_{j=1}^{n_c} c_{ii}}{\sum_{i=1}^{n_c}\sum_{j=1}^{n_c} c_{ij}} \quad (4)$$

where, n_c is the number of classes, and Cij is the confusion matrix.

Topic Significance Metrics [33] play a crucial role in gauging the relevance and importance of topics, with a specific emphasis on scrutinizing both document-topic and topic-word distributions to discern and assess the significance of individual topics [15]. These metrics encompass the following key components:

a) KL-Uniform: This metric compares the topic and W-Uniform distribution using KL-Divergence. The underlying assumption is that genuine topics should be characterized by concisely selecting highly relevant words.

b) KL-Vacuous: In this case, the metric involves evaluating the topic distribution about the W-Vacuous distribution through KL-Divergence. The expectation is that authentic topics should exhibit distinct characteristics compared to a distribution that combines various elements from the sample set.

c) KL-Background: This metric assesses the topic distribution vis-à-vis the W-Background distribution, again utilizing KL-Divergence. The premise here is that genuine topics should only be present in a subset of the documents within the corpus and should not be predominantly composed of background noise.

Topic Coherence Metrics [32] assess how well the top-k words in a topic relate to each other, indicating topic interpretability. This process involves segmentation, probability estimation, confirmation, and aggregation. Standard coherence metrics, detailed in Formulas (5) to (10), are commonly used in the literature. Researchers in past studies have commonly used perplexity or held-out likelihood to evaluate models when comparing different topic numbers 'k.' However, it's crucial to recognize that while perplexity is helpful for this comparison, it primarily assesses predictive performance, not the exploratory goals of topic modeling [16].

$$C_{UCI} = \frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} log\left(\frac{P(w_i,w_j)+\epsilon}{P(w_j)P(w_i)}\right)}{\frac{N(N-1)}{2}} \qquad (5)$$

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

$$C_{NPMI} = \frac{\sum_{i=2}^{N}\sum_{j=i+1}^{N} NPMI(w_i,w_j)}{\frac{N(N-1)}{2}} \qquad (7)$$

$$\underset{v\ m,\gamma}{\rightarrow} (W') = \left\{\sum_{w_i \epsilon W'} m(w_i, w_j)^{\gamma}\right\}_{j=1\ldots|W|} \qquad (8)$$

$$C_V = \frac{\sum_{i=1}^{|W|} \vec{v}_{NMPI,1}(W')_i \cdot \vec{v}_{NMPI,1}(W^*)_i}{\left\|\vec{v}_{NMPI,1}(W')\right\|_2 \left\|\vec{v}_{NMPI,1}(W^*)\right\|_2} \qquad (9)$$

$$\mathrm{NPMI}(\mathrm{w_i}, \mathrm{w_j}) = \frac{\log\left(\frac{\mathrm{P}(\mathrm{w_i},\mathrm{w_j})+\epsilon}{\mathrm{P}(\mathrm{w_j})\mathrm{P}(\mathrm{w_i})}\right)}{-\log(\mathrm{P}(\mathrm{w_i},\mathrm{w_j})+\epsilon)} \qquad (10)$$

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
$\mathrm{w_i}, \mathrm{w_j}$	words (specific to a topic)
$\mathrm{P(w_i)P(w_j)}$	word probabilities
$\mathrm{P(w_i, w_j)}$	joint Probability of observing words $\mathrm{w_i}, \mathrm{w_j}$
$\underset{v\ m,\gamma}{\rightarrow} (W')$	context vector for words in W', direct confirmation measure m and power γ
ϵ	epsilon for avoiding indeterminate log (0)

In addition to conventional coherence metrics, the research literature has introduced coherence measures for individual topics using word embeddings with the emergence of distributed word representations. These metrics [30], [32] are calculated through pairwise or centroid-based methods, as elaborated in the provided from Formula (11) to Formula (14).

$$W_k = \left\{w_{ki} \mid w_{ki} \in t_k, i \in argsort\ \left\{ P(w_{kj}) \,\middle|\, w_{kj} \in t_k \right\}[: n_{max}]\ \right\}\# \qquad (11)$$

$$WE - Pairwise_k = \left(\frac{2}{|W_k|(|W_k|-1)} \sum_{i=1}^{|W_k|}\sum_{j=i+1}^{|W_k|} \frac{(e(w_i))^T e(w_j)}{\|e(w_i)\|\|e(w_j)\|}\right);\ w_i, w_j \in W_k\# \qquad (12)$$

$$e(w_k^*) = \frac{1}{|W_k|}\sum_{w \in W_k} e(w)\ \# \qquad (13)$$

$$WE - Centroid_k = \left(\frac{1}{|W_k|} \sum_{i=1}^{|W_k|} \frac{(e(w_i))^T e(w_k^*)}{\|e(w_i)\|\|e(w_k^*)\|}\right);\ w_i \in W_k\# \qquad (14)$$

where,

e(w_i)	Word embedding of word i
n	Maximum number of words per topic in consideration
W_k	Multiset of top n words (in terms of Probability) for topic k
$e(w_k^*)$	Centroid word embedding for the set W_k

Diversity Metrics [30] quantify the variation among the top k-words within a topic, focusing on identifying redundancies by evaluating the recurrence of words. These metrics employ a Symmetric KL-Divergence measure applied to normalized document-topic and topic-word distributions. The primary objective is to assess the diversity within the generated document-topic and topic-word distributions, emphasizing their variability [31]. The mathematical expressions for these diversity metrics, namely KL-Divergence and Topic Diversity, are provided in Formula (15) and Formula (16).

$$\mathrm{KL}(\mathrm{R_{l1}} \parallel \mathrm{R_{l2}}) = \sum_{\mathrm{i=1}}^{\mathrm{T}} \mathrm{R_{l1}} * \log\left(\frac{\mathrm{R_{i1}(i)}}{\mathrm{R_{i2}(i)}}\right) \qquad (15)$$

$$\text{Topic Diversity} = \left(\frac{1}{\mathrm{K}}\sum_{\mathrm{k=1}}^{\mathrm{K}} \frac{|\{\mathrm{w_{ki}}|\mathrm{w_{ki}} \in \mathrm{t_k}, \mathrm{i} \in \text{argsort}\{\mathrm{P(w_{ki})}|\mathrm{w_{ki}} \in \mathrm{t_k}\}[:\mathrm{n_{max}}]\}|}{\mathrm{n_{max}}}\right) \times 100 \qquad (16)$$

where,

K	number of Topics
$\mathrm{t_k}$	topic k
$\mathrm{w_{ki}}$	word i of topic k
$\mathrm{P(w_{ki})}$	probability of word i in topic k as per the topic-word distribution
$\mathrm{n_{max}}$	maximum number of words per topic in consideration

Topic Similarity Metrics [33] come in lexical and semantic forms. Lexical similarity deals with shared word sequences or structures, while semantic similarity relates to shared meaning. Cosine similarity evaluates text similarity by representing documents as term vectors and measuring it as the cosine of the angle between these vectors. Jaccard similarity calculates similarity based on the ratio of shared terms to total unique terms in both texts [3]. Jaccard's similarity [33] is computed to compare topics (Topic A and Topic B) using the Formula in (17).

$$J(Topic\ A, Topic\ B) = \frac{|TopicA \cap TopicB|}{|TopicA \cup TopicB|} \qquad (17)$$

where,

$TopicA \cap Topic\ B$	shared words in both topics
$TopicA \cup Topic\ B$	all unique words in both topics

2) External evaluation metrics: On the other hand, external metrics assess topics' performance in terms of classification and clustering of documents based on topic association. The study computed the Adjusted RAND Index (ARI) [29] and the Adjusted Mutual Information (AMI) [35]. While one external clustering metric would typically suffice, both are presented here to compare with findings from other research endeavors. Consider clustering documents as a series of pairwise decisions. If two documents fall in the same class and cluster, or both in distinct classes and clusters, the choice is regarded as accurate; otherwise, it is false. The Rand index calculates the percentage of correct decisions. The adjusted Rand index is the corrected-for-chance version of the Rand index, with an expected value of 0 and a maximum value of 1 for an exact match. On the other hand, the AMI takes a value of 1 when the two partitions are identical and 0 when the MI between two partitions equals the value expected due to chance alone. The mathematical formulation for ARI and AMI is provided in Formula (18) and Formula (19), respectively.

$$\mathrm{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}} \tag{18}$$

where,

n_{ij},a_i,b_j are values from the contingency table where each entry n_{ij} denotes the number of objects common in clusters.

$$AMI(U,V) = \frac{MI(U,V) - E\{MI(U,V)\}}{max\{H(U),H(V)\} - E\{MI(U,V)\}} \tag{19}$$

where,

$MI(U,V)$	mutual information between two clusters
$E\{MI(U,V)\}$	expected mutual information between two clusters
$H(U), H(V)$	entropy associated with clusters U and V, respectively

By scrutinizing a range of facets related to the identified topics using internal and external metrics, it becomes feasible to examine the strengths and weaknesses of different topic models in various dimensions or aspects, as discussed in the next section.

C. Experimental Setup

The process of identifying the most suitable topic model adheres to a conventional evaluation approach, encompassing the subsequent stages, as shown in Fig. 4. The topic models selected for consideration were trained separately over each choice of the dataset, followed by an evaluation of the performance of the topics learned by computation of internal and external metrics to determine the most effective topic model across datasets with diverse characteristics. Each data set is preprocessed to eliminate irregular word forms across documents. The topic models were trained over the preprocessed datasets with the best possible choices of hyperparameters: the number of topics being the optimum for the dataset, along with other parameters selected appropriately (Dirichlet priors selected to be symmetric and as per the value of number of topics). Following the training of topic models, the models were evaluated using internal and external metrics covering various aspects of topic quality: coherence, significance, diversity, similarity, usability for classification, and clustering information. The process of training and evaluation was jointly performed by use of the OCTIS framework [5], which incorporates implementations of multiple topic models and provides a suite of diverse evaluation metrics for comparing and contrasting the same. The experiment was repeated for all datasets chosen for this work. Within each experiment, apart from listed internal metrics covering topic diversity, topic coherence, topic significance, topic classification, and topic similarity, external metrics covering the extent of information conveyed by topics for determining a suitable clustering were also employed to determine the efficacy of learned topics.

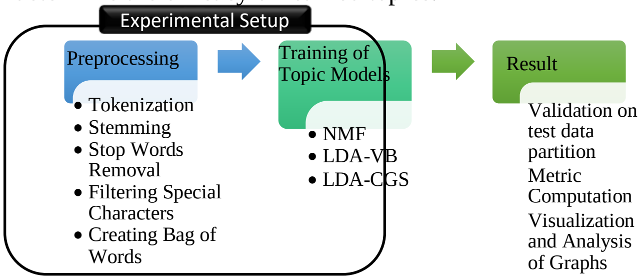


Fig. 4. Stages of experimental setup and result.

1) Data preprocessing: Each examined dataset has undergone preprocessing to eliminate irregular word forms across documents. This preprocessing encompasses standard procedures such as tokenization, stemming, removing stop words, and filtering special characters. Additionally, the preprocessing involves generating a bag-of-words representation that can be utilized in various topic models.

2) Training of topic models: Following the essential preprocessing phase, the dataset is divided into distinct training and testing segments, each allocated with specific roles and objectives. The training portion is the foundation for training the selected topic models, enabling them to gain insights and patterns from the provided data. On the other hand, the testing portion plays a pivotal role in the evaluation process, serving as an independent set of documents against which the models' performance is scrutinized and assessed.

During the evaluation phase, each chosen topic model is subjected to rigorous scrutiny over these training and testing datasets. This assessment entails meticulously examining their output in light of internal and external evaluation metrics. Internal evaluation metrics encompass criteria that gauge the coherence, diversity, and overall quality of the topics generated by the models. Topic classification metrics are another category of popular metrics employed to assess the quality of topic models by assessing the extent of how useful topics are as features for the classification of the source documents using standard classification techniques. Our experiments observed poor performance of the learned topics for classifying documents using an SVM classifier. Thus, we have included only a relative comparison based on the values highlighting the best topic model per dataset and metric. A possible explanation for the poor metric values is the sensitivity of the metrics to the choice of classifier and the corresponding hyperparameters. So, these metrics may not accurately represent the usefulness of topics for classifying documents. A thorough analysis for contrasting performance must also compare diverse classifiers to determine the overall effectiveness of topics learned by the different topic models for classification, which is beyond the scope of this study of contrasting topic models.

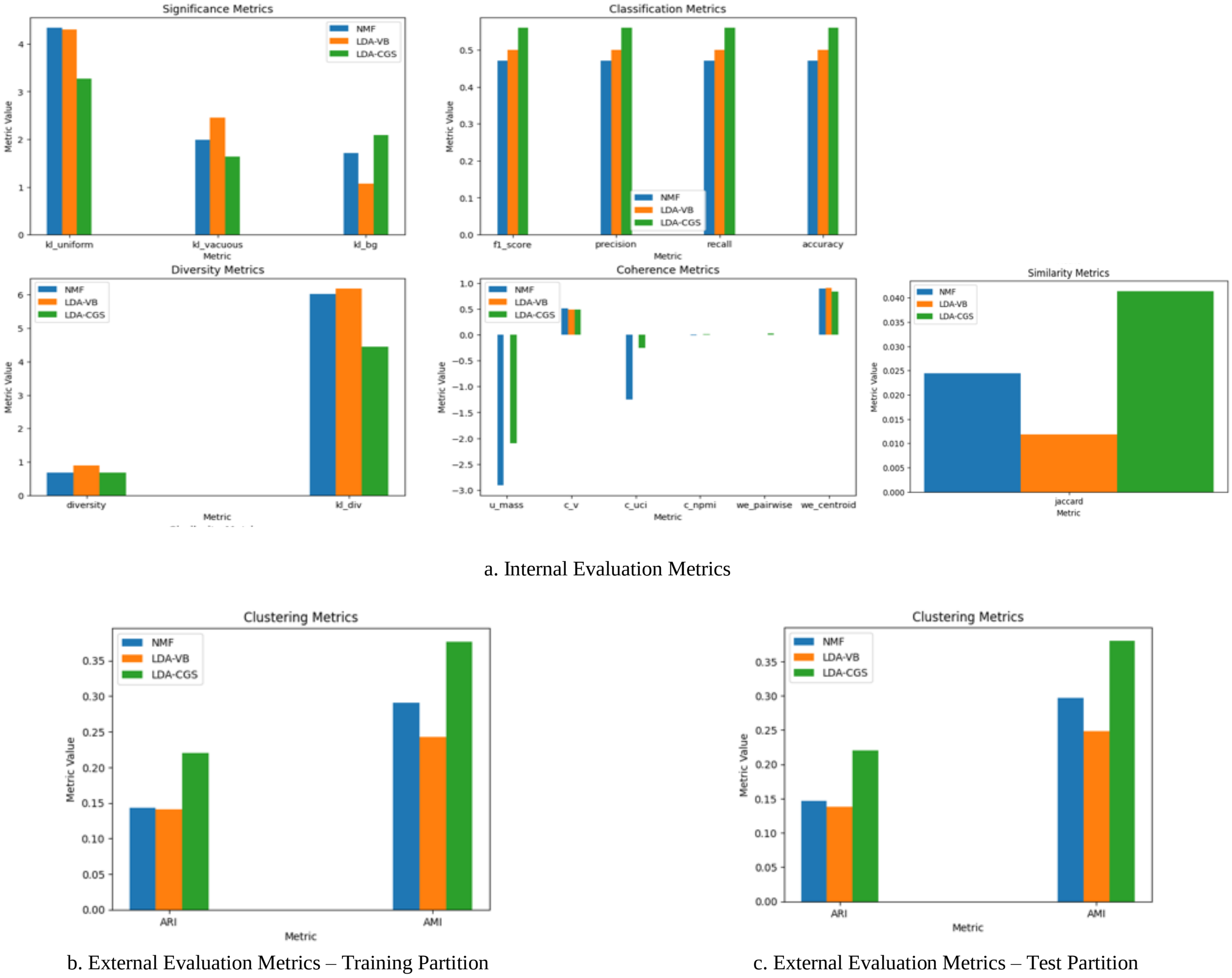
External evaluation metrics, on the other hand, pivot towards assessing the model's efficacy in practical applications. They evaluate the model's ability to categorize, classify, and cluster documents based on the topics it infers. This dual-pronged evaluation strategy, involving internal and external metrics, forms a comprehensive and well-rounded approach to ascertain the effectiveness and applicability of the topic models under scrutiny.

Once the models have been trained successfully, they are run on test data. The output is analyzed using the graphs discussed in the next section.

IV. Results and Discussion

A. Results over Short Text Datasets

Varying results regarding the most performant topic model across different internal metrics are observed across short text datasets. Regarding internal metrics, NMF is significantly more performant than LDA implementations across many metrics. Across external metrics, however, LDA-CGS outperforms both NMF and LDA-VB for both datasets.



a. Internal Evaluation Metrics

b. External Evaluation Metrics – Training Partition

c. External Evaluation Metrics – Test Partition

Fig. 5. Comparison of topic models over Internal and External Metrics for the Conceptual Captions dataset

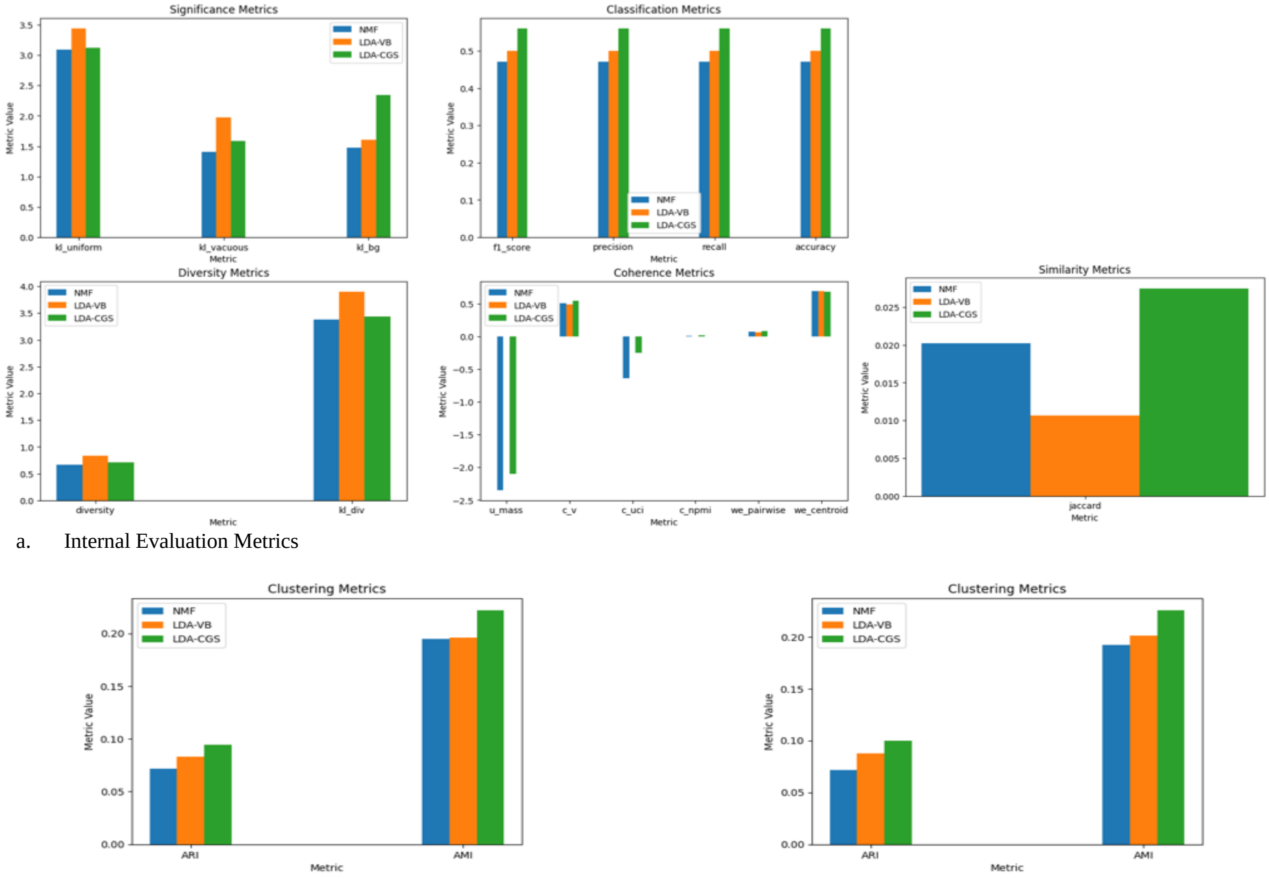
Fig. 5(a), 5(b), and 5(c) compare the performance of the three topic models over the Conceptual Captions dataset regarding internal and external evaluation metrics for training and test partitions, respectively. Across internal metrics, while there exists no consensus for the best topic models, NMF demonstrates high performance across a significant number of metrics. This behavior is prominent in topic classification metrics, where NMF outperforms both LDA variants by a minute margin. NMF also demonstrates leading performance across coherence metrics C_V and C_{NPMI} and topic significance compared to the Uniform distribution (KL-Uniform). Across other metrics, while NMF is not the best model, the difference between the metric values of NMF and that of the best model is minimal. Across the remaining metrics, LDA-CGS and LDA-

VB perform in contrast to one another across different metrics. While the topics of LDA-CGS are more coherent in terms of C_{UMass} and WE-Pairwise, more significant in terms of KL-Divergence for the Background distribution, the topics are more similar and less diverse than LDA-VB topics. Across external evaluation metrics, however, LDA-CGS is the best performant, demonstrating an ability to produce a prominent quality clustering based on inferred topics, with an agreement to the ground truth of the document labels.

The performance of the topic models in terms of internal and external metrics for the Wider Captions dataset over training and test partitions is compared in Fig. 6(a), 6(b), and 6(c), respectively. For the dataset, the behavior of NMF is much more prominent, with leading results across many metric categories. NMF outperforms both variants of LDA across topic significance (KL-Uniform), topic classification, topic similarity, and topic diversity when measured with KL-divergence. However, unlike prior observations, topics inferred by NMF for the dataset are not as coherent as those inferred by other models, and instead of NMF, LDA-CGS results in the most coherent topics as observed by the coherence values for all coherence metrics. However, the divergence and similarity of the LDA-CGS topics are still significantly lower than the other models. LDA-VB demonstrates comparable results to the most compelling topic model across most metrics and leads over other models only in terms of diversity and topic significance (KL-Vacuous). Across external metrics, a similar trend as the Conceptual Captions dataset is observed, with LDA-CGS significantly outperforming other topic models, demonstrating the reasonable effectiveness of LDA-CGS in inferring high-quality topics that produce high-quality clusters.

B. Results over Long Text Datasets

Across long text datasets, NMF is superseded by LDA models across internal and external evaluation metrics. It demonstrates the effectiveness of LDA in general; most corpora on which topic modeling is utilized comprise long text documents with multiple topics per document. LDA is significantly more effective than NMF, and short text document corpora are rare and application-specific.

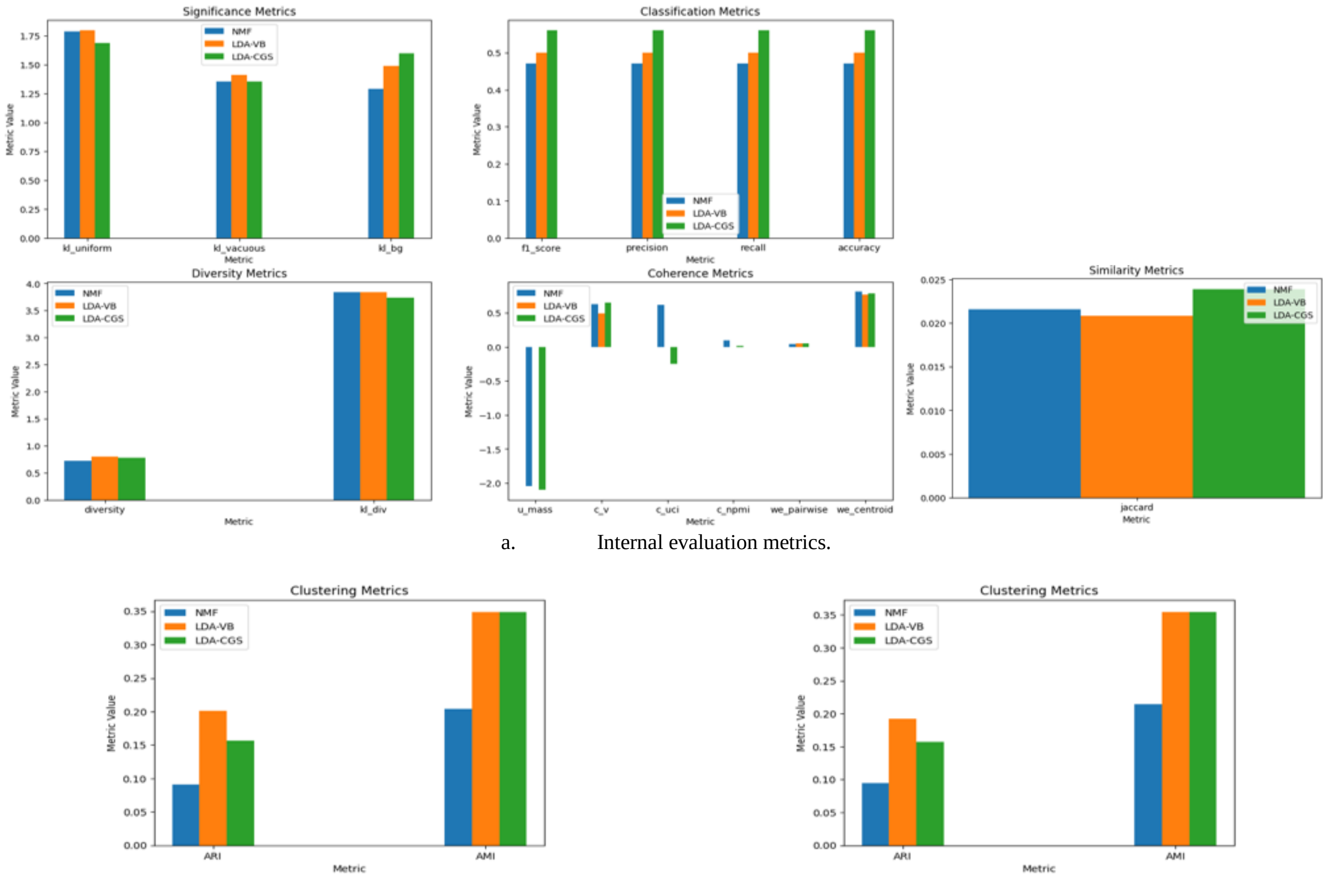


a. Internal Evaluation Metrics

b. External Evaluation Metrics – Training Partition

c. External Evaluation Metrics – Test Partition

Fig. 6. Comparison of topic models over internal and external metrics for the wider captions dataset.



a. Internal evaluation metrics.

b. External evaluation metrics – training partition.

c. External evaluation metrics – test partition.

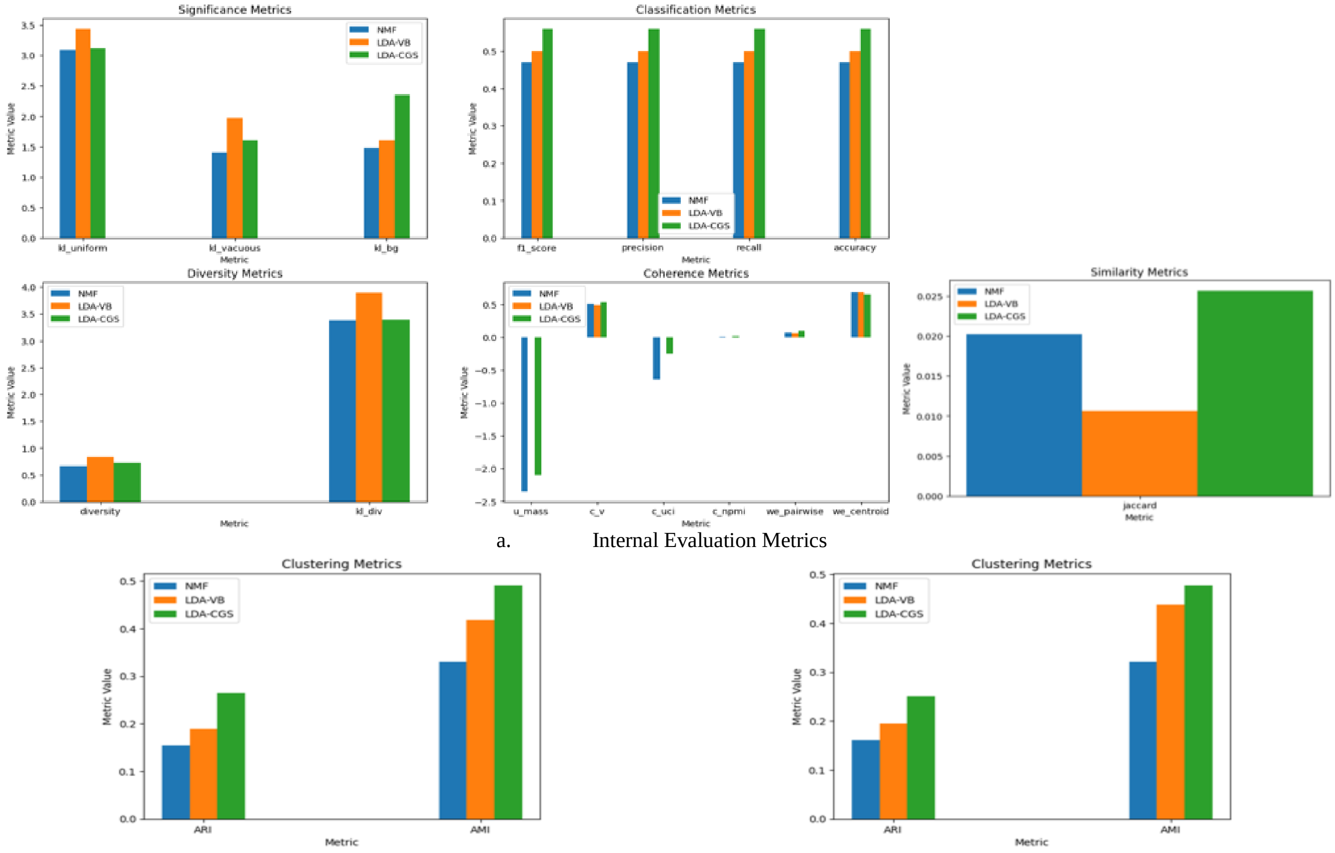
Fig. 7. Comparison of topic models over internal and external metrics for the 20-newsgroups dataset.

Fig. 7(a), 7(b), and 7(c) compare the performance of the three topic models over the 20-Newsgroups dataset regarding internal and external evaluation metrics over training and test partitions, respectively. Regarding internal evaluation metrics, the results regarding determining the best topic model are inconclusive as the metrics appear to disagree. However, across all metrics, NMF is outperformed by either variant of LDA. Classification and Diversity metrics prefer topics inferred by the variational Bayes implementation of LDA. In contrast, topics produced by LDA implemented with Collapsed Gibbs Sampling are found to be the most coherent and significant compared to the background distribution. Across external evaluation metrics, LDA-CGS outperforms LDA-VB in terms of AMI but not in terms of ARI scores. Both variants of LDA perform similarly for this dataset, and either can be preferred for inferring latent topics. It is noted that the difference across metric values for both LDA variants is insignificant, strengthening the assertion that either model is suitable for obtaining high-quality topics.

The comparison of the performance of topic models over the Web of Science dataset in terms of internal and external metrics over training and test partitions is demonstrated in Fig. 8(a), 8(b), and 8(c) respectively. Similar to the trend across other datasets, internal evaluation metrics do not demonstrate a unified consensus for the choice of the topic model. However, LDA variants significantly outperform NMF in all metrics. However, unlike 20-Newsgroups, LDA-CGS is found to be more performant than LDA-VB across classification and coherence metrics. The trend for significance and diversity metrics, however, is similar to the trend observed in 20-Newsgroups and other short text datasets, which indicates that LDA-CGS, in general, infers topics that are highly coherent and more significant than background information but are not significantly diverse in general. Further, the trend across external metrics is similar to the observations across short text datasets, with LDA-CGS outperforming other topic models, demonstrating the reasonable quality of LDA-CGS inferred topics and their effectiveness in clustering documents.

C. Discussion and Interpretation of Results

Based on the results, LDA-CGS is suitable for producing high-quality topics across corpora with diverse text characteristics. However, to improve the quality of topics specific to the application, LDA-CGS produces reasonable quality topics that demonstrate high performance over external clustering metrics and classification over documents, especially with long text data. Further, the topics have high coherence, divergence, significance, and reasonable similarity, making the algorithm a suitable default choice for most applications. The best performant topic model across each dataset and metric has been summarized in Table II.



a. Internal Evaluation Metrics

b. External evaluation metrics – training partition.

c. External evaluation metrics – test partition.

Fig. 8. Comparison of topic models over internal and external metrics for the web of science dataset.

TABLE II. COMPARISON OF TOPIC MODELS NMF, LDA-VB, AND LDA-CGS) ACROSS INTERNAL AND EXTERNAL EVALUATION METRICS SET

Evaluation Metrics		**Datasets**			
Metric Category	**Metric**	**Conceptual Captions**	**Wider Captions**	**20-Newsgroups**	**Web of Science (WOS11697)**
Topic Classification [9] [10]	**Precision**	NMF	NMF	LDA-VB	LDA-CGS
	Recall	NMF	NMF	LDA-VB	LDA-CGS
	F1-Score	NMF	NMF	LDA-VB	LDA-CGS
	Accuracy	NMF	NMF	LDA-VB	LDA-CGS
Topic Significance [9] [13]	**KL-Uniform**	NMF	NMF	LDA-VB	LDA-VB
	KL-Vacuous	LDA-VB	LDA-VB	LDA-VB	LDA-VB
	KL-Background	LDA-CGS	LDA-CGS	LDA-CGS	LDA-CGS
Topic Coherence [14] [20]	**C_U_Mass**	LDA-CGS	LDA-CGS	LDA-CGS	LDA-CGS
	C_UCI	LDA-VB	LDA-CGS	LDA-CGS	LDA-CGS
	C_NPMI	NMF	LDA-CGS	LDA-CGS	LDA-CGS
	C_V	NMF	LDA-CGS	LDA-CGS	LDA-CGS
	WE-Pairwise	LDA-CGS	LDA-CGS/ LDA-VB	LDA-CGS	LDA-CGS
	WE-Centroid	LDA-VB	NMF	NMF	LDA-VB/ NMF
Topic Diversity [21] [29]	**KL-Divergence**	LDA-VB	NMF	LDA-VB	LDA-VB
	Diversity	LDA-VB	LDA-VB	LDA-VB	LDA-VB
Topic Similarity [33]	**Jaccard Score**	LDA-VB	NMF	LDA-VB	LDA-VB
External Topic / Clustering Quality [12] [32]	**Adjusted Rand Index (ARI)**	LDA-CGS	LDA-CGS	LDA-VB	LDA-CGS
	Adjusted Mutual Information (AMI)	LDA-CGS	LDA-CGS	LDA-CGS	LDA-CGS

V. CONCLUSION

Topic modeling techniques exhibit versatility in handling both short and long text data. While models like LDA excel with longer documents, approaches such as NMF demonstrate efficiency in understanding the context within shorter texts. This research performs empirical analysis of state-of-the-art topic models through statistical metrics in the context of varied text length data to determine the best model irrespective of text length. Based on the experimental results, LDA-CGS produces high quality topics over external clustering metrics for both long and short text data. The topics produced by LDA-CGS have high coherence, divergence, significance, and similarity, making it a suitable choice for most datasets.

Future research directions could explore hybrid approaches that combine the strengths of multiple algorithms to enhance topic modeling performance across different text lengths. The results show that NMF is still a strong contender for short text data, and a hybrid model may show much better performance for varying text length datasets. These models can also be applied on Twitter or image caption datasets to discover relevant information for the classification process.

Conflict of Interest: The authors declare no competing interests and confirm that neither the manuscript nor any parts of its content are currently under consideration or published in another journal.

REFERENCES

[1] B. A. H. Murshed, S. Mallappa, J. Abawajy, M. A. N. Saif, H. D. E. Al-ariki and H. M. Abdulwahab, Short-text topic modelling approaches in the context of big data: taxonomy, survey, and analysis, vol. 56, Springer Science and Business Media LLC, 2022, p. 5133–5260.

[2] S. Sbalchiero and M. Eder, Topic modeling, long texts and the best number of topics. Some Problems and solutions, vol. 54, Springer Science and Business Media LLC, 2020, p. 1095–1108.

[3] S. Athukorala and W. Mohotti, An effective short-text topic modelling with neighbourhood assistance-driven NMF in Twitter, vol. 12, Springer Science and Business Media LLC, 2022.

[4] R. Egger and J. Yu, A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts, vol. 7, Frontiers Media SA, 2022.

[5] S. Terragni, E. Fersini, B. G. Galuzzi, P. Tropeano and A. Candelieri, OCTIS: Comparing and Optimizing Topic models is Simple!, Association for Computational Linguistics, 2021.

[6] M. Rüdiger, D. Antons, A. M. Joshi and T.-O. Salge, Topic modeling revisited: New evidence on algorithm performance and quality metrics, vol. 17, D. R. Amancio, Ed., Public Library of Science (PLoS), 2022, p. e0266325.

[7] R. Albalawi, T. H. Yeap and M. Benyoucef, Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis, vol. 3, Frontiers Media SA, 2020.

[8] P. Sharma, N. Ding, S. Goodman and R. Soricut, Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning, Association for Computational Linguistics, 2018.

[9] Y. Xiong, K. Zhu, D. Lin and X. Tang, Recognize complex events from static images by fusing deep channels, IEEE, 2015.

[10] K. Lang, NewsWeeder: Learning to Filter Netnews, Elsevier, 1995, p. 331–339.

[11] K. Kowsari, Web of Science Dataset, Mendeley, 2018.

[12] A. Goyal and I. Kashyap, Latent Dirichlet Allocation - An approach for topic discovery, IEEE, 2022.

[13] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer and R. Harshman, Indexing by latent semantic analysis, vol. 41, Wiley, 1990, p. 391–407.

[14] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," Nature, vol. 401, p. 788–791, October 1999.

[15] D. Rugeles, Z. Hai, J. F. Carmona, M. Dash and G. Cong, Improving the Inference of Topic Models via Infinite Latent State Replications, arXiv, 2023.

[16] Y. Chen, H. Zhang, R. Liu, Z. Ye and J. Lin, Experimental explorations on short-text topic mining between LDA and NMF based Schemes, vol. 163, Elsevier BV, 2019, p. 1–13.

[17] K. Devarajan, Nonnegative Matrix Factorization: An Analytical and Interpretive Tool in Computational Biology, vol. 4, B. Bryant, Ed., Public Library of Science (PLoS), 2008, p. e1000029.

[18] Z.-Y. Zhang, Nonnegative Matrix Factorization: Models, Algorithms and Applications, Springer Berlin Heidelberg, 2012, p. 99–134.

[19] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li and L. Zhao, Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey, vol. 78, Springer Science and Business Media LLC, 2018, p. 15169–15211.

[20] M. D. Hoffman, D. M. Blei, C. Wang and J. Paisley, Stochastic Variational Inference, vol. 14, 2013, p. 1303–1347.

[21] A. U. Rehman, Z. Rehman, J. Akram, W. Ali, M. A. Shah and M. Salman, Statistical Topic Modeling for Urdu Text Articles, IEEE, 2018.

[22] Y. W. Teh, D. Newman and M. Welling, "A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation," in Advances in Neural Information Processing Systems 19, The MIT Press, 2007, p. 1353–1360.

[23] M. O. Ajinaja, A. O. Adetunmbi, C. C. Ugwu and O. S. Popoola, "Semantic similarity measure for topic modeling using latent Dirichlet allocation and collapsed Gibbs sampling," Iran Journal of Computer Science, vol. 6, p. 81–94, November 2022.

[24] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth and M. Welling, Fast collapsed gibbs sampling for latent dirichlet allocation, ACM, 2008.

[25] J. P. Lim and H. Lauw, "Large-Scale Correlation Analysis of Automated Metrics for Topic Models," in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023.

[26] A. Hosseiny Marani and E. P. S. Baumer, "A Review of Stability in Topic Modeling: Metrics for Assessing and Techniques for Improving Stability," ACM Computing Surveys, vol. 56, p. 1–32, November 2023.

[27] I. Harrando, P. Lisena and R. Troncy, "Apples to Apples: A Systematic Evaluation of Topic Models," in Proceedings of the Conference Recent Advances in Natural Language Processing - Deep Learning for Natural Language Processing Methods and Applications, 2021.

[28] M. Hoffman, F. Bach and D. Blei, Online learning for latent dirichlet allocation, vol. 23, 2010.

[29] J. Han, M. Kamber and J. Pei, Data mining concepts and techniques third edition, 2012.

[30] A. B. Dieng, F. J. R. Ruiz and D. M. Blei, Topic Modeling in Embedding Spaces, arXiv, 2019.

[31] F. Bianchi, S. Terragni and D. Hovy, Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence, arXiv, 2020.

[32] M. Röder, A. Both and A. Hinneburg, Exploring the Space of Topic Coherence Measures, ACM, 2015.

[33] L. AlSumait, D. Barbará, J. Gentle and C. Domeniconi, Topic Significance Ranking of LDA Generative Models, Springer Berlin Heidelberg, 2009, p. 67–82.

[34] X.-H. Phan, L.-M. Nguyen and S. Horiguchi, "Learning to classify short and sparse text {\&}amp$\mathsemicolon$ web with hidden topics from large-scale data collections," in Proceedings of the 17th international conference on World Wide Web, 2008.

[35] S. Romano, N. X. Vinh, J. Bailey and K. Verspoor, Adjusting for Chance Clustering Comparison Measures, arXiv, 2015.