Combining BERT with LDA: Improved Topic Modeling in Bengali Language

Pintu Chandra Paul, Maqsudur Rahman*, Amena Begum, Md. Tofael Ahmed,
Dulal Chakraborty and Md. Saifur Rahman

***Abstract*—Topic modeling is a widely used technique for extracting hidden patterns from unlabeled text data, facilitating various functionalities such as document organization, content suggestion, and information retrieval. While traditionally applied to English-language text, topic modeling has recently gained traction in other languages, including Bengali, driven by the growing availability of Bengali content online. While recent research has applied some topic modeling methods to Bengali, their effectiveness in terms of performance has not been thoroughly validated. This paper introduces BERT-LDA, a hybrid approach to topic modeling, applied to a Bengali news corpus comprising articles from various categories collected from online Bengali news portals. Latent Dirichlet Allocation (LDA) is a probabilistic model that represents each document as a mixture of topics, whereas BERT-LDA leverages the semantic richness of BERT's contextual embeddings combined with LDA's robust topic modeling capabilities. By integrating the strengths of both methods, our approach seeks to enhance the performance of topic modeling for Bengali text. Experimental results demonstrate that the proposed BERT-LDA model consistently outperforms traditional topic modeling techniques across various evaluation metrics, offering a significant improvement in extracting meaningful insights from Bengali text data.**

***Index Terms*—BERT-LDA, Sentence-BERT, Topic modeling, Bengali news corpus.**

I. Introduction

Every day, a substantial volume of data is generated online as people increasingly prioritize digital interactions over offline activities. The unstructured nature of this data presents significant challenges for researchers, particularly in the realm of data analysis. While a large portion of internet data is in English, owing to its widespread accessibility, there has been a noticeable rise in data generated in resource-constrained languages like Bengali. This increase is attributed to the extensive use of social media platforms, the proliferation of blogging sites, and the growing presence of online Bengali news portals. To effectively extract valuable insights from this huge amount of unstructured data, it is required to be organized in such ways that it simplifies the information retrieval for efficient decision-making and problem-solving. To achieve this, it is necessary to understand the relationships between words within a document and uncover the underlying patterns among them. Topic modeling serves as a method for identifying latent abstract information within large datasets by uncovering hidden thematic relationships among words. It can also be termed as a text-mining approach for finding patterns in textual documents [1]. Topic modeling has a wide range of applications, including the analysis of news trends and research trends [2] in specific domains. Most topic modeling tasks are done by clustering a group of documents based on some common textual patterns. While numerous robust topic modeling tools exist for English-language texts, similar tools for languages like Bengali remain underdeveloped. However, recent advancements have significantly improved Bengali topic modeling capabilities.

As a generative model, topic modeling can be characterized as a probability distribution of topics because it uses probabilistic measures to generate connections between documents [3]. To cluster a set of similar documents, various topic modeling methods have been proposed by scholars, including Latent Dirichlet Allocation (LDA) [4], Latent Semantic Indexing (LSI) [5], and Hierarchical Dirichlet Process (HDP) [6]. To implement these methods, it is necessary to construct a Document Term Matrix (DTM) that quantifies the occurrences of terms within each document. The DTM subsequently serves as input for the topic modeling algorithms. LDA, the most widely utilized probabilistic topic modeling algorithm, identifies the underlying topics present in a text corpus. LSI is another prominent topic modeling algorithm that emphasizes capturing the semantic relationships between words in textual data. HDP, a Bayesian nonparametric approach, automatically infers the optimal number of topics within a dataset. In 2018, Google introduced Bidirectional Encoder Representations from Transformers (BERT), a sophisticated pre-trained model designed to enhance the contextual understanding of unlabeled text across a broad range of tasks. BERT's ability to generate precise contextual word embeddings has made it a cornerstone in nearly all subfields of Natural Language

Manuscript received November 2, 2023; revised March 26, 2025. This work was supported by the research cell of Comilla University, Bangladesh.

Pintu Chandra Paul is lecturer of Information and Communication Technology, Comilla University, Cumilla-3506, Bangladesh. (e-mail: pintu@cou.ac.bd).

Maqsudur Rahman is assistant professor of Information and Communication Technology, Comilla University, Cumilla-3506, Bangladesh. (Corresponding author, e-mail: mrrajon@cou.ac.bd).

Amena Begum is assistant professor of Information and Communication Technology, Comilla University, Cumilla-3506, Bangladesh. (e-mail: amenaict@cou.ac.bd).

Md. Tofael Ahmed is associate professor of Information and Communication Technology, Comilla University, Cumilla-3506, Bangladesh. (e-mail: tofael@cou.ac.bd).

Dulal Chakraborty is associate professor of Information and Communication Technology, Comilla University, Cumilla-3506, Bangladesh. (e-mail: dulal.ict.cou@gmail.com).

Md. Saifur Rahman is associate professor of Information and Communication Technology, Comilla University, Cumilla-3506, Bangladesh. (e-mail: saifurice@cou.ac.bd).

Processing (NLP), including text mining [7] for sentiment analysis and various other applications. In several application areas of natural language processing, BERT models have demonstrated superiority over competing models [8]. Its ability to capture nuanced word relationships and context has led to its widespread adoption across NLP subfields. This research focuses on incorporating the potentials of BERT and LDA for improving topic modeling in Bengali by proposing a novel hybrid approach that combines the strengths of both techniques. Although online resources in Bengali are now readily accessible, working with these resources poses several challenges due to limited datasets and the complexity of diverse grammatical rules. Furthermore, the process of vectorizing and training a substantial amount of Bengali documents is inherently difficult. Despite these challenges, researchers are actively engaged in enhancing the efficiency and accuracy of natural language processing tasks in the Bengali language. This research investigates the effectiveness of the BERT-LDA model in improving the performance of topic modeling in Bengali. The key contributions of our study is to,

(a) Create a novel dataset of Bengali news articles gathered from various online news portals to effectively utilize topic modeling techniques in the Bengali language.
(b) Evaluate the performance of the BERT-LDA model based on different evaluation metrics such as Silhouette Score, Coherence Score, Jaccard Similarity, Davies-Bouldin Index (DBI), Calinski-Harabasz Index (CHI), Topic Diversity, as well as Homogeneity and Completeness Scores.
(c) Investigate the effectiveness of the BERT-LDA model compared to traditional topic modeling methods such as LDA, LSI, and HDP on the novel dataset.
(d) Present a detailed comparative analysis of state-of-the-art topic modeling methods on the dataset.

The structure of the paper is organized as follows: Section II provides an overview of related works on topic modeling. In Section III, we introduce the methodologies employed in our study. Section IV presents the experimental results and discusses the findings. Finally, Section V concludes the study and offers suggestions for future research directions.

II. Related Works

Numerous studies on topic modeling have been conducted in recent years across various domains of interest. Currently, social media data is increasingly utilized for document clustering, particularly through the application of a multi-objective genetic algorithm [9] for analyzing twitter interactions. The majority of research in this area has been conducted in the English language, primarily due to the abundance of available materials on the internet. However, there has been a notable increase in research focused on the Bengali language because of the advancements in Bengali NLP.

P. C. Paul et al. [10] evaluated LDA and BERT-LDA models using a corpus of 51,016 news articles collected from popular Bengali online news portals. They have used an n-gram profile for creating Bag-of-Word for LDA and sentence-BERT embedding for BERT-LDA. They have achieved a coherence score of 0.63 for LDA and a coherence score of 0.66 for BERT-LDA. They have shown that BERT-LDA is contextually stronger than LDA in terms of document clustering.

M. Al Helal et al. [11] used LDA with Bigram to find the core topic of the Bengali news corpus, which has 7,134 news articles. They have used coherence measures to find the optimal number of topics. They have also explored cosine similarity measures using the Doc2Vec model along with LDA.

M. Hasan et al. [12] analyzed the performance of the LDA and lda2vec models using a dataset of 22,675 Bengali news documents. They have found a classification accuracy of 62.45% for LDA and 85.66% for lda2vec, and the lda2vec model outperformed LDA.

The authors of [13] used topic modeling to find topics of interest using historical newspaper data that was published from 1829 to 2008 in Texas. They have tried to evaluate the result of topic modeling from (MAchine Learning for LanguagE Toolkit) MALLET compared with a human historian expert. They have achieved 60% of accuracy in topic modeling.

K. M. Alam et al. [14] have analyzed topic modeling to find the news trend in Bangladeshi media. The outcome shows that their approach can be used to track media trends over a period of time. They have used a dataset containing 70,000 Bengali news articles. A labeled LDA model is used to evaluate the coherence score and find the highest coherence value for 6 as the number of topics.

S. H. Mohammed et al. [15] evaluated LDA and LSA topic modeling approaches for clustering similar topics in a group. They have used 300 text files of books and articles with their full content. For the performance analysis, coherence UCI and coherence UMass scores were used for both LDA and LSA. LDA achieved a coherence UCI score of 0.59, while LSA's score is 0.49 for 20 topics. The UMass score is -0.37 and -0.92 for LDA and LSA respectively.

M. Grootendorst [16] introduced a deep neural topic model named BERTopic that uses modified TF-IDF for clustering documents with similar topics. Three different datasets were used in that study. The author has compared the clustering performance of BERTopic with some other well-known topic modeling approaches like LDA, NMF, CTM, and Doc2Vec in terms of coherence NPMI score. It was found that BERTopic had a higher coherence score than the other four models for all datasets.

S. Palani et al. [17] implemented LDA and BERT topic modeling approaches for clustering microblog data based on sentiment analysis. They have used around 40,000 microblogs for the evaluation of the coherence equation, C_V, and silhouette score. The values of the coherence score are 0.50, 0.52, and 0.56 for LDA, BERT, and LDA+BERT respectively. The silhouette scores for BERT and LDA+BERT are 0.04 and 0.46.

S. S. Panigrahi et al. [18] implemented the Word2Vec model for clustering Hindi corpus using skip-gram and Continuous Bag-of-Word (CBoW) feature extraction methods. They have used a dataset of 21,681 unique

TABLE I: Summary of the Related Works

Reference and Year	Methods	Dataset Domain	Dataset Size	Findings
P. C. Paul et al. [**10**], 2022	LDA, BERT-LDA	Bengali news articles	51,016	Coherence score, • LDA: 0.63 • BERT-LDA: 0.66
M. Al Helal et al. [**11**], 2018	LDA, Doc2Vec	Bengali news articles	7,134	LDA performs better than Doc2Vec
M. Hasan et al. [**12**], 2019	LDA, lda2vec	Bengali news articles	22,675	Accuracy, • LDA: 62.45% • Lda2vec: 85.66%
T. Yang et al. [**13**], 2011	MALLET	English newspaper article	-	Accuracy 60%
K. M. Alam et al. [**14**], 2020	LDA	Bengali news article	70,000	Topic score 49.12
S. H. Mohammed et al. [**15**], 2020	LDA, LSA	English books and articles	300	Coherence score, • LDA: 0.59 • LSA: 0.49
M. Grootendorst [**16**], 2022	BERTopic, LDA, NMF, CTM, Doc2Vec	20 NewsGroups BBC news Trump's tweet	16,309 2,225 44,253	Coherence score, • LDA: 0.058 • NMF: 0.089 • CTM: 0.096 • Doc2Vec: 0.192 • BERTopic: 0.166
S. Palani et al. [**17**], 2021	LDA, BERT	English tweets	40,000	Coherence score, • LDA: 0.50 • BERT: 0.52 • LDA+BERT: 0.56
S. S. Panigrahi et al. [**18**], 2018	Word2Vec	Hindi Wikipedia data	21,681	12 cluster for hierarchical clustering.
S. K. Ray et al. [**19**], 2019	LSI, LDA, NMF	Hindi news articles	10,400	Coherence score, • LSI: 0.48 • LDA: 0.66 • NMF: 0.79
A. Abuzayed et al. [**20**], 2021	BERTopic, LDA, NMF	Arabic news articles	108,789	BERTopic performed better than LDA and NMF
L. George et al. [**21**], 2023	LDA, BERT, BERT-LDA	CORD-19 dataset	40,000	Silhouette score on UMAP, • LDA: 0.38 • BERT: 0.49 • BERT-LDA: 0.52
M. H. Asnawi et al. [**22**], 2023	CTM-MPNet	User-review dataset	15,000	Coherence score, • C_V: 0.7091 • C_{UCI}: -0.6407 • C_{NPMI}: 0.0752
M. C. Wijanto et al. [**23**], 2024	BERTopic, RoBERTa, DistilRoBERTa	Research articles	20,972	Coherence score, • BERTopic: 0.5412 • RoBERTa: 0.5554 • DistilRoBERTa: 0.4950

tokens created from 7.2GB of Wikipedia documents. Using hierarchical clustering, 12 clusters were found to be the optimum number of clusters in that study.

S. K. Ray et al. [19] applied LSI, LDA, and NMF algorithms for topic modeling on the Hindi news corpus. The corpus consists of 10,400 documents. The coherence C_V score and perplexity were used for the performance analysis in that study. The coherence scores were 0.48, 0.66, and 0.79 for LSI, LDA, and NMF respectively.

A. Abuzayed et al. [20] have experimented with the BERTopic model on an Arabic dataset. The dataset was contained a total of 108,789 Modern Standard Arabic (MSA) documents. The BERTopic model has performed better than LDA and NMF based on coherence NPMI score.

L. George et al. [21] proposed a hybrid model combining BERT with LDA which focuses on enhancing coherence and interpretability by integrating clustering techniques, such as k-means, along with dimensionality reduction methods like PCA, t-SNE, and UMAP. They have evaluated their model using silhouette score on the COVID-19 Open Research Dataset (CORD-19). They have found that UMAP based dimensionality reduction outperformed other methods with the silhouette score of 0.52.

M. H. Asnawi et al. [22] presented a combination of the Contextualized Topic Model (CTM) and the Masked and Permuted Pre-training for Language Understanding (MPNet) model to improve the analysis of user feedback data. The approach starts by determining the optimal number of topics on a user-review dataset of different apps from the app store. By optimizing hyperparameters, they have found that the model outperforms the traditional topic models with a coherence C_V score of 0.7091. Their emphasis was on extracting meaningful insights from user feedback for developers to prioritize improvements.

M. C. Wijanto et al. [23] proposed topic modeling for scientific articles, focusing on BERT-based approaches with optimized hyper parameters. Their experiments were run across different combination of word embedding, dimension reduction techniques and clustering methods. They have achieved superior results over traditional methods like LDA with the combination of

RoBERTa for word embedding, PCA for dimension reduction and K-Means for clustering. A detailed summary of our related works is given in TABLE I.

III. Methodology

This section explains the detailed information about our proposed methodologies as shown in Fig. 1.

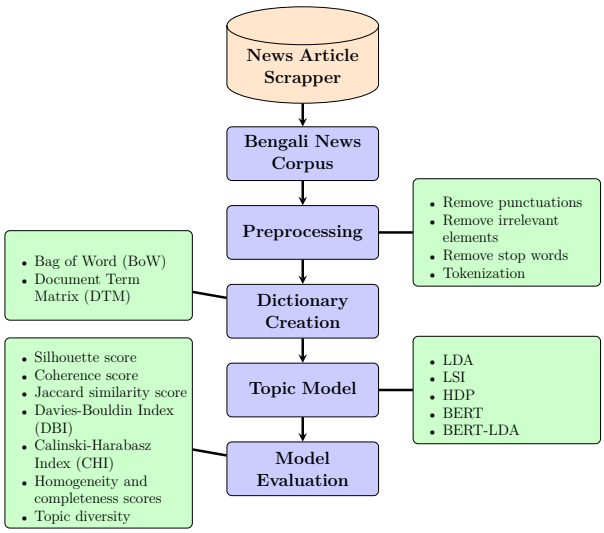

Fig. 1: Proposed Methodology

A. Data Collection

A high quality dataset is essential in the field of NLP, as the output is significantly influenced by the quality of the dataset utilized. Collecting data in Bengali has always been challenging due to limited resources. For our research, we gathered data by extracting news articles from various Bangladeshi news portals using a web crawler based on a Python library. Around 60,652 news articles from 10 different categories were collected to construct the corpus. The distribution of data across these 10 categories is illustrated in a pie chart presented in Fig. 2.

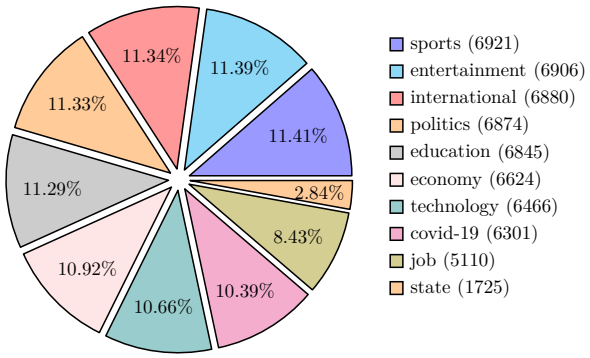

Fig. 2: Category-wise data distribution in the corpus

B. Data Preprocessing

Preprocessing is the initial and most crucial step in improving the accuracy of NLP tasks [8]. It is essential to preprocess data while preserving the underlying meaning of documents, particularly in the context of topic modeling [10]. The following steps are involved in the preprocessing of our dataset:

- Punctuations are removed along with other irrelevant elements such as extra white spaces, special characters, and non-Bengali words. Due to the significant meaning of certain numbers in our dataset, some important numbers were not removed. For example, "১৯৭১" (1971, the Liberation War of Bangladesh), "৬ দফা" (the six point movement), "৭ মার্চ" (the historical 7^{th} march speech) were kept as it is in our dataset.
- Significantly less important and frequently occurring words in a document are known as stop words. For the stop words removal task, a list of 425 Bengali stop words is used in our study.
- Tokenization is a process of splitting documents into sentences and subsequently breaking those sentences down into a list of words. Each news article is tokenized into word tokens.

After preprocessing, the statistical information of our Bengali news corpus is shown in TABLE II.

TABLE II: Statistical information about the dataset

Total news document	60652
Total word in dataset	10987697
Average words per document	181
Total characters in dataset	67514214
Average characters per document	1113
Maximum words in a document	2746
Maximum characters in a document	17075
Unique words in dataset	323219

C. Topic Modeling

Topic modeling works in an unsupervised manner for clustering a collection of documents. It provides methodologies for automatically organizing, summarizing, and understanding large corpora. Commonly used topic modeling algorithms are LSI, LDA, and HDP. For the training of these algorithms, a Bag of Word (BoW) corpus is often used, where each document is represented as a vector of word frequencies. A document-term matrix (DTM) is then created based on this frequency count, which serves as input for the topic modeling algorithms. This matrix allows the algorithms to identify patterns and relationships between words, thereby facilitating the extraction of topics from the textual data.

Recently, deep learning transformer models, such as BERT, have also been used in topic modeling. The LSI, LDA, and HDP models utilize the document-term matrix to identify relationships between documents effectively and assemble similar documents together. The use of document-term matrix is simple because it uses linear algebraic calculations and is therefore easy to understand. Unlike HDP, LDA and LSI models perform poorly when the corpus is large. In contrast, BERT has demonstrated superior performance on large datasets. Word level embeddings, like BoW, are inadequate for extracting highly semantic topics because the same word

PROCEDURE : BERT-LDA

1. **Preprocess the data:**
   CleanText(data)
   Tokenize(data)
2. **Apply LDA for topic modeling:**
   lda = LDA(num_topics)
   lda.fit(data)
   lda_vectors = lda.transform(data)
3. **Apply BERT for document embeddings:**
   bert_embeddings = BERT(data)
4. **Merge LDA vector with BERT embeddings:**
   merged_embeddings = Concatenate(lda_vectors * γ, bert_embeddings)
5. **Train an autoencoder with hyperparameter tuning:**
   Define hyperparameters
   autoencoder = Autoencoder(input_dim, latent_dim)
   optimizer = Adam(learning_rate)
   loss_fn = MSE()
   encoded_embeddings = autoencoder.encode(merged_embeddings)
6. **Perform clustering using K-means:**
   kmeans = KMeans(num_clusters)
   kmeans.fit(encoded_embeddings)
   cluster_labels = kmeans.labels
7. **Visualize clustering results using UMAP:**
   umap_embeddings = UMAP(encoded_embeddings)
   plot(umap_embeddings)

may be expressed differently across various sentences. For BERT, we have used a sentence transformer pre-trained model for extracting contextual topics [24].

We implemented a hybrid model that combines LDA and BERT, which exhibits enhanced performance compared to the use of either LDA or BERT in isolation. In this hybrid approach, the LDA topic vectors are merged together with the BERT sentence embedding vectors. Since the BERT vectors are significantly larger than the LDA vectors, scaling is performed on the LDA vectors to even out their relative importance using the gamma hyper-parameter. The merged vector is fed into an auto-encoder to ensure dimensionality reduction. Clustering is then performed to separate different topics using the default KMeans algorithm. The hyper-parameters that were used in our study for the auto encoder are listed in TABLE III.

TABLE III: Auto encoder hyper parameters

Parameters	Values
Learning Rate	1e-5
Latent dimension	32
Activation Function	RELU
Epochs	200
Batch size	128
Gamma	10
Optimizer	Adam

D. Evaluation Metrics

Different approaches exist for evaluating topic models. While no evaluation can fully replicate the assessment of a human expert, it is necessary to employ mathematical evaluation methods due to the inherent limitations of human evaluation. Topic models are commonly evaluated with coherence measures, as these correlate well with human judgment. This study specifically examines four coherence measures: C_{UCI} [25], C_{UMass} [26], C_{NPMI} [27], and C_V [28]. These measures were selected for their high correlation with human judgment and widespread use in evaluating new topic models.

- **C_{UCI}** : The coherence measure CUCI calculates coherency by considering the top N words from each topic and sum a confirmation measure over all word pairs. The precise definition of UCI is as follows:

$$C_{UCI} = \frac{2}{N.(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} PMI(w_i, w_j)$$

$$PMI(w_i, w_j) = log \frac{P(w_i, w_j) + 1}{P(w_i).P(w_j)}$$

  where the probabilities are estimated from an external reference corpus moving a sliding window through each document. $P(w)$ is the probability of a single word in each document while $P(w_i, w_j)$ determines the probability of words w_i and w_j co-occuring in each document of the reference corpus.
- **C_{UMass}** : The C_{UMass} is an intrinsic measure of coherence that relies on the word co-occurrences within the corpus being analyzed. Unlike C_{UCI}, it is independent of any external reference corpus. C_{UMass} is defined as:

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

  Although the formula of C_{UCI} and C_{UMass} seem alike, they actually employ different computational approaches. That is, $P(w_j)$ is the document frequency containing word w_j while $P(w_i, w_j)$ is the document frequency containing both w_i and w_j in the corpus.
- **C_{NPMI}** : The C_{NPMI} replaces the PMI in C_{UCI} by rescaling the probabilities of word co-occurrences known as normalized PMI. It is defined as:

$$NPMI(w_i, w_j) = \frac{log \frac{P(w_i, w_j)+1}{P(w_i).P(w_j)}}{-log(P(w_i, w_j) + 1}$$

  The normalization of PMI is achieved by applying a negative log of co-occurrence probability to each PMI result.
- **C_V** : The C_V is an extrinsic measure that uses the context vectors rather than the frequency of co-occurring words. Context vectors compute the frequency of words from the top N words only, ensuring uniform vector length. The C_V measure combines the indirect confirmation measure and a Boolean slide window over some external reference corpus. It then uses Normalized Point-wise Mutual Information (NPMI) and the cosine similarity of the context vectors to calculate the coherence score.

Topic diversity is another important metric in topic modeling evaluation to quantify the diversity of terms

TABLE IV: Clustering result analysis metrics and their significance

Name of the metric	Significance	Bounds	Ideally expected result
Silhouette score	Measures the quality of clusters	[-1.0, 1.0]	1= best prediction
Jaccard similarity score	Determines overlap between clusters	[0.0, 1.0]	0.0 = best clustering
Davies-Bouldin Index (DBI)	Measures the average similarity ratio of each cluster	[0.0, ∞]	0.0 = best clustering
Calinski-Harabasz Index (CHI)	Measures the ratio of cluster variance	[0.0, ∞]	higher value is desirable
Homogeneity score	Measures whether each cluster contains only members of a single class	[0.0, 1.0]	1.0= homogeneous prediction
Completeness score	Measures whether all members of a given class are assigned to the same cluster	[0.0, 1.0]	1.0= complete prediction
Topic diversity	Measures the distinctness of topics from each other	[0.0, 1.0]	1.0= topics are distinct and unique

across different topics. Topic diversity is typically calculated as the proportion of unique words across topics to the total number of words across all topics. This metric helps to assess whether the model has captured a broad range of distinct themes rather than repeating similar words across different topics. The formula for topic diversity is defined as:

$$Topic\,diversity\ =\ \frac{Unique\ words\ in\ all\ topics}{Total\ words\ in\ all\ topics}$$

Furthermore, the clustering results of our proposed BERT-LDA model can be evaluated using a range of metrics, including the silhouette score, Jaccard similarity score, Calinski-Harabasz Index, Davies-Bouldin Index, as well as homogeneity and completeness scores. All of these metrics are implemented within the scikit-learn library [29]. The significance of these metrics is illustrated in TABLE IV.

IV. Result and Discussion

In this section, we evaluate the results and related findings of the BERT-LDA model through several numerical experiments. The performance of a topic model is often evaluated by the number of topics it generates; therefore, it is essential to identify the optimal number of topics for our dataset. To identify this optimal number, we computed the coherence score for each topic count (k) ranging from 1 to 17 using the LDA topic model. As shown in Fig. 3, the coherence score increases gradually with the number of topics, peaking at k=11. This indicates that 11 topics yield the most coherent representation of our dataset. Based on these findings, we evaluated the proposed BERT-LDA model using the optimal value of 11 topics across various evaluation metrics.

A. Coherence score

Topic coherence provides a useful metric for evaluating the effectiveness of a particular topic model. In the context of coherence, C_{UCI}, C_{NPMI}, and C_V measure the degree of semantic similarity among words within a topic based on their co-occurrences in a reference corpus. A higher score indicates stronger semantic coherence among the words within a topic. On the other hand, the coherence C_{UMass} score measures the pairwise word similarity among the top N words of each topic. A higher negative C_{UMass} score, closer to zero, indicates stronger pairwise word similarity, thereby indicating improved topic coherence. TABLE V shows the coherence scores for all models, including the hybrid BERT-LDA model with k=11. The maximum coherence scores achieved for C_{UCI} and C_{NPMI} are **0.65** and **0.33** respectively. The C_{UMass} score ranges approximately from 0 to -4 across all models, with the optimal value being **-0.88**. Accordingly, the C_V score for all topic models falls within the range of 0.2 to 1.0 having the optimum score of **0.92** for BERT-LDA. These results demonstrate that the BERT-LDA model outperforms other topic models in terms of topic coherence, making it a superior choice for generating

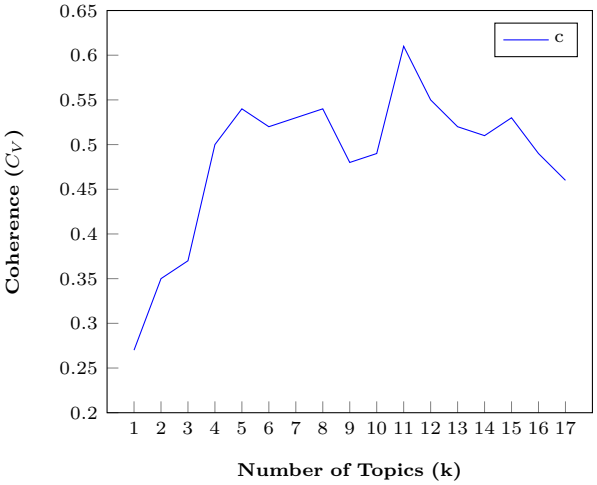

Fig. 3: Optimum coherence score for k=11 using LDA

TABLE V: Coherence scores of all topic models

Model	Coherence			
	C_{UCI}	C_{NPMI}	C_{UMass}	C_V
LDA	0.35	0.07	-3.73	0.61
LSI	0.39	0.15	-2.28	0.52
HDP	0.64	0.22	-1.88	0.85
BERT	0.53	0.15	-2.57	0.82
BERT-LDA	**0.65**	**0.33**	**-0.88**	**0.92**

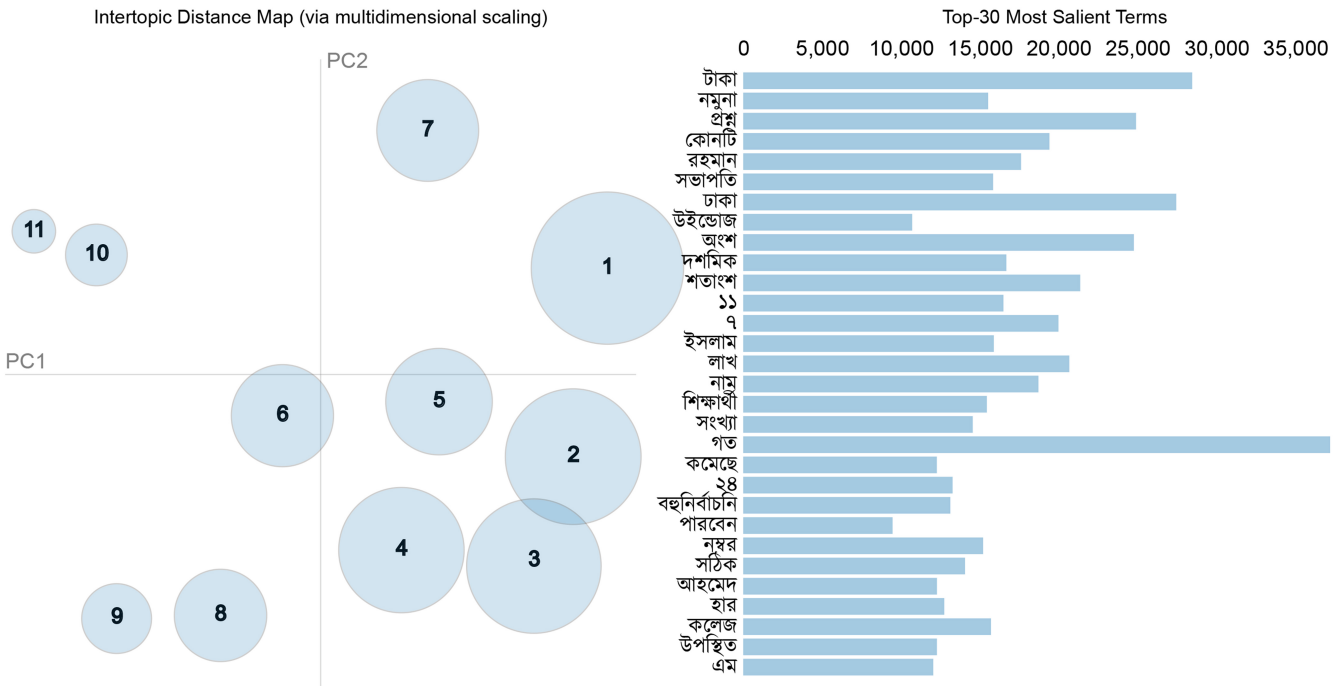

Fig. 4: Visualization of LDA topic model

semantically meaningful topics.

B. Silhouette score

The silhouette score is an important metric for evaluating the quality of clustering results. It measures how similar a document is to its own cluster compared to other clusters. It ranges from -1 to 1, where a higher score indicates that the document is well-matched to its own cluster and poorly matched to neighboring clusters. For our proposed BERT-LDA model, the silhouette score is **0.49**, suggesting that the identified topics are distinct and well-separated, demonstrating the effectiveness of the clustering process.

C. Jaccard similarity score

The Jaccard similarity score is particularly useful in determining the degree of overlap between clusters. A high score indicates significant overlap between clusters, while a low score suggests that the generated clusters are distinct and exhibit minimal overlap. For the proposed BERT-LDA model, the Jaccard similarity score is **0.31**, indicating relatively distinct clusters with limited overlap.

D. Davies-Bouldin Index

The Davies-Bouldin Index (DBI) measures how well-separated the clusters are from each other by calculating the distance between clusters. A lower DBI score reflects better separation between clusters. In our evaluation, the BERT-LDA model achieved a DBI score of **1.42**, signifying a satisfactory level of cluster separation.

E. Calinski-Harabasz Index

The Calinski-Harabasz Index (CHI), commonly referred to as the Variance Ratio Criterion, serves as a metric for assessing both the compactness of individual clusters and the degree of separation among them. A higher CHI score signifies that the clusters are both dense and well-separated. The BERT-LDA model yielded a CHI score of **8974.43**, demonstrating its ability to form cohesive and distinct clusters.

F. Homogeneity and Completeness score

The homogeneity score measures how uniformly each cluster contains only members of a single class while the completeness score measures how well all members of a given class are assigned to the same cluster. In practice, there is often a trade-off between homogeneity and completeness. Achieving high homogeneity may sometimes lead to lower completeness and vice versa. The goal is to find a balance where the topics are both pure (homogeneous) and comprehensive (complete). The homogeneity and completeness scores of our proposed BERT-LDA model are **0.51** and **0.52** respectively.

G. Topic diversity

To find topic diversity, we used the top 25 words of all topics. The topic diversity of the proposed BERT-LDA model is **0.71** which indicates that the topics are sufficiently different from each other.

H. Topic visualization

An interactive visual representation of interpreting topics [30] using the LDA model is shown in Fig. 4. The visual interactive chart is produced using the pyLDAvis package. The chart displays the topics on the left side as bubbles, while the top 30 most frequent words associated with each topic are presented on the right side. The size of each bubble reflects the relevance of that topic within the corpus, with larger bubbles indicating greater

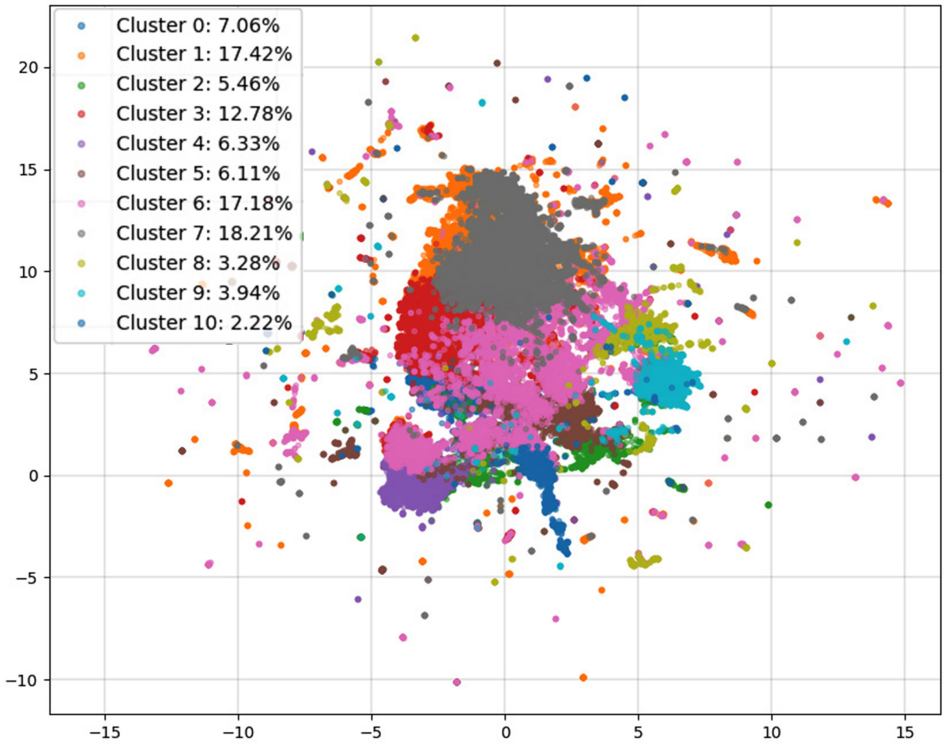

Fig. 5: Visualization of BERT clustering results

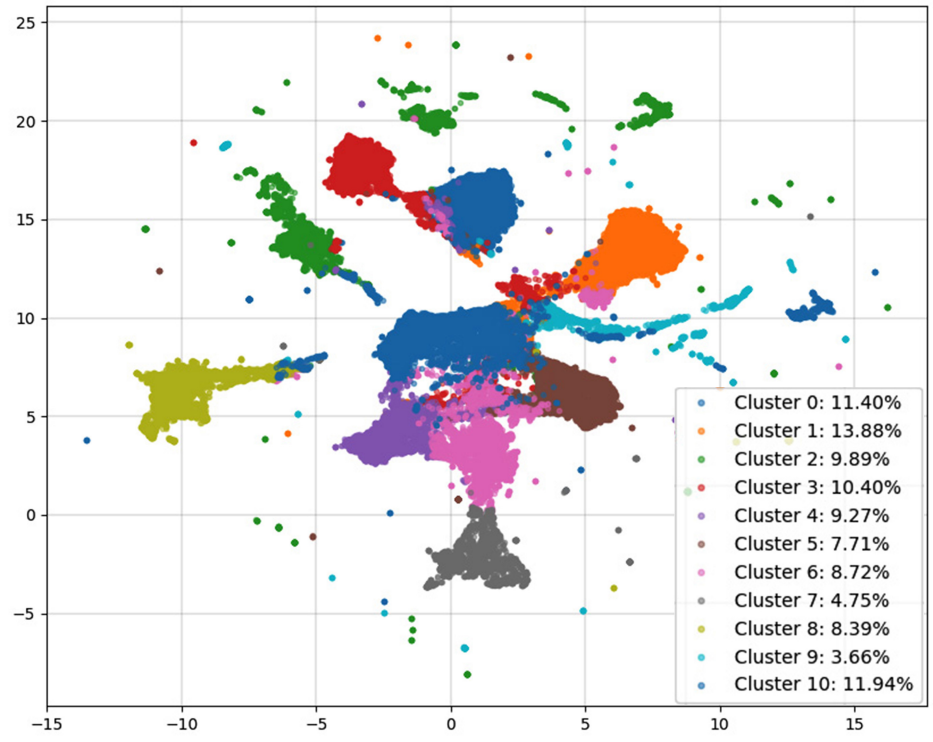

Fig. 6: Visualization of BERT-LDA clustering results

relevance. When a bubble representing a topic on the left side is selected, the most frequently occurring words for that topic are displayed on the right side. The figure shows minimum overlap between clusters.

The clustering results of the BERT and BERT-LDA models are visualized in Fig. 5 and Fig. 6 respectively. The clustering results obtained from the BERT-LDA model, as depicted in Fig. 6, demonstrate a superior performance relative to the clustering outcomes of the BERT model illustrated in Fig. 5.

The visualization of word clouds for the final 11 topics, as shown in TABLE VI, illustrates that the most frequently used words within each topic are grouped together cohesively. The size of each word is directly

TABLE VI: Visualization of word clouds of BERT-LDA for the cluster size 11

Category: Business
Top words meaning: Dhaka stock exchange, trading, price of share, share in taka, leading trader, etc.

Category: Covid
Top words meaning: died, last hour, corona detected, infection rate, recovery, sample test, etc.

Category: Economy
Top words meaning: percent in decimal, Bangladesh bank, dollar in lakh, taka in crore, government, etc.

Category: Sports
Top words meaning: finish, team, goal, excellent, yesterday, match, Messi, Ronaldo, Bangladesh, etc.

Category: Education
Top words meaning: college, Dhaka, which one is right, publish, tommorow, MCQ, remaining part, etc.

Category: Entertainment
Top words meaning: Zee TV, Zee cinema, Ramadan, Sakib Khan, Amitabh Bachchan, Gazi TV, etc.

Category: International
Top words meaning: local time, United States, India, in a report, in a statement, Pakistan, government, etc.

Category: Technology
Top words meaning: Google, made of, computer, Facebook, technology, besides, category, etc.

Category: Job
Top words meaning: apply, institutions, job types, name of the post, salary scale, last date, qualification, etc.

Category: Politics
Top words meaning: member of committee, Awami league, Mirza Fakhrul, Obaidul Qader, government, activists, etc.

Category: State
Top words meaning: village, topic, region, medical college, based on, local, officer-in-charge, etc.

proportional to its probability of occurrence within the topic; thus, as the frequency of a word increases within a given topic, its size also increases. The word clouds indicate the presence of 11 distinct categories in the dataset, while the original dataset comprises 10 categories. It is noteworthy that the business category is derived from the economy category, while the other categories remain unchanged. For better understanding, a concise explanation of each word cloud is also included in TABLE VI.

I. Performance on English datasets

To evaluate our proposed BERT-LDA model on English datasets, we utilized two publicly available datasets: the 20 Newsgroups [31] and BBC News [32] datasets. The 20 Newsgroups dataset consists of 18,846 articles across 20 categories in English, while the BBC News dataset contains 2,225 articles from 5 categories sourced from the BBC News website. The evaluation results for these two datasets are presented in TABLE VII. The results clearly

TABLE VII: Evaluation of BERT-LDA model on 20 Newsgroups and BBC News datasets

Evaluation metric		Datasets	
		20 Newsgroups	BBC News
Silhouette score		0.21	0.47
Coherence score	C_{UCI}	0.17	0.15
	C_{NPMI}	0.07	0.03
	C_{UMass}	-0.67	-1.03
	C_V	0.59	0.48
Jaccard similarity score		0.03	0.05
Davies-Bouldin index		3.68	1.19
Calinski-Harabasz index		561.73	708.15
Homogeneity score		0.36	0.31
Completeness score		0.37	0.31
Topic diversity		0.45	0.61

TABLE VIII: Performance comparisons with previous works

Research	Dataset Size	Algorithm	Coherence			
			C_{UCI}	C_{NPMI}	C_{UMass}	C_V
P. C. Paul et al. [**10**], 2022	51,016	LDA	-	-	-	0.63
		BERT-LDA	-	-	-	0.66
S. H. Mohammed et al. [**15**], 2020	300	LDA	0.59	-	-0.37	-
		LSA	0.49	-	-0.92	-
M. Grootendorst [**16**], 2022	16,309	LDA	-	-	-	0.058
		NMF	-	-	-	0.089
		CTM	-	-	-	0.096
		Doc2Vec	-	-	-	0.192
		BERTopic	-	-	-	0.166
S. Palani et al. [**17**], 2021	40,000	LDA	-	-	-	0.50
		BERT	-	-	-	0.52
		LDA+BERT	-	-	-	0.56
S. K. Ray et al. [**19**], 2019	10,400	LSI	-	-	-4.118	0.485
		LDA	-	-	-2.808	0.662
		NMF	-	-	-1.628	0.797
M. H. Asnawi et al. [**22**], 2023	15,000	CTM-MPNet	-0.6407	0.0752	-	0.7091
M. C. Wijanto et al. [**23**], 2024	20,972	BERTopic	-	-	-2.4856	0.5412
		RoBERTa	-	-	-1.8291	0.5554
		DistilRoBERTa	-	-	-1.7787	0.4950
This Research	60,652	LDA	0.35	0.07	-3.73	0.61
		LSI	0.39	0.15	-2.28	0.52
		HDP	0.64	0.22	-1.88	0.85
		BERT	0.53	0.15	-2.57	0.82
		BERT-LDA	**0.65**	**0.33**	**-0.88**	**0.92**

indicate that our proposed model performs effectively on English datasets, suggesting its suitability for application in other languages as well.

J. Performance comparison with previous works

The performance comparisons of the topic models employed in our study with several notable previous works are illustrated in TABLE VIII. Based on the evaluation metrics, our findings suggest that BERT-LDA demonstrates greater effectiveness than the other models utilized in prior researches.

V. Conclusion

Topic modeling is a powerful method for uncovering thematic patterns and latent topics within textual datasets. Various topic modeling approaches have been employed to extract meaningful insights from Bengali text, supporting applications such as information retrieval, document clustering, and recommendation systems in the Bengali language domain. BERT-LDA is a hybrid fusion of BERT and LDA, which provides a unique advantage through the combination of the semantic contextual embedding power of BERT with the probabilistic modeling of LDA. BERT-LDA, a hybrid fusion of BERT and LDA, offers a significant advantage by combining BERT's ability to generate rich semantic contextual embeddings with LDA's probabilistic topic modeling capabilities. While still an emerging approach in Bengali topic modeling, BERT-LDA demonstrates enhanced topic interpretability, as measured by various evaluation metrics. Our experiments yielded promising results with traditional probabilistic models like LDA, LSI, and HDP, while BERT-LDA outperformed them all in terms of coherence scores. Additional evaluation metrics, including Silhouette Score, Jaccard Similarity, Davies-Bouldin Index (DBI), Calinski-Harabasz Index (CHI), Topic Diversity, as well as Homogeneity and Completeness Scores, further validate BERT-LDA's effectiveness for Bengali topic modeling. However, BERT-LDA may incur considerable computational costs, especially during the inference process, due to the large number of parameters associated with BERT and the iterative nature of LDA inference. Future research aims to address these challenges and extend the scope to multi-modal and cross-lingual topic modeling, facilitating knowledge discovery across diverse languages and cultures.

References

[1] 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, pp. 147-153, 2015.

[2] P. Shah, D. Sharma, and R. Sekhar, "Analysis of Research Trends in Fractional Controller Using Latent Dirichlet Allocation," *Engineering Letters*, vol. 29, no. 1, pp. 109-119, 2021.

[3] B. V. Barde and A. M. Bainwad, "An overview of topic modeling methods and tools," *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, Jun. 2017, pp. 745-750.

[4] D. Blei, B. Edu, A. Ng, M. Jordan, and J. Edu, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[5] 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, 1990.

[6] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet Processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[7] M. M. Samia, A. Rajee, Md. R. Hasan, M. O. Faruq, and P. C. Paul, "Aspect-based Sentiment Analysis for Bengali Text using Bidirectional Encoder Representations from Transformers (BERT)," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 12, pp. 978-986, 2022.

[8] S. Arts, J. Hou, and J. C. Gomez, "Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures," *Research Policy*, vol. 50, no. 2, p. 104144, 2021.

[9] R. Alfred, L. Y. Jie, J. H. Obit, Y. Lim, H. Haviluddin, and A. Azman, "Social Media Mining: A Genetic Based Multi-objective Clustering Approach to Topic Modelling," *IAENG International Journal of Computer Science*, vol. 48, no. 1, pp. 32-42, 2021.

[10] P. C. Paul, M. Shihab Uddin, M. T. Ahmed, M. Moshiul Hoque, and M. Rahman, "Semantic Topic Extraction from Bangla News Corpus Using LDA and BERT-LDA," *2022 25th International Conference on Computer and Information Technology (ICCIT)*, Dec. 2022, pp. 512-516.

[11] M. Mouhoub and M. Al Helal, "Topic Modelling in Bangla Language: An LDA Approach to Optimize Topics and News Classification," *Computer and Information Science*, vol. 11, no. 4, pp. 77-83, 2018.

[12] M. Hasan, M. M. Hossain, A. Ahmed and M. S. Rahman, "Topic Modelling: A Comparison of The Performance of Latent Dirichlet Allocation and LDA2vec Model on Bangla Newspaper," *2019 International Conference on Bangla Speech and Language Processing (ICBSLP), Sylhet, Bangladesh*, 2019, pp. 1-5.

[13] T. Yang, A. Torget, and R. Mihalcea, "Topic Modeling on Historical Newspapers," *in Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, USA*, 2011, pp. 96–104.

[14] K. M. Alam, Md. T. H. Hemel, S. M. Muhaiminul Islam, and A. Akther, "Bangla News Trend Observation using LDA Based Topic Modeling," *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, Dec. 2020, pp. 1-6.

[15] S. H. Mohammed and S. Al-augby, "LSA & LDA Topic Modeling Classification: Comparison study on E-books," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 19, no. 1, pp. 353–362, 2020.

[16] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[17] S. Palani, P. Rajagopal, and S. Pancholi, "T-BERT–Model for Sentiment Analysis of Micro-blogs Integrating Topic Model and BERT," *arXiv preprint arXiv:2106.01097* , 2021.

[18] S. S. Panigrahi, N. Panigrahi, and B. Paul, "Modelling of Topic from Hindi Corpus using Word2Vec," *in 2018 Second International Conference on Advances in Computing, Control and Communication Technology (IAC3T)*, 2018, pp. 97–100.

[19] S. K. Ray, A. Ahmad, and C. A. Kumar, "Review and Implementation of Topic Modeling in Hindi," *Applied Artificial Intelligence*, vol. 33, no. 11, pp. 979–1007, 2019.

[20] A. Abuzayed and H. Al-Khalifa, "BERT for Arabic Topic Modeling: An Experimental Study on BERTopic Technique," *Procedia Computer Science*, vol. 189, pp. 191–194, 2021.

[21] L. George and P. Sumathy, "An integrated clustering and BERT framework for improved topic modeling," *International Journal of Information Technology*, vol. 15, no. 4, pp. 2187–2195, 2023.

[22] M. H. Asnawi, A. A. Pravitasari, T. Herawan, and T. Hendrawati, "The Combination of Contextualized Topic Model and MPNet for User Feedback Topic Modeling," *IEEE Access*, vol. 11, pp. 130272–130286, 2023.

[23] M. C. Wijanto, I. Widiastuti, and H.-S. Yong, "Topic Modeling for Scientific Articles: Exploring Optimal Hyperparameter Tuning in BERT," *International Journal on Advanced Science, Engineering and Information Technology (IJASEIT)*, vol. 14, no. 3, pp. 912–919, 2024.

[24] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2019, pp. 3982–3992.

[25] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic Evaluation of Topic Coherence," *in Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010, pp. 100–108.

[26] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing Semantic Coherence in Topic Models," *in Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 2011, pp. 262–272.

[27] N. Aletras and M. Stevenson, "Evaluating Topic Coherence Using Distributional Semantics," *in Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, 2013, pp. 13–22.

[28] M. Röder, A. Both, and A. Hinneburg, "Exploring the Space of Topic Coherence Measures," *in Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 2015, pp. 399–408.

[29] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[30] C. Sievert and K. Shirley, "LDAvis: A method for visualizing and interpreting topics," *in Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 2014, pp. 63–70.

[31] "7.2. Real world datasets - scikit-learn 1.5.2 documentation," [Online]. Available: https://scikit-learn.org/stable/datasets/real_world.html#the-20-newsgroups-text-dataset.

[32] B. Bose, "BBC News Classification," 2019. [Online]. Available: https://kaggle.com/competitions/learn-ai-bbc.

Date of modification: 26 March, 2025
Description of Changes: Removed the corresponding author mark (*) from the first author (Pintu Chandra Paul).