GHTM: A Graph based Hybrid Topic Modeling Approach in Low-Resource Bengali Language

Anonymous ACL submission

Abstract

Topic modeling is a Natural Language Processing (NLP) technique that is used to identify latent themes and extract topics from text corpora by grouping similar documents based on their most significant keywords. Although widely researched in English, topic modeling remains understudied in Bengali due to its morphological complexity, lack of adequate resources and initiatives. In this contribution, a novel Graph Convolutional Network (GCN) based model called GHTM (Graph-Based Hybrid Topic Model) is proposed. This model represents input vectors of documents as nodes in the graph, which GCN uses to produce semantically rich embeddings. The embeddings are then decomposed using Non-negative Matrix Factorization (NMF) to get the topical representations of the underlying themes of the text corpus. This study compares the proposed model against a wide range of Bengali topic modeling techniques, from traditional methods such as LDA, LSA, and NMF to contemporary frameworks such as BERTopic and Top2Vec on three Bengali datasets. The experimental results demonstrate the effectiveness of the proposed model by outperforming other models in topic coherence and diversity. In addition, we introduce a novel Bengali dataset called 'NCTBText' sourced from Bengali textbook materials to enrich and diversify the predominantly newspaper-centric Bengali corpora.

1 Introduction

Topic modeling is a powerful unsupervised text mining technique that helps make sense of unstructured and unlabeled real-world data, without the manual labor of going through large volumes of documents. Through clustering words that tend to co-occur frequently across multiple documents, topic models generate insightful and thematic set of words that can enlighten us about the topics from any massive amount of text corpora, which can further contribute in other Natural Language Processing (NLP) tasks like document classification, information retrieval, sentiment analysis, exploratory data analysis, etc.

The field of topic modeling has advanced a lot recently since its inception, from conventional models like Latent Dirichlet Allocation (LDA) (Blei et al., 2003), Latent Semantic Indexing (LSI) (Deerwester et al., 1990), Non-negative Matrix Factorization (NMF) (Lee and Seung, 1999) etc. to high-quality sentence embedding based models like BERTopic (Grootendorst, 2022), Top2vec (Angelov, 2020) etc. and neural models like ETM (Dieng et al., 2020), ProdLDA (Srivastava and Sutton, 2017), NeuralLDA (Card et al., 2018) etc. Earlier the text extraction only depended on either Bag-of-Words (BOW) or Term Frequency-Inverse Document Frequency (TF-IDF), which has later evolved into word-based embeddings and sentence level embeddings. As text vectorization methods improved, topic models also progressed over time from probabilistic and algebraic form to neural network and embedding-based models.

Even though recent benchmarks for topic modeling in English are leveraging cutting-edge tools and techniques, Bengali is lagging in this field. Bengali topic modeling has mostly been explored around LDA and extensions of LDA like LDA2Vec (Hasan et al., 2019) and BERT-LDA (Paul et al., 2025). Not only is there a lack of advanced research in Bengali topic modeling, but there is also no widely accepted benchmark dataset for evaluating such models. Most of the available Bengali datasets are scraped from online newspapers, often lacking in variation of topics.

This paper presents a novel topic modeling approach, especially for Bengali language, called Graph-based Hybrid Topic Model **(GHTM)** which is an innovative fusion of TF-IDF weighted GLoVE (Pennington et al., 2014) embeddings, Graph Convolutional Network (GCN) (Kipf and Welling, 2017)and NMF, that is evaluated against currently available models and outperforms them in topic coherence and topic diversity. GHTM constructs a k-nearest-neighbor graph, where each document serves as a node, and uses cosine similarity to measure document similarity to connect edges. Documents are represented as TF-IDF-weighted GloVe embeddings, which are subsequently refined through GCN. Finally, NMF is applied to the refined embeddings to extract interpretable topics. To ensure compliance with NMFs non-negativity constraint, the refined embeddings are converted to their absolute values.

We evaluated a wide range of models, including our proposed approach, on three differently sized datasets. Two of these are publicly available newspaper datasets, while the third is a novel Bengali textbook dataset curated from materials provided by Bangladesh's National Curriculum and Textbook Board (NCTB) website[1].

The major contributions of this study are as follows:

- Development of a novel graph-based topic modeling approach called **GHTM**.
- Generation of a novel dataset, **NCTBText** that introduces diversity into the currently dominant newspaper-centric Bengali corpora.
- A comprehensive comparison of topic modeling methods using a wide range of existing and newly generated datasets.

2 Related Work

This section reviews related studies, by showing their contributions and identifying key research gaps, that we aim to address in this work.

Helal and Mouhoub (2018) took the first step towards Bengali topic modeling by applying LDA on Bengali dataset, which shows the efficacy of LDA with bi-grams for Bengali news classification and topic extraction. Hasan et al. (2019) proposes LDA2Vec, which is a combination of the parts of LDA and word2vec. The study shows LDA2Vecs high accuracy over LDA itself in a comparison between these two models on a Bengali newspaper dataset but did not employ more than one dataset. Alam et al. (2020) curated a Bengali dataset consisting of 70K news articles and applied LDA to uncover latent topics and observe media trend evolution over time in Bengali news. The study offers an in-depth analysis and demonstrates how different topics prevail across weeks. Ahmed et al. (2021) presents a structured overview of topic modeling research from 2003 to 2020, covering a wide range of techniques. Its strength lies in highlighting the disparity between English and Bengali topic modeling efforts, answering some of the most important questions regarding topic modeling such as "What are the techniques that have been used in English topic modeling but not yet used in Bangla?", "What are the sources of the datasets used?" etc. through rigorous research. The study also effectively outlines future research scopes for Bengali topic modeling. But the review lacks quantitative analysis and remains mostly descriptive. Paul et al. (2025) compiled a novel Bengali news dataset and proposed a hybrid model combining the potentials of both LDA and BERT, called BERT-LDA, advancing topic modeling in Bengali. The study compares its proposed hybrid model with traditional models like LDA, LSI, Hierarchical Dirichlet Process (HDP) etc. in terms of topic coherence. The authors also applied their model on English benchmark datasets (20NewsGroup, BBC) for topic modeling and demonstrated the results. Dawn et al. (2024) proposes a Dirichlet-polynomial clustering model called Likelihood Corpus Distribution (LCD), which is based on a Bayesian numerical prototype that evaluates the probability distribution of words in a document to identify topics. Experiments are done to show the efficiency of LCD over conventional topic models on five real-world datasets of Bengali corpora.

Graph Neural Networks (GNN) have been proven to be effective across many scientific tasks including NLP and have also caught the attention of the topic modeling community in recent years. Thus, many researchers in this field have successfully incorporated GNN in topic modeling. Some graph-based topic modeling studies are discussed below. Shen et al. (2021) proposed a novel method called Graph Neural Topic Model (GNTM), which represents documents as semantic graphs and uses the Neural Variational Inference (NVI) approach with GNN for topic modeling. They evaluated their model against baseline topic modeling methods on four benchmark English datasets and showed promise in performance. Graph Contrastive Neural Topic Model (GCTM) (Luo et al., 2024) integrates contrastive learning with topic modeling via graph-based sampling to resolve semantic redundancy and false negatives in topic discovery. Their model treats document data augmentation as

[1] http://www.nctb.gov.bd

a graph data augmentation problem and conducts graph contrastive learning (GCL) based on instructive positive and negative samples generated by a graph-based sampling strategy. GCTM significantly outperforms existing neural topic models in coherence and representation quality across benchmark datasets in English. Graph Enhanced Autoencoded Variational Inference for Biterm Topic Model (GraphBTM) (Zhu et al., 2018) represents bi-terms as graphs and design GCNs with residual connections to extract transitive features from bi-terms, resulting in more coherent topics. They also propose a dataset called "All News" which has larger documents than 20 Newsgroups. Topic Modeling with Graph Isomorphism Network (GINopic) (Adhya and Sanyal, 2025) is another approach that takes the word similarity graphs for each document, where the word similarity graph is constructed using word embeddings to capture the complex correlations between the words. The study performs extrinsic evaluations on diverse benchmark datasets, showing the effectiveness of GINopic.

While GNN-based models have shown their prominence in topic modeling, it has not been adapted for Bengali language yet, to the best of our knowledge. To bridge this gap, we propose a novel method using GCN, especially tailored for Bengali.

3 Methodology

The proposed GHTM model consists of three stages. First, text features are extracted using a combination of TF-IDF and GloVe representations. The second stage constructs a graph from the extracted vectors, which is then processed by a graph convolutional network to produce similarity-aware embeddings. Finally, matrix factorization is applied to the embeddings to generate diverse and coherent topic keywords.

3.1 Text Vectorization

The text vectorization process is based on the combination of TF-IDF and GloVe. First, the TF-IDF vectorizer constructs a sparse document-term matrix of size $N \times V$, where N is the number of documents and V is the vocabulary size. Separately, the GloVe model generates a dense $V \times D$ matrix of pre-trained word embeddings for the same vocabulary, where D denotes the embedding dimension. The matrix dot product of the TF-IDF matrix and the GloVe embeddings produces the final document representation matrix of size $N \times D$. This effectively merges the statistical relevance of TF-IDF with the semantic richness of GloVe. As a result, the input vector becomes the GloVe embeddings weighted by TF-IDF for each document, where each vector reflects the importance of its constituent words based on their TF-IDF scores.

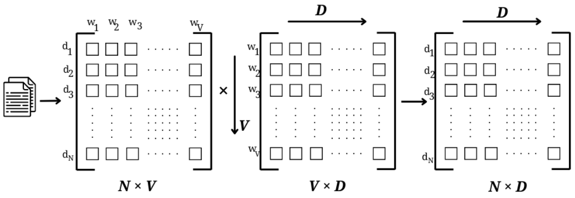

Figure 1: **Text Vectorization** Stage. TF-IDF produced sparse document-term matrix of size $N \times V$ and a dense $V \times D$ matrix generated by GloVe goes through matrix multiplication. It yields a document representation matrix of size $N \times D$.

3.2 Graph Convolutional Network (GCN)

GCN (Graph Convolutional Network) is a powerful neural architecture that is used in GHTM to learn the similarity of documents for topic modeling. Since, GCN requires graph representation of input, a K-Nearest-Neighbor (KNN) graph is constructed from the previously calculated vectors. To construct this graph, we represent the documents as graph nodes and connect these nodes with edges based on document similarity. As this is a KNN graph, the edges are connected depending on neighbors of each node and the distance is calculated using cosine similarity. Therefore, we get a graph representation of our input vectors, which is similarity-aware. The GCN here acts as a influential intermediary which learns the way to keep close similar nodes together and push the dissimilar nodes apart and reduces the dimensionality along the way.

Cluster-GCN (Chiang et al., 2019) is used to utilize the GPU memory efficiently, dividing the input graph into sub-graphs. This way, the GPU does not run out of memory trying to process the whole graph at once, when the training data is substantially large. The performance does not suffer because of Cluster-GCN as we preserve the inter-cluster edges.

The architecture employs a joint loss function, combining margin-based hinge loss, enforcing local edge structure preservation through positive/negative node pair contrast, with a global con-

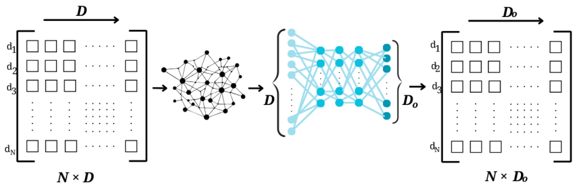

Figure 2: **Graph Convolutional Network (GCN)** Stage. The document representation matrix of size $N \times D$, converted to a KNN graph and passed through Cluster-GCN to get enriched similarity-aware document embeddings. Here, D_o denotes the output dimension of the GCN.

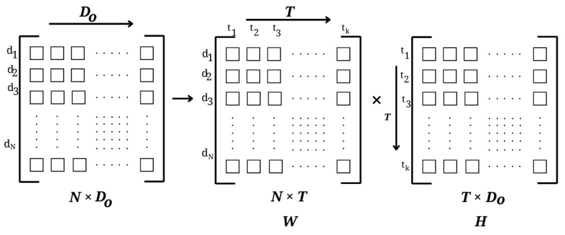

Figure 3: **Matrix Factorization** Stage. Refined embeddings from GCN stage are factorized using NMF to derive document-topic distribution matrix (W) and topic-embedding matrix (H).

trastive loss that sharpens embedding distinctiveness via self-supervised discrimination. To simply put, hinge loss penalizes dissimilar connected nodes, and contrastive loss makes each nodes embedding unique, balancing local relational fidelity (via edges) and global semantic separation (via contrastive pull-push). The architecture employs edge dropout and residual connections, enabling scalable processing of graph-structured data and stabilizes training via graph normalization.

At the end of this stage, GCN produces refined, dimensionality reduced and semantically enriched embeddings which now can be applied to the next stage.

3.3 Matrix Factorization

This is the final stage of GHTM, where Non-negative Matrix Factorization (NMF) is applied to decompose the GCN produced embeddings for extracting topics. Unlike traditional NMF-based topic modeling, which factorizes a sparse document-term matrix, our approach factorizes a dense document embedding matrix.

Since NMF requires non-negative inputs, several transformation techniques—such as ReLU, SoftPlus, Global Minimum Shift, and Absolute Value Transformation—are explored to ensure non-negativity. Among these, Absolute Value Transformation yields the best performance in our case.

$$X_{\text{non-negative}} = |X| \tag{1}$$

where X is the input matrix.

Following factorization, a document-topic distribution matrix (W) is obtained, where each row corresponds to a document and each column reflects the document's association with a topic. Additionally, a topic-embedding matrix (H) is generated; however, this matrix does not directly map to words, as it captures abstract relationships in the embedding space rather than explicit vocabulary terms. Therefore, instead of directly retrieving topic words from the factorized matrices, a post-processing workaround is employed to map the learned topics to human-interpretable keywords, as described in the following steps.

1. From the document-topic matrix (W), we identify the representative documents for each topic.

2. We aggregate the term frequencies from those documents using the original sparse matrix from the vectorization stage of size $N \times V$.

3. Finally, we select the most frequent terms in these aggregated weights as the topic words.

This bypasses the need to interpret the abstract embedding-based topic components, rather, this method assumes that documents strongly associated with a topic will contain words relevant to that topic in their raw text data.

4 Results and Analysis

This section elaborately discusses the datasets and evaluation metrics used in this study, experimental setup, comparative analysis results and the findings.

This study benchmarks the proposed model against diverse topic-modeling approaches: traditional methods like **LDA** (Blei et al., 2003) (probabilistic generative modeling), **LSA** (Deerwester et al., 1990) (linear algebra via Singular Value Decomposition), and **NMF** (Lee and Seung, 1999) (non-negative matrix factorization); neural variants including **ProdLDA** (Srivastava and Sutton, 2017) (Variational Autoencoder with multinomial decoding), **NeuralLDA** (Card et al., 2018) (neural parameterization of Dirichlet priors), and **ETM** (Dieng et al., 2020) (topic-word embedding alignment); embedding-enhanced models such as **CombinedTM** (Bianchi et al., 2021a) (BoW + SBERT inputs) and **ZeroShotTM** (Bianchi et al., 2021b) (SBERT-based zero-shot transfer); and clustering-driven frameworks **Top2Vec** (Angelov, 2020) (joint document-word embedding with UMAP/HDBSCAN) and **BERTopic** (Grootendorst, 2022) (BERT embeddings, UMAP, HDBSCAN, and c-TF-IDF for data-driven topics). The comparison spans probabilistic, neural, embedding-integrated, and clustering-based paradigms to evaluate cross-methodological performance.

4.1 Datasets

This research utilizes three datasets of varying sizes to evaluate the proposed GHTM approach alongside existing topic modeling techniques.

Jamuna News is curated from the Jamuna TV website. It is a balanced collection of short documents, obtained from *Kaggle*[2].

BanFakeNews (Hossain et al., 2020), is compiled from Bengali newspaper. This dataset was released to combat the spread of fake news in Bengali. This study uses its *"Authentic-48K"* subset, sourced from Kaggle and stripped the labels for topic modeling purposes.

NCTBText, is a novel dataset developed for this research. It comprises unlabelled text from textbooks available on the Bangladesh Government's NCTB website. The dataset includes diverse subjects such as Religion, Bengali, Science, Agriculture, Information and Communication Technology (ICT), Business, Social Science, and Home Science. The attributes of the datasets are presented as a summary in Table 1.

4.2 Evaluation Metrics

We evaluate the performance of the models in terms of both topic diversity and topic coherence. The metrics used in this study are introduced below.

Normalized Pointwise Mutual Information (NPMI) (Newman et al., 2010) is statistical measure of word association that measures topic coherence internally using a sliding window to count word co-occurrence patterns. The measure ranges from [-1, 1] where 1 indicates a perfect relevance of words in a topic.

$$\mathrm{NPMI}\left(w_i, w_j\right) = \frac{\log \frac{P(w_i,w_j)}{P(w_i)P(w_j)}}{-\log P\left(w_i, w_j\right)} \qquad (2)$$

where $P(w_i)$ and $P(w_j)$ are marginal probabilities of the words w_i and w_j. $P(w_i, w_j)$ is the joint probability of co-occurrence.

Topic Coherence (CV) (Röder et al., 2015) is a variant of NPMI that measures the semantic relatedness of topic words. It uses the one-set segmentation to count word co-occurrences and the cosine similarity as the similarity measure. It ranges from [0, 1], where 1 means closely related words identified as topic representatives.

Topic Diversity (TD) (Dieng et al., 2020) measures the uniqueness of the words across all topics and the measure ranges from [0, 1] where 0 indicates redundant topics and 1 indicates highly diverse topics.

$$\mathrm{TD} = \frac{\left|\bigcup_{k=1}^{K} W_k\right|}{K \times T} \qquad (3)$$

where W_k = Set of top T words in topic k, and K = Total number of topics.

Inverted Rank-Biased Overlap (IRBO) (Webber et al., 2010), a diversity metric that evaluates inter-topic dissimilarity, derived from Rank-Biased Overlap (RBO). While RBO quantifies the similarity of ranked word lists across topics, IRBO inverts RBO to penalize overlapping top words.

We used Gensim's CoherenceModel (Řehůřek and Sojka, 2010) to compute coherence metrics (NPMI and CV) and OCTIS (Optimizing and Comparing Topic models is Simple) (Terragni et al., 2021) to calculate diversity metrics (TD and IRBO). **Runtime (RT)** is also recorded for each run to observe how long models take to train, as the size of the dataset grows.

Usually, the topic modeling studies experiment with the topic number, *k*, but we set the number of topics according to the ground truth to ensure uniformity across the models. Except BERTopic and Top2vec, all the models that are being compared here expects the topic number to be set beforehand. Therefore, we add another metric in our comparison table called, **Topics Identified (*n*)** to observe how many topics the models (BERTopic, Top2Vec) think are there in the dataset.

4.3 Experimental Setup

This section outlines the experimental settings used in this research, including data preparation and

[2] https://www.kaggle.com/datasets/durjoychandrapaul/over-11500-bangla-news-for-nlp

Dataset	Count of Documents	Vocabulary Size	Classes	Size	Avg. Word Count
Jamuna News	11 904	34 101	4	19.2 MB	89.00
NCTBText	8 650	84 269	8	37.6 MB	271.73
BanFakeNews	48 678	130 227	12	244.4 MB	304.59

Table 1: Dataset Summary

Embedding	Model Name	Dimension
Word2Vec (W2V)	bnwiki_word2vec	100
Doc2Vec	bangla_news_article_doc2vec	100
GloVe	bn_glove.39M.300d	300
FastText	fasttext_cc.bn.300	300
ST	bangla-sentence-transformer (Uddin et al., 2024)	768

Table 2: Embedding Model Summary

Hyper-parameter	Jamuna News	NCTBText	BanFakeNews
# Hidden Layers	3	2	2
# Clusters in GCN	4	8	12
Hidden Dimension	32	32	128
Output Dimension	64		
Epochs	100		
# Neighbors in KNN	15		
Learning Rate	0.005		
Dropout	0.4		
Edge Dropout	0.2		

Table 3: GHTM hyper-parameter values for the datasets

model configuration.

4.3.1 Data Preparation

We prepared the datasets in both tokenized sequences and raw sentence forms to accommodate all the models used in the comparative analysis. While BERTopic and Top2Vec accepts raw sentences as they work with sentence-level embeddings, other conventional and neural models accept tokenized list of words as documents. CombinedTM and ZeroShotTM, on the other hand leverages both formats. The tokenized format of the datasets went through rigorous pre-processing steps like tokenization, stop words removal, lemmatization, while raw sentence format only had unnecessary punctuations and numerics removed. We used unigrams across all the models for this study to keep the comparison straightforward.

4.3.2 Model Configuration

All the models employed in the study were set to their default configurations and hyper-parameters, but we had to modify sklearns CountVectorizer parameters for the models that uses this module for vectorization to accurately handle Bengali text tokenization. The hyper-parameter values used in GHTM, on the other hand, are shown in Table 3 for each dataset. The embedding models that are used throughout the experiment are summarized in Table 2.

4.4 Findings

This section discusses and analyzes the results of the experiment that is shown in Table 4.

Among the **conventional models** LSA performed poorly than the others. In case of BanFakeNews dataset, LDA and NMF failed to find the topics *k*, which was set to ground truth, and therefore the metric Topics Identified (*n*) is mentioned on the comparison table, despite being irrelevant for traditional models. This metric is only applicable for BERTopic and Top2Vec.

Considering the **neural models**, ETM has the inferior results than the others, despite using FastText for word embeddings, while CombinedTM performs well leveraging both BoW and sentence transformer embeddings, which also thrives in generating diverse topics.

Among the **cluster-based models**, Top2Vec based on Doc2Vec embeddings is way ahead than others. Despite performing well on Doc2Vec, while tried on other available embedding options for Bengali such as universal-sentence-encoder-multilingual, distiluse-base-multilingual-cased and paraphrase-multilingual-MiniLM-L12-v2, Top2Vec fails by only producing 1 topic. We also learn from Table 4 that, as the size of the dataset grows, the runtime for Top2Vec increases rapidly. BERTopic on the other hand takes a reasonable amount of time based on dataset size, but its performance lags behind Top2Vec. As BERTopic offers modularity and flexibility in choosing embedding model, dimensionality reduction model and clustering model, it gave us the chance to explore a lot of combinations for Bengali. And our experiments show that Bengali sentence transformer model along with UMAP and KMeans works best for Bengali in BERTopic, rather than the authentic BERTopic combination of UMAP and HDBSCAN.

Model	Jamuna News						NCTBText						BanFakeNews					
	CV	NP MI	TD	IR BO	RT	*n*	CV	NP MI	TD	IR BO	RT	*n*	CV	NP MI	TD	IR BO	RT	*n*
Traditional Models																		
LDA (BOW)	0.62	0.08	0.89	0.95	3	-	0.50	0.03	0.83	0.95	3	-	0.64	0.12	0.86	0.97	33	10
LDA (TF-IDF)	0.57	0.03	0.91	0.95	3	-	0.51	-0.04	0.96	0.99	3	-	0.65	0.09	0.97	1.00	30	10
LSA (BOW)	0.51	-0.04	0.60	0.63	1	-	0.34	-0.05	0.47	0.78	1	-	0.44	-0.01	0.39	0.79	12	-
LSA (TF-IDF)	0.49	-0.05	0.68	0.67	1	-	0.42	-0.07	0.61	0.78	2	-	0.45	-0.03	0.42	0.83	21	-
NMF (BOW)	0.62	0.09	0.91	0.95	1	-	0.54	0.06	0.82	0.96	1	-	0.67	0.13	0.84	0.97	8	10
NMF (TF-IDF)	0.65	0.08	0.98	0.99	1	-	0.57	0.04	0.95	0.99	1	-	0.72	0.17	0.94	0.99	16	10
Neural Models																		
ProdLDA	0.60	-0.02	1.00	1.00	166	-	0.66	0.05	1.00	1.00	219	-	0.39	-0.19	0.89	0.98	2355	-
Neural LDA	0.55	-0.24	1.00	1.00	163	-	0.54	-0.30	1.00	1.00	214	-	0.42	-0.22	0.96	0.99	2187	-
ETM (FastText)	0.49	-0.02	0.92	0.95	14	-	0.40	-0.14	0.81	0.94	21	-	0.49	-0.12	0.08	0.00	411	-
Combined TM	0.67	0.04	1.00	1.00	174	-	0.64	0.05	1.00	1.00	174	-	0.71	0.14	1.00	1.00	686	-
ZeroShot TM	0.66	0.04	0.87	0.93	104	-	0.64	0.05	1.00	1.00	115	-	0.60	0.05	0.37	0.75	353	-
Cluster-Based Models																		
Top2vec	0.83	0.18	1.00	1.00	32	72	0.74	0.11	1.00	1.00	134	87	0.80	0.12	0.99	1.00	1651	502
BERTopic (W2V+UMAP+HDBSCAN)	0.26	-0.15	0.78	0.81	35	3	0.44	-0.05	0.73	0.90	34	9	0.39	-0.12	0.85	0.95	176	6
BERTopic (GloVe+UMAP+HDBSCAN)	0.53	0.03	0.83	0.88	19	120	0.56	0.13	0.68	0.91	21	110	0.43	-0.11	0.83	0.95	278	7
BERTopic (Doc2Vec+UMAP+HDBSCAN)	0.25	-0.16	0.67	0.72	36	4	0.44	-0.05	0.72	0.90	34	11	0.36	-0.12	0.84	0.94	157	6
BERTopic (FastText+UMAP+HDBSCAN)	0.25	-0.18	0.77	0.80	47	3	0.42	-0.07	0.79	0.93	43	11	0.40	-0.12	0.86	0.96	167	6
BERTopic (ST+UMAP+HDBSCAN)	0.48	-0.06	0.85	0.93	68	150	0.82	-0.08	0.44	0.52	56	6	0.62	0.03	0.93	0.99	356	773
BERTopic (ST+PCA+HDBSCAN)	0.55	-0.03	0.99	1.00	82	4	0.78	-0.05	0.49	0.61	53	6	0.62	0.00	0.93	0.99	239	10
BERTopic (ST+SVD+HDBSCAN)	0.66	0.01	1.00	1.00	82	5	0.78	-0.04	0.51	0.62	53	6	0.56	0.03	1.00	1.00	285	3
BERTopic (ST+HDBSCAN)	0.65	0.02	1.00	1.00	82	4	0.77	-0.04	0.55	0.74	52	7	0.53	0.02	0.97	0.99	282	3
BERTopic (ST+UMAP+KMeans)	0.71	0.08	0.99	1.00	82	4	0.69	0.02	0.97	0.99	54	8	0.67	0.07	0.94	0.99	307	12
BERTopic (ST+UMAP+Agglomerative)	0.54	-0.09	0.93	0.96	81	4	0.58	-0.09	0.86	0.96	53	8	0.65	0.00	0.97	0.99	232	12
BERTopic (ST+UMAP+DBSCAN)	0.43	-0.12	0.91	0.92	82	4	0.65	-0.09	0.76	0.86	52	4	0.59	0.01	0.94	0.99	233	43
BERTopic (ST+UMAP+Spectral)	0.56	0.02	0.93	0.97	72	6	0.66	0.01	0.93	0.98	55	10	0.66	0.07	0.92	0.98	354	337
GHTM (Proposed)	**0.91**	**0.31**	1.00	1.00	23	-	**0.87**	**0.27**	0.99	1.00	15	-	**0.82**	**0.28**	0.96	0.99	230	-

Table 4: Comprehensive model comparison across the datasets. Metrics: Coherence Value (CV), Normalized PMI (NPMI), Topic Diversity (TD), IRBO, Runtime in seconds (RT), and Topics Identified by model (*n*). Here, ST = Sentence-Transformer. The results are averaged across 3 runs. All RT and *n* values have been ceilinged to nearest integer.

k	CV	NPMI	TD	IRBO
20	0.735	0.173	0.910	0.993
50	0.659	0.135	0.580	0.972
100	0.624	0.105	0.523	0.980
Avg.	0.673	0.138	0.671	0.982

Table 5: Evaluation metrics for different values of k. Last row shows the average of metrics across $k = 20$, 50, and 100.

GHTM performed really well in terms of both coherence and diversity, as we can see in Table 4, Figure 4 and Figure 5. It also doesn't require much runtime, even when the dataset size grows. The overall results verify GHTM's efficacy over Bengali text data. The topics generated by GHTM along with the original categories is demonstrated in Table 5 for the dataset NCTBText.

Bigrams for performance boost: While generating the vocabulary, if we consider bi-grams along with unigrams, the performance of GHTM increases in every aspect. However, bi-grams were not employed in this study, to ensure uniformity in experimental setup across the models.

Cluster Size for Cluster-GCN: Too many clusters fragment the graph, degrading model performance, whereas too few clusters cause memory issues. Thus, while setting the 'num_clusters' parameter for Cluster-GCN, we've to maintain a balance depending on the dataset size, number of hidden layers, memory size etc. In our case, this parameter was set to k (the number of topics) after empirical testing showed it yielded the best model performance.

GHTM on English: We also tried GHTM on English benchmark dataset for topic modeling called 20NewsGroup. We tried different number of topic numbers, k = {20,50,100} and averaged the re-

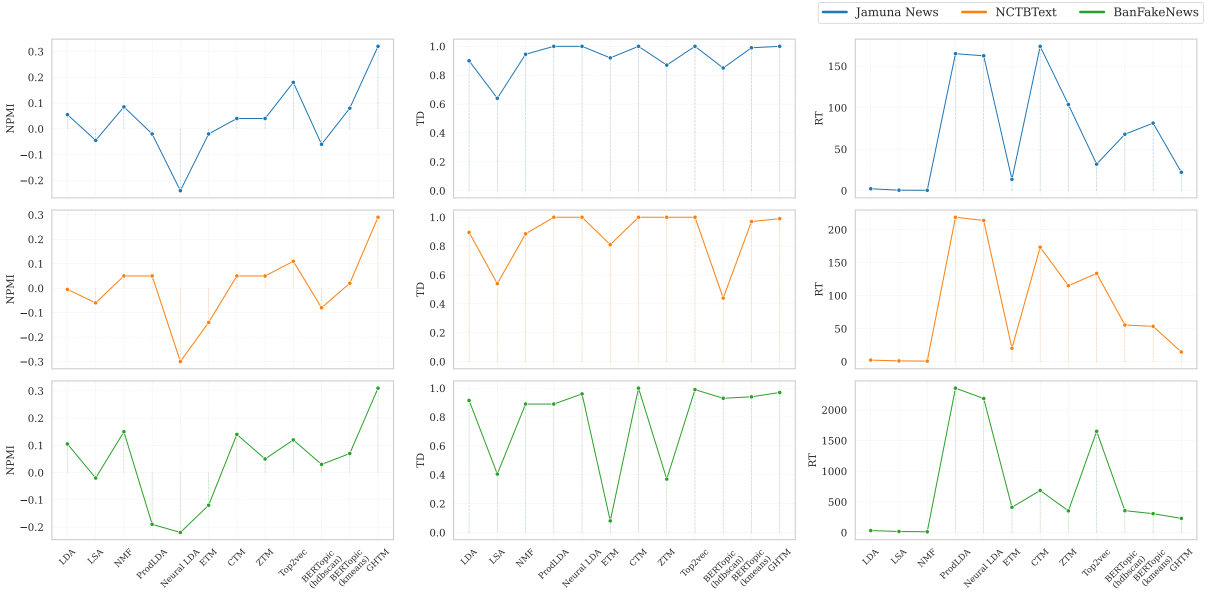

Figure 4: Comparative analysis of topic modeling performance across Jamuna News, NCTBText, and BanFakeNews datasets. Each row represents a dataset, with columns illustrating the metrics NPMI, TD and RT respectively. GHTM consistently outperforms baseline models in coherence and diversity.

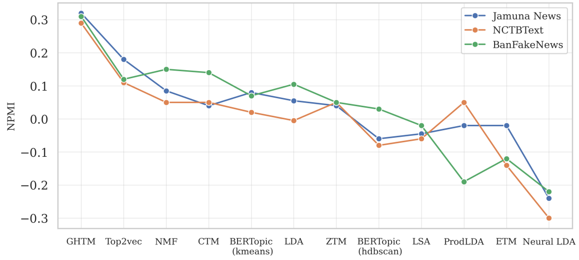

Figure 5: NPMI score comparison across datasets. Models are sorted across X axis based on performance. Results for BoW and TF-IDF were averaged here for LDA, LSA and NMF

Category	Keywords
Home Science	খাদ্য, খাবার, খাওয়া, মাংস, মাছ, আটা, ভিটামিন, পরিমাণ, দুধ, লবণ
Science	যৌগ, সমযোজী, যোজনী, আয়নিক, পরমাণু, মৌল, কপার, দ্রবীভূত, আয়ন, বন্ধন
Business	টাকা, ব্যাংক, নগদ, লেনদেন, ট্রেডার্স, আমানত, হিসাব, চেক, পরিশোধ, প্রদান
Social Science	বাংলা, খান, শাহ, শাসন, মুঘল, খলজি, দিল্লি, সুলতান, সুবাদার, সম্রাট
Religion	নবি, কুরআন, হাদিস, আল্লাহ, আল, রাসুল, তায়ালা, হযরত, কিতাব, নবুয়ত
Bengali Literature	বাংলা, রাষ্ট্রভাষা, পাকিস্তান, আন্দোলন, সাল, ঢাকা, খান, ভাষা, উর্দু, সুবাদার
Agriculture	জাত, ফসল, চাষ, ধান, পাট, ব্রি, সরিষা, জমি, ঈশ্বরদী, দোআঁশ
ICT	নেটওয়ার্ক, কম্পিউটার, মোবাইল, ফোন, তথ্য, যোগাযোগ, ইন্টারনেট, হাব, ক্লায়েন্ট

Figure 6: Visualization of topics generated by GHTM for the NCTBText dataset, distributed across original categories to demonstrate topic quality.

sults. The results are shown in Table 5. Considering 20NewsGroup dataset, GHTM outperforms GINopic (Adhya and Sanyal, 2025) in terms of CV (0.647) and NPMI (0.102).

5 Conclusion

Despite having a lot of potential in Bengali NLP, topic modeling is rarely utilized and studied for Bengali. This study verifies that, we can use topic models for generating coherent and latent themes from Bengali corpus. The topics extracted can be used to initially make sense of an unlabeled text dataset and further utilized for data annotation, which can contribute a lot in NLP tasks, especially for low-resource languages. Leveraging GCN and NMF, we have developed a hybrid topic modeling approach called GHTM, which significantly improves performance in topic coherence and diversity for Bengali dataset compared to existing models. The model shows promise and can be further advanced in the future for even better results and adapted for any language.

Limitations

This paper assumes topic numbers as a constant based on ground truth rather than a hyperparameter to be tuned. This approach forces models to generalize the topics to align with the authentic categories, because we intended find out which models can do it best. We strongly believe that topic models can act as a great starter for annotating unlabeled documents for low-resource languages. However, setting the topic number beforehand while not knowing the exact categories of an unstructured or unlabeled dataset, can be count as a limitation.

We hope to experiment further in future on how clustering works on GCN refined embeddings, so that we don't have to rely on topic numbers and let the model identify the latent topics by itself. Moreover, GHTM, being a neural network-based model, intensively uses GPU for the Cluster-GCN part and the runtime can become slower if ran on CPU only.

References

Sayan Adhya and Debarshi Sanyal. 2025. Ginopic: Topic modeling with graph isomorphism network. *arXiv preprint arXiv:2404.02115*.

Md. Basim Uddin Ahmed, Ananta Akash Podder, Mahruba Sharmin Chowdhury, and Mohammad Abdullah Al Mumin. 2021. A systematic literature review on english and bangla topic modeling. *Journal of Computer Science*, 17(1):1–15.

Md. Shahin Alam, Md. Saiful Rahman, and Md. Nazrul Islam. 2020. Topic modeling and trend analysis of bengali news articles. *International Journal of Computer Applications*, 176(34):1–7.

Dimo Angelov. 2020. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021a. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766, Online. Association for Computational Linguistics.

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021b. Cross-lingual contextualized topic models with zero-shot learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1676–1683, Online. Association for Computational Linguistics.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Dallas Card, Chenhao Tan, and Noah A. Smith. 2018. Neural variational document model for semi-supervised classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2153–2162.

Wei-Lin Chiang, Xuanqing Liu, Si Si, Yang Li, Samy Bengio, and Cho-Jui Hsieh. 2019. Cluster-gcn: An efficient algorithm for training deep and large graph convolutional networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 257–266. Association for Computing Machinery.

Sifat Dawn, Md. Saiful Rahman, and Md. Nazrul Islam. 2024. Likelihood corpus distribution: A dirichlet-polynomial clustering model for bengali topic modeling. In *Proceedings of the 2024 International Conference on Natural Language Processing (ICON)*, pages 100–110.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS)*.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Md. Mahmudul Hasan, Md. Saiful Rahman, and Md. Nazrul Islam. 2019. Lda2vec: Combining lda and word2vec for topic modeling in bangla. *International Journal of Computer Applications*, 177(28):1–7.

Mustakim Al Helal and Malek Mouhoub. 2018. Topic modelling in bangla language: An lda approach to optimize topics and news classification. *Computer and Information Science*, 11(4):77–77.

Md Zobaer Hossain, Md Ashraful Rahman, Md Saiful Islam, and Sudipta Kar. 2020. BanFakeNews: A dataset for detecting fake news in bangla. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020)*, pages 2862–2871, Marseille, France. European Language Resources Association (ELRA).

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*.

Daniel D. Lee and H. Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.

Yujie Luo, Hao Zhang, Yuxuan Wang, Ming Li, and Qiang Liu. 2024. Graph contrastive neural topic model. *arXiv preprint arXiv:2307.02078*.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108, Los Angeles, California. Association for Computational Linguistics.

Pintu Chandra Paul, Maqsudur Rahman, Amena Begum, and Md. Tofael Ahmed. 2025. Combining bert with lda: Improved topic modeling in bengali language. *IAENG International Journal of Computer Science*, 52(2):383–393.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 399–408. ACM.

Dazhong Shen, Chuan Qin, Chao Wang, Zheng Dong, Hengshu Zhu, and Hui Xiong. 2021. Topic modeling revisited: A document graph-based neural network perspective. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*.

Silvia Terragni, Bruno Galuzzi, Pietro Tropeano, Antonio Candelieri, Fabio Archetti, and Elisabetta Fersini. 2021. Octis: Comparing and optimizing topic models is simple! In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 263–270. Association for Computational Linguistics.

Md. Shihab Uddin, Mohd Ariful Haque, Rakib Hossain Rifat, Marufa Kamal, Kishor Datta Gupta, and Roy George. 2024. Bangla sbert - sentence embedding using multilingual knowledge distillation. In *2024 IEEE 15th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pages 495–500.

William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems*, 28(4):20:1–20:38.

Junxian Zhu, Yichao Jiang, Zhiting Li, Chengqing Zong, Qun Liu, and Eduard Hovy. 2018. Graphbtm: Graph enhanced autoencoded variational inference for biterm topic model. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4663–4672.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.