Human-Centric Intelligent Systems (2024) 4:53–76
https://doi.org/10.1007/s44230-023-00058-8

RESEARCH ARTICLE

# A Local Explainability Technique for Graph Neural Topic Models

**Bharathwajan Rajendran**[1] · **Chandran G. Vidya**[1] · **J. Sanil**[1] · **S. Asharaf**[1]

Received: 14 July 2023 / Accepted: 13 December 2023 / Published online: 12 January 2024


**Abstract**
Topic modelling is a Natural Language Processing (NLP) technique that has gained popularity in the recent past. It identifies word co-occurrence patterns inside a document corpus to reveal hidden topics. Graph Neural Topic Model (GNTM) is a topic modelling technique that uses Graph Neural Networks (GNNs) to learn document representations effectively. It provides high-precision documents-topics and topics-words probability distributions. Such models find immense application in many sectors, including healthcare, financial services, and safety-critical systems like autonomous cars. This model is not explainable. As a matter of fact, the user cannot comprehend the underlying decision-making process. The paper introduces a technique to explain the documents-topics probability distributions output of GNTM. The explanation is achieved by building a local explainable model such as a probabilistic Naïve Bayes classifier. The experimental results using various benchmark NLP datasets show a fidelity of 88.39% between the predictions of GNTM and the local explainable model. This similarity implies that the proposed technique can effectively explain the documents-topics probability distribution output of GNTM.

**Keywords** Explainable neural network · Graph neural topic model · Local explainable · Natural language processing · Topic modelling

**Abbreviations**

| | |
|---|---|
| 20NG | 20 News groups |
| BoW | Bag of words |
| XAI | Explainable artificial intelligence |
| GloVe | Global vectors for word representation |
| GNNs | Graph neural networks |
| GNTM | Graph neural topic model |
| LDA | Latent Dirichlet allocation |
| LIME | Local interpretable model agnostic explanation |
| NLTK | Natural langauge toolkit |
| NLP | Natural language processing |
| NTM | Neural topic model |
| PGM | Probabilistic graphical model |
| TMN | Tag my news |

Bharathwajan Rajendran, Chandran G. Vidya and S. Asharaf have contributed equally to this work.

✉ J. Sanil
sanil.j@duk.ac.in

Bharathwajan Rajendran
bharathwajan.cs21@duk.ac.in

Chandran G. Vidya
vidya.g@duk.ac.in

S. Asharaf
asharaf.s@duk.ac.in

[1] School of Computer Science and Engineering, Kerala University of Digital Sciences, Innovation and Technology, Thiruvananthapuram, Kerala, India

## 1 Introduction

Topic modelling as evidenced by [1–4] is a well-established probabilistic generative model popular within computer science, focusing mainly on text mining, document classification, information retrieval, summarization, and many others. Some diverse applications involving topic modelling are presented in these articles [1, 5–11]. Topic modelling is an unsupervised learning technique that finds the latent topics of a document corpus using the word co-occurrence patterns present in the documents. Developing a complete generative model, such as the Latent Dirichlet Allocation (LDA), marked the beginning of topic modelling [5, 12–15]. Even though topic modelling using LDA is a popular and effective technique, LDA is considered relatively complex, tends to be topic model specific, and may not provide an accurate model in practice.

Recent development in deep learning has given rise to much efficient topic modelling techniques, such as neural topic models, recurrent neural networks, variational

autoencoders, and transformer-based models. The research community considers the neural network-based models a black box function due to their complicated structure and nonlinearity. The exceptional precision of the models rarely justifies their reliability and complex user interactions. This obscure aspect might lead to issues if the model delivers incorrect results or malfunctions, particularly in fields like agriculture, forestry, health, and climate that affect human lives [16]. Explaining the neural network-based models can justify their decision-making process and outputs. With the advent of Explainable Artificial Intelligence (XAI), explaining the inner workings/predictions of obscure and sophisticated machine learning models was possible [17–21]. It helped researchers, developers, subject matter experts, and users better comprehend the model while utilizing its high performance and accuracy.

GNTM [22, 23] is a neural topic modelling technique [24, 25] that combines a graph neural network with a variational autoencoder. GNTM models the document corpus's underlying graph structure and generates the topic distribution for each document. The document relation graph of the corpus serves as the input of the graph neural network that extracts the relationship between documents. The documents and words in the corpus become a node in the document relation graph. The nodes are connected based on how often a document and a word appear together. Finally, the variational autoencoder finds the latent representations of documents and topics. The output of the GNTM consists of documents-topics and topics-words probability distribution matrices.

The practical ramifications of GNTM are significant. This technology demonstrates exceptional performance when documents display complex linkages since it effectively identifies topics that include content-based relevance and relational importance. The GNTM model represents a notable progression in topic modeling, effectively addressing the challenge of integrating content and context while analyzing large collections of documents.

However, the applicability of GNTM is often questioned due to its inherent black-box nature, not only within the context of topic modeling but also in various other application domains. This lack of transparency present in GNTM decision-making processes can impede its utility and explainability, limiting its applicability. In the context of topic modelling, researchers and domain experts require a clear understanding of how topics are generated and how specific words and documents are assigned to a particular topic to impart confidence and be able to use the model's results. Recognizing these challenges and limitations, the paper investigates the feasibility of explaining individual predictions of GNTM applied in a topic modelling context.

The main contribution of the paper is to develop a local explanation model equivalent to the GNTM. Decision trees, Naïve Bayes, and random forests are some probable candidates for the explanation model. This paper selects the Naïve Bayes classifier as the explanation model for its suitability to the specific NLP problem being addressed and as its predictions are more straightforward probabilistic calculations, making it relatively easy to understand how it arrives at its classifications or predictions. Experimental analyses demonstrate the similarities between the GNTM and the proposed explanation technique and how explanations are obtained.

The remainder of the paper is structured as follows: Sect. 2 describes relevant research works in the topic modelling context. The background theory of GNTM, Naive Bayes, the proposed methodology, and pseudocode are mentioned in Sect. 3. Section 4 discusses the details of NLP datasets, the experimental results showing the resemblance between the GNTM and trained local explainable model based on individual prediction, comparison with baseline topic modelling techniques, theoretical and practical implications, and limitations. The Section also demonstrates how explanation is achieved. Finally, Sect. 5 forms the conclusions and possible future directions.

## 2 Related Works

The inability of artificial intelligence and machine learning techniques to promote trust and acceptability due to their functional opaqueness has led to the concept of explainability. The international community has developed various techniques and methodologies to bring explainability to the present framework. This Section discusses relevant research on topic modelling, including graph neural network and explainable topic modelling techniques. The Section also identifies the research gap present in the context of topic modelling using GNTM.

A generative probabilistic model of a corpus known as the LDA is one of the most popular topic modelling techniques [12–15, 26, 27]. Modelling documents as a random mixture over latent topics, with each topic represented by a distribution over words, is the basic idea behind LDA. The need for an appropriate choice of distribution for each latent variable or the tolerance of laborious and case-by-case customized theoretical formulation restricts the flexibility and scalability of the model design. Traditional topic modelling techniques, including Markov Chain Monte Carlo [28, 29] or probabilistic variational inference [30], also suffer from this problem of flexibility and scalability.

The success of inferring topic models with a Variational autoencoder [31] has attracted more people to use

deep learning techniques in topic modelling contexts. A variational autoencoder is a parameterized neural network that uses variational distributions to estimate the posterior of latent variables. A generalization of LDA known as the Neural Topic Model (NTM) is proposed in the article [32]. Neural topic models are rapidly growing, resulting from combining topic modelling and deep neural networks. Neural understanding problems such as text generation, summarization, and language models have witnessed many neural topic models achieving excellent performance. A summary of the research progress and a discussion of outstanding issues and potential future approaches is presented in the article [33]. A NTM particularly suited for conversational scenarios known as the Conversational Neural Topic Model (ConvNTM) is proposed in [34], in which the topics are discovered by formulating the multi-turn structure in dialogues. Various variants of neural topic models for topic modelling [35] has been developed recently. Graph neural network is one such neural topic model. The literature [36] gives a comprehensive overview of neural graph networks categorized into convolution, recurrent, spatial-temporal graph neural networks, and graph autoencoder. The article also addresses the possibility of applying graph neural networks in various applications, including natural language processing.

Instead of viewing documents in a corpus as a bag of words or sequences, it is possible to represent them as a graph. This representation enables graph neural networks to extract latent topics. The work proposed in the literature [22] introduces a topic modelling technique using a graph neural network. The proposed Graph Neural Topic Model (GNTM) transforms each document in the corpus into directed graphs with edges representing word dependency between word nodes. The model learns using a neural variational inference approach to encode document graphs. Instead of using the word co-occurrence, the work proposed in [23] uses commonsense relationships to explicitly imply semantic relevance. A relational graph neural network is used to capture the relational information present inside the graph. Furthermore, manifold regularization imposes constraints on the documents' topic distributions. Another neural topic model, Graph Topic Model (GTM) [37], also represents the corpus as a graph representing the relationships between documents. In this graph, both documents and words in the corpus are indicated as nodes and linked to one another depending on the co-occurrence of words inside documents. The topical representation of a document node in GTM is derived by aggregating information from its multi-hop neighborhood, which consists of both document and word nodes. This aggregation process utilizes the Graph Convolutional Network (GCN) algorithm.

Several approaches, such as GNNExplainer [38], XGNN [39], SubgraphX [40], and Probabilistic Graphical Model (PGM)-Explainer [41], have been developed recently to explain the predictions obtained from graph neural networks. These methods are capable of providing instance and model-level explanations. Examples of instance-level explanation techniques include gradient/feature, perturbation, surrogate, and decomposition-based methods. An extension of the Local Interpretable Model Agnostic Explanation (LIME) [42] approach known as GraphLime [43], RelEx, and PGM-Explainer are surrogate methods. On the other hand, model-level approaches seek to provide general insights and high-level expertise to explain deep graph models. In particular, they look at what kinds of input graph patterns might cause a graph neural network to act in a given way, like maximizing a target prediction. Due to graph models' discontinuous graph topology information, a model-level explanation of the graph neural network will be complicated. XGNN belongs to the model-level explanation technique using graph generation. The article [44] explains the challenges, different instance and model-level methods of graph neural networks, and comparative analysis using other evaluation metrics. Graph Neural Networks (GNNs) often struggle to explain the discovered latent relations in a manner that ensures their reasonableness and independence. Additionally, GNNs often need textual content of edges, which is often not present in real-world datasets. Topic-Disentangled Graph Neural Network (TDG) proposed in the literature [45] aims to overcome these limitations. The proposed approach involves using a topic module to efficiently manage node attributes for the purpose of constructing distinct and explainable semantic subspaces. Subsequently, a neighborhood routing mechanism assigns appropriate relation topics to each graph connection based on their association with these subspaces.

A multi-modal causability technique in medical analysis using GNNs is developed by introducing multi-modal embeddings and interactive explainability in the research work [46]. Another work [47] presents a novel approach called the Semantic Reinforcement Neural Topic Model (SR-NSTM) that addresses the challenges of sparse and explainable text representation. To enhance the quality of text representations, the SR-NSTM model takes a closer look at the generation process of sparse topic models and integrates contextual information using Bi-LSTM. The literature introduces a Topic-Disentangled Graph Neural Network (TDG) to address the challenge posed by GNNs, which often struggle to explain the extracted latent relations and ensure their plausibility and independence. Additionally, this technique aims to overcome the need

for detailed textual information about these relationships, which is frequently absent in real-world datasets.

It's apparent from examining the existing research papers that numerous resources are dedicated to explaining GNNs and NTM. However, there's a noticeable lack of similar work explaining the outputs of the GNTM. This gap in providing explanations for GNTM results is identified as the research challenge that is being tackled in this paper.

## 3 Local Explainability Technique for Graph Neural Topic Models

This Section delineates the methodology of the proposed explainability technique for GNTM. The Section begins by providing a concise overview of the GNTM, local explainable model-Naive Bayes. The Section also gives the pseudo-code and a graphical representation of the proposed explainability technique.

### 3.1 Graph Neural Topic Model (GNTM)

GNTM [22] is a machine learning technique that combines graph neural networks and topic models to analyze and model large amounts of graph-structured text data. Graph neural network embeds the graph-structured text data formed from the datasets into a low-dimensional vector space. Furthermore, each vector in the low-dimensional vector space represents a document from the dataset. Furthermore, the documents are represented as directed semantic graphs, including word dependency as edges connecting the word nodes. The extraction of topics depends not only on the information inside individual documents but also on the evaluation of the impact of adjacent documents that are related by edges in the graph. The semantic structure of the texts is captured by propagating information along the network in accordance with the co-occurrence relationships between words in the documents. The GNTM tries to learn the low-dimensional vector representation (embedding) of each node in the graph. It then uses these embeddings to model the topics and relationships between the documents in the graph. Using both the graph structure and the text content of the documents, the GNTM can capture complex dependencies between the documents and make more accurate topic models than traditional topic models. The crux of the GNTM framework is in this interplay between the graph structure and topic modeling.

During the training process, GNTM acquires knowledge in two fundamental aspects, namely, document-topic distributions and topic-word distributions. The former examines the distribution of topics within individual documents, considering both the content and relational context. The latter pertains to characterizing the probability of each word's occurrence inside each topic. Using a dual learning process guided by the graph structure results in enhanced precision and context sensitivity in topic representations.

Graph neural topic model achieves better performance than other traditional topic modelling techniques at the cost of being highly complex. This complexity makes the GNTM opaque, making it difficult for the user to understand why the model arrived at a particular output. A proper explainability technique can make GNTM even more fruitful in human-centric applications.

### 3.2 Local Explainable Model: Naive Bayes

In this work, Naive Bayes classifier [48, 49] is selected to function as the local explainable model. The Naive Bayes Classifier is a probabilistic classification technique based on the concepts of the Bayesian Theorem, which Thomas Bayes first introduced. Its core objective in classification tasks is establishing an optimal mapping between a new data instance and a predefined set of classifications within a specific domain. Mathematical operations are used to transform joint probabilities into the product of prior and conditional probabilities to facilitate probabilistic computation for this mapping. The term "naive" in Naive Bayes refers to the assumption of conditional independence among the features included in the model. This implementation is a versatile toolkit applicable across a wide spectrum of classification domains.

Consider a dataset $x$ having n instances, $x_i, i = 1, 2, \ldots, n$ with 'm' attributes i.e., $x_i = (x_{i1}, x_{i2}, x_{i3}, \ldots, x_{im})$. Assume that each instance belongs to one and only class, $y \epsilon (y_{i1}, y_{i2}, y_{i3}, \ldots, y_{in})$. Naive Bayes learning pertains to the development of a Bayesian probabilistic model that gives a posterior class probability to a particular data instance $P(Y = y_i|X = x_i)$. The basic Naive Bayes classifier utilizes these probabilities to assign an instance to a class. The basic formulation of Naive Bayes applying Bayes theorem [50] is given in Eq. 1

$$P(y_i|x_i) = \frac{P(x_i|y_i) \times P(y_i)}{P(x_i)} \tag{1}$$

The numerator in Eq. 1 represents the joint probability between $x_i$ and $y_i$. The numerator can be written as given in Eq. 2

$$P(x|y_i) \times P(y_i) = P(x_1|x_2, x_3, \ldots, x_m, y_i) \cdot P(x_2|x_3, x_4, \ldots, x_m, y_i) \ldots, \cdot P(x_m|y_i) \times P(y_i) \tag{2}$$

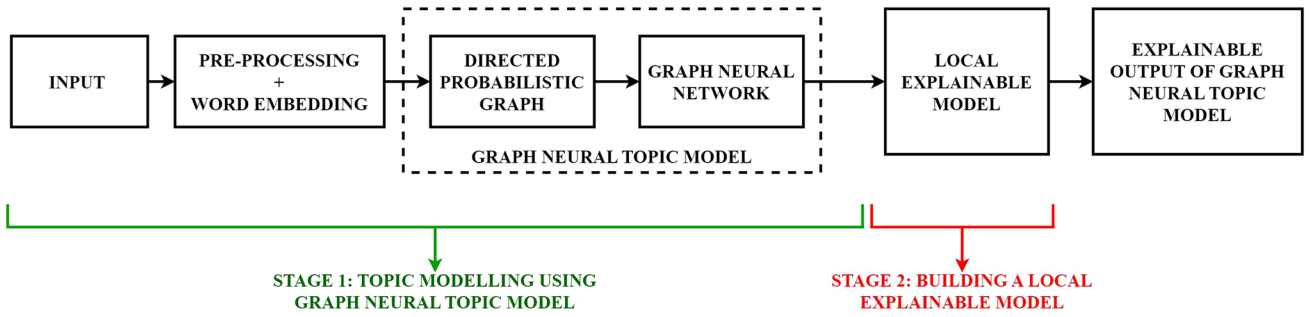



**Fig. 1** Schematic representation of proposed local explainability technique for graph neural topic model

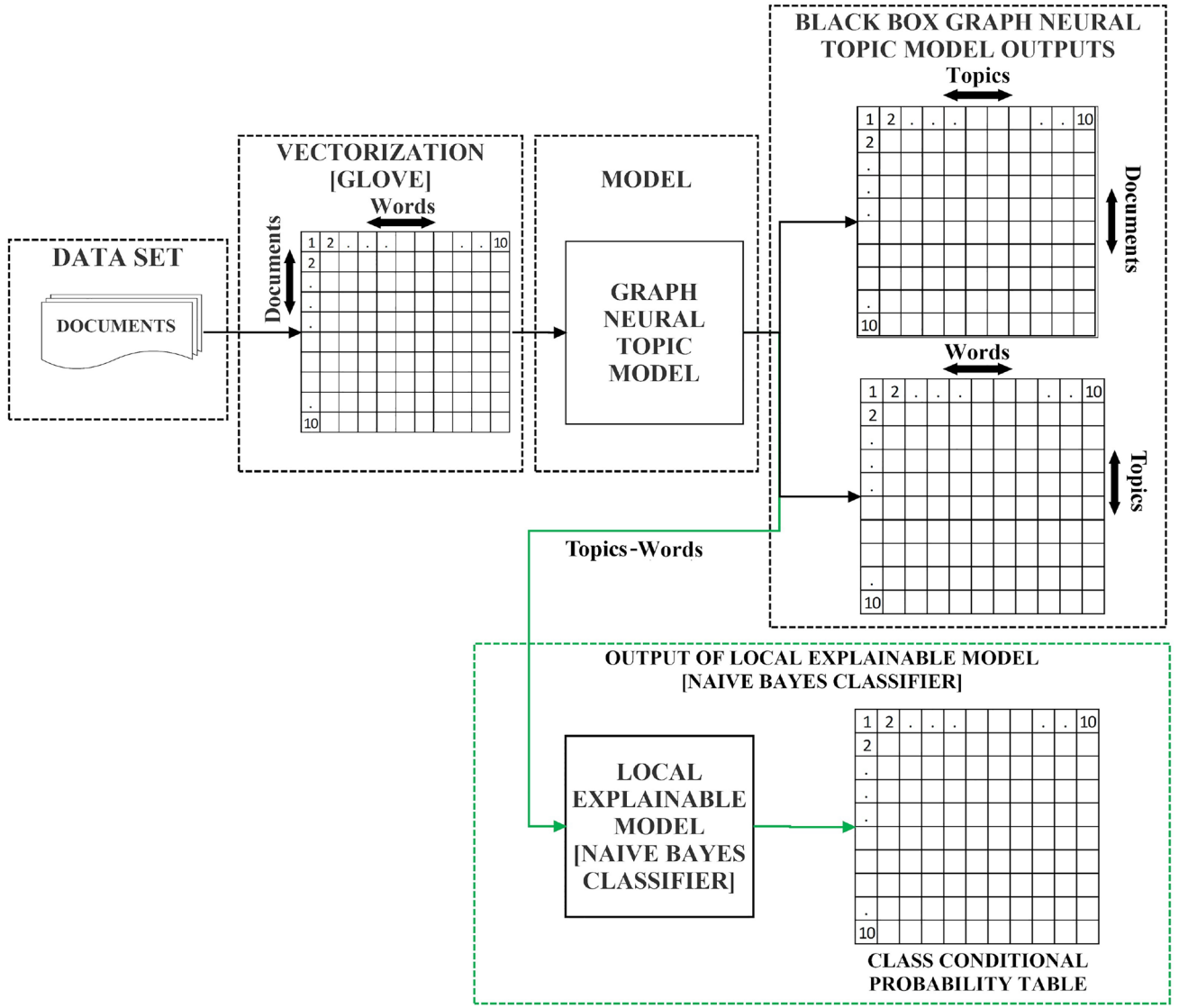



**Fig. 2** Training of proposed local explainability technique for graph neural topic models

Assuming the data instances $x_i$ are independent to each other, $P(x_1|x_2, x_3, \ldots, x_m, y_i) = P(x_1|y_i)$. Thus, the numerator in Eq. 1 reduces to Eq. 3.

$$P(x|y_i) \times P(y_i) = P(x_1|y_i) \cdot P(x_2|y_i) \ldots \cdot P(x_m|y_i) \cdot P(y_i) = \prod_{k=1}^{m} P(x_k|y_i) \cdot P(y_i) \quad (3)$$

Substituting Eq. 3 in Eq. 1, the basic formulation of Naive Bayes becomes as given in Eq. 4.

$$P(y_i|x) = \frac{\prod_{k=1}^{m} P(x_k|y_i) \cdot P(y_i)}{P(x)} \quad (4)$$

The term $P(x)$ in the denominator of Eq. 4 is independent of any output class $y_i$. The term serves as a scaling factor and can be excluded from calculation due to the assumption that each instance belongs to one and only one class. Thus, Eq. 4 the mathematical formulation of Naive Bayes becomes,

$$P(y_i|x) = \prod_{k=1}^{m} P(x_k|y_i) \cdot P(y_i). \quad (5)$$

### 3.3 Proposed Explainability Technique for Graph Neural Topic Models

The GNTM in a topic modelling context generate two outputs, namely,

1. Documents topics distribution
2. Topics words distribution

The proposed approach tries to provide an explanation of the documents topics distribution output of the GNTM. Explainability is achieved by training a local explainable model. Suppose the trained model produces a documents topics distribution similar to the GNTM. In that case, it is possible to say that, a black box model such as the GNTM can be explained by introducing an inherently explainable local model. The overall flow of the proposed approach is as follows.

1. The corpus is first preprocessed.
2. Using the preprocessed corpus, topics words distributions are generated using a GNTM.
3. A vectorized labelled training dataset is generated with the topics words distribution.
4. A local explainable model is built using the labelled training dataset.

This approach builds a local explainable model equivalent to a GNTM. The schematic representation of the proposed explainability technique, highlighting the various steps involved, is depicted in Fig. 1. The proposed technique comprises of two stages: topic modelling using a GNTM and building a local explainable model. In this paper, the function of the local explainable model is carried out using a Naive Bayes classifier.

*Stage 1:* Topic Modelling using Graph Neural Topic Model: The pre-processed training dataset is converted to a dense vector representation of words using Global Vectors for Word Representation (GloVe). The GNTM uses the graph-structured data obtained by converting the vector representations to form the topic models. The topic models contain a list of words associated with different topics. The formation of the topic model, including various pre-processing techniques, is adapted from the literature [22].

*Stage 2:* Building a Local Explainable Model: A labelled training dataset with words as features and topics as class labels are formed using the topic model generated by the GNTM. The labelled training dataset is vectorized using Bag of Words (BoW). The local explainable model is built by training the model using the labelled vectorized training dataset. The trained local model classifies the new, unseen documents into any one of the predefined topics.

Figure 2 is a pictorial illustration of the training of the proposed explainability technique using the Naive Bayes classifier. From Fig. 2, it can be noted that the vectorized documents is used to train the GNTM. The GNTM provides two outputs in matrices. The first matrix provides the probability distribution of words over topics. The second matrix, a probability distribution of topics over documents, is the individual prediction of GNTM, which the proposed technique explains with the help of a local explainable model. The Naïve Bayes classifier is trained using the probability distribution of words over topics (first matrix). During topic modelling of a new unknown document, the class conditional probability table of the trained Naïve Bayes classifier is used.

Figure 3 illustrates the testing of the proposed topic modelling explainability technique. The test documents are vectorized (BoW) using the trained GNTM. Naïve Bayes classifier acts on these vectorized documents generating the individual prediction by forming the documents- topics probabilities.

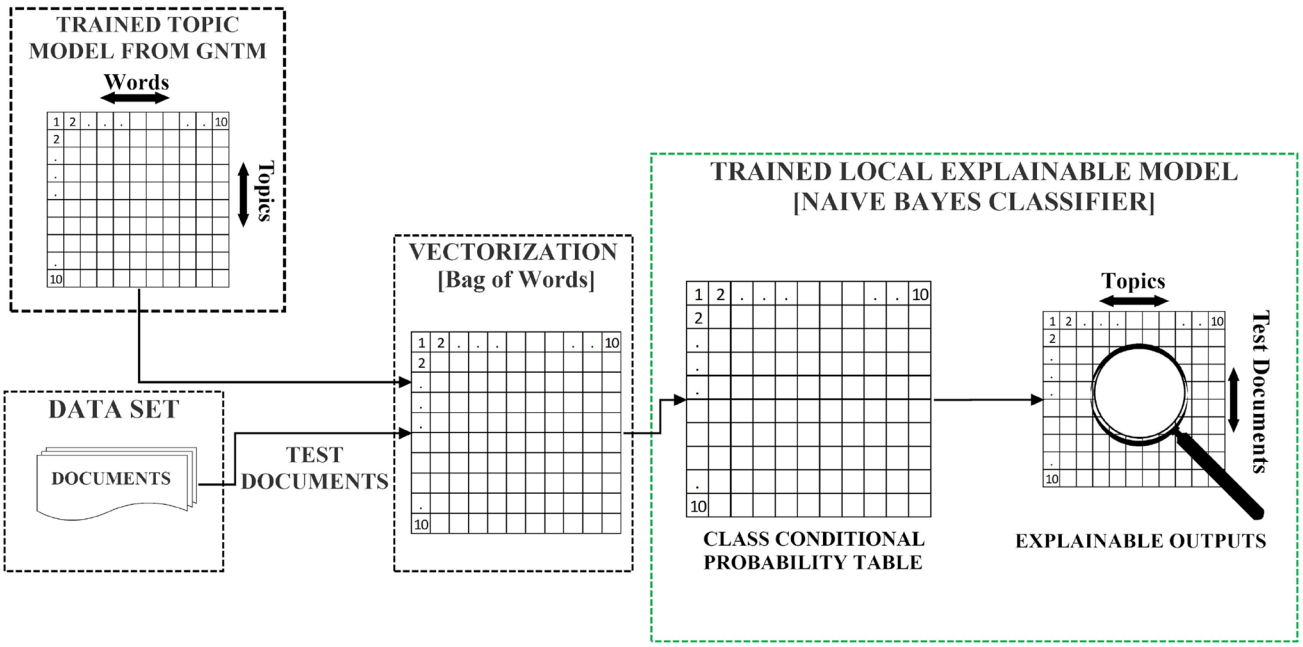



**Fig. 3** Testing of proposed local explainability technique for graph neural topic models

**Algorithm 1** Pseudo code of the Local Explainability Technique For Graph Neural Topic Model

**Input** : $tr, ts$
**Output** : $d\_td$
**begin**
Training
$trp = pre\_proc(tr)$
**for** each $trp\_doc$ **do**
  $vct\_tr(trp\_doc) = vect\_GloVe(trp\_doc)$
  $gsd(trp\_doc) = gsd\_fn\big(vct\_tr(trp\_doc)\big)$
**end for**
$tm = gnn\_fn(gsd)$
$tmlab = label\_fn(tm)$
**for** each $tmlab\_rec$ **do**
  $vct\_tmlab(tmlab\_rec) = vect\_BoW(tmlab\_rec)$
**end for**
$tr\_mod = LM(vct\_tmlab)$
Testing
$tsp = pre\_proc(ts)$
**for** each $tsp\_doc$ **do**
  $vct\_ts(tsp\_doc) = vect\_BoW(tsp\_doc)$
**end for**
$d\_td = tr\_mod(vct\_ts)$
**end**

**Table 1** Notations and definitions of attributes used in Algorithm 1

| Sl. no. | Notations | Definition |
|---|---|---|
| 1. | *tr*, *ts* | Training and testing datasets |
| 2. | *d_td* | Document topic distributions |
| 3. | *trp*, *tsp* | Pre-processed training and testing datasets |
| 4. | *trp_doc*, *tsp_doc* | Documents in pre-processed training and testing datasets |
| 5. | *pre_proc* | Pre-processing function |
| 6. | *vect_GloVe* | GloVe vectorization function |
| 7. | *vct_tr*, *vct_ts* | Vectorized training and testing datasets |
| 8. | *gsd_fn* | Graph structured data function |
| 9. | *gsd* | Graph structured data |
| 10. | *gnn_fn* | Graph neural topic model function |
| 11. | *tm* | Topic model |
| 12. | *label_fn* | Topic model labelling function |
| 13. | *tmlab* | Labelled Topic model |
| 14. | *tmlab_rec* | Records in labelled topic model |
| 15. | *vect_BoW* | BoW vectorization function |
| 16. | *vct_tmlab* | Vectorized labelled topic model |
| 17. | *LM* | Local Model (Naïve Bayes) |
| 18. | *tr_mod* | Vectorized labelled topic model |

Algorithm 1 gives a pseudo-code of the proposed explainability technique, and the definition of notations used in Algorithm 1 is given in Table 1. Pseudocode implementation of the proposed local explainability technique begins with two inputs, namely training and testing datasets *tr*, *ts*, respectively. The training dataset *tr* is first preprocessed before performing vectorization and forming the graph-structured data. Vectorization is done using GloVe. The graph-structured data is then be applied to the GNTM function *gnn_fn* to form the topic model *tm* associated with the dataset. Using this topic model, a labelled topic model *tmlab* is generated, and each record in the labelled topic model contains a specified number of words associated with each topic. The labeled topic model, vectorized using BoW, forms the vectorized labeled topic model denoted as *vct_tmlab*. This vectorized representation is used to train the local explainable model, which, in this case, is Naive Bayes. This completes the training of the proposed technique. Testing is done by first performing preprocessing and vectorization using BoW. It should be noted here that the BoW used for vectorization of the testing dataset should align with that in the training phase for consistent performance. The vectorized testing dataset is given to the trained model to form the topics-document distribution. A process flow diagram of Algorithm 1 is illustrated in Fig. 4.

## 4 Experimental Results and Discussion

The experimental analyses are carried out to establish the similarity between GNTM and the proposed local explainability technique for GNTM. The similarity is evaluated at the corpus level and document level. At the corpus level, similarity is assessed in terms of the Euclidean distance among the normalized average topic mix and topic-wise word cloud. Regarding document level similarity evaluation, the document level percentage of topics matched is calculated. Once the similarity is established, the details of how the explanations are achieved are demonstrated. The similarity is evaluated using three NLP datasets of varying dimensions. Also, the performance of the proposed explainability technique for GNTM is compared with a few baseline topic modelling techniques. The datasets include Reuters-21578 https://www.kaggle.com/datasets/nltkdata/reuters, 20 News Groups (20NG) https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_20newsgroups.html, and Tag My News (TMN) https://www.kaggle.com/datasets/rmisra/news-category-dataset. The dimensional details of the datasets and the train-test split-up are given in Table 2.

The experimental analyses are carried out using a laptop with an 11th-generation Intel(R) Core(TM) i5-1135G7 CPU running at 2.40 GHz and 16 GB of RAM. The scripts of the techniques are implemented in Python 3.9.7, and various natural language processing functions are used from the Natural Langauge Toolkit (NLTK) (NLTK 3.8.1) available at https://www.nltk.org/.

### 4.1 Experimental Results

This Section focuses on the various experiments to evaluate the similarities between GNTM and the proposed local explainability model. Evaluating similarity is a key part of determining if the proposed local explainability model can successfully mimic GNTM. The similarity is evaluated at the corpus and document levels. In corpus-level similarity evaluation, the similarity between the GNTM and local explainability technique for the GNTM is calculated using Euclidean distance among the normalized average topic mix and topic-wise word cloud. At the document level, a similar evaluation is done by assessing the percentage of matched topics and with respect to fidelity between GNTM and the proposed technique. The latter part of the experimentation also expresses an evaluation of GNTM and the local explainable model using a specific test sample from different datasets. This evaluation of similarities will facilitate the development of an explanation for the GNTM, as demonstrated in Sect. 4.2. A comparison with a few baseline topic

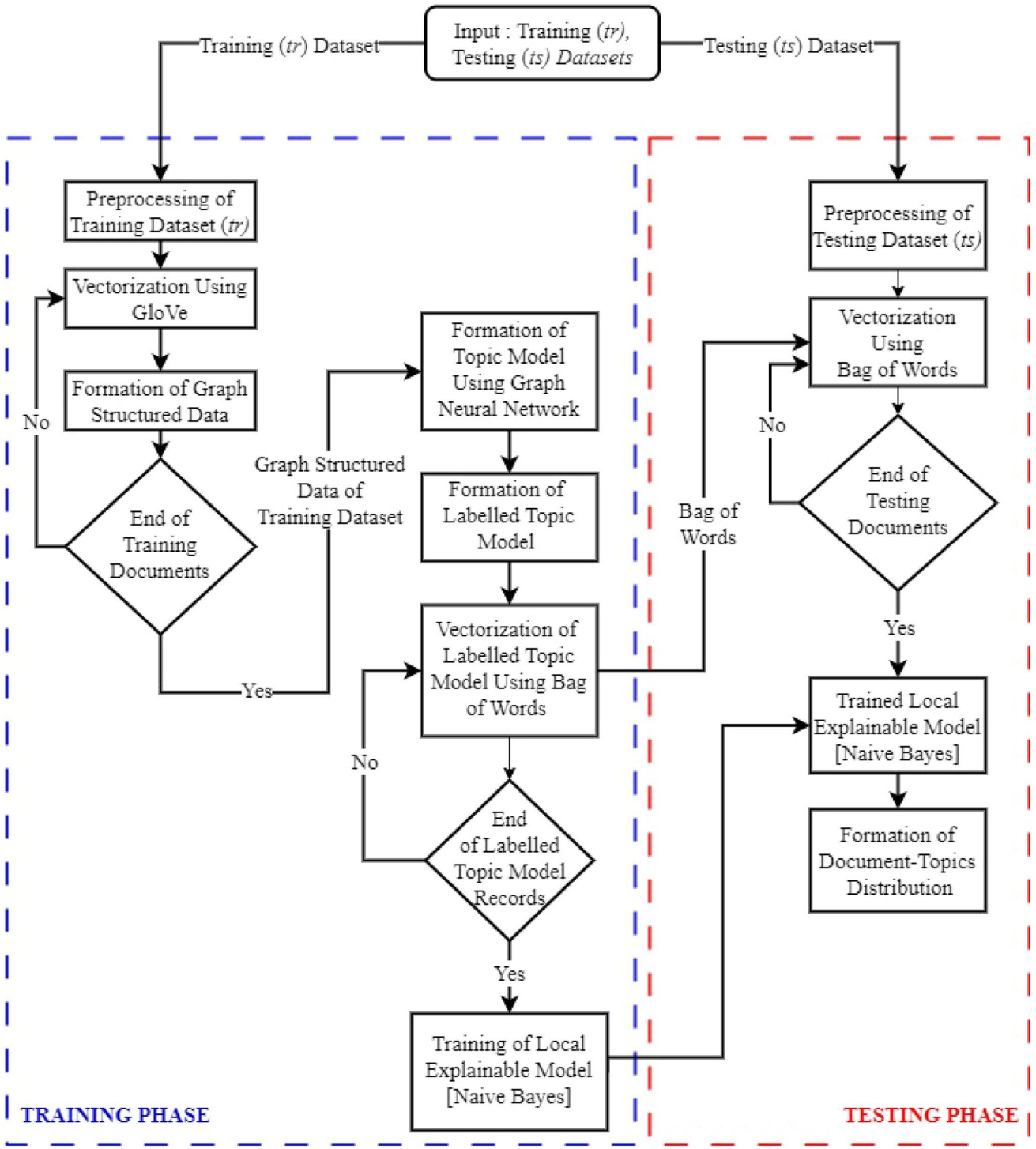



**Fig. 4** Process flow diagram of Algorithm 1

**Table 2** Specifics of benchmark datasets

| Dataset | Total no. of documents | Documents taken | Data split (train/val/test) | Vocabulary | Word token | Edge set | Edge token |
|---|---|---|---|---|---|---|---|
| Tag my news (TMN) | 20006 | 18926 | 12094/2626/4206 | 8402 | 272719 | 2381 | 58031 |
| 20 news groups (20NG) | 18846 | 16506 | 6649/3319/6538 | 12866 | 1276916 | 16767 | 334903 |
| Reuters-21578 | 10788 | 10717 | 6619/950/2897 | 5228 | 1362647 | 28803 | 2631581 |

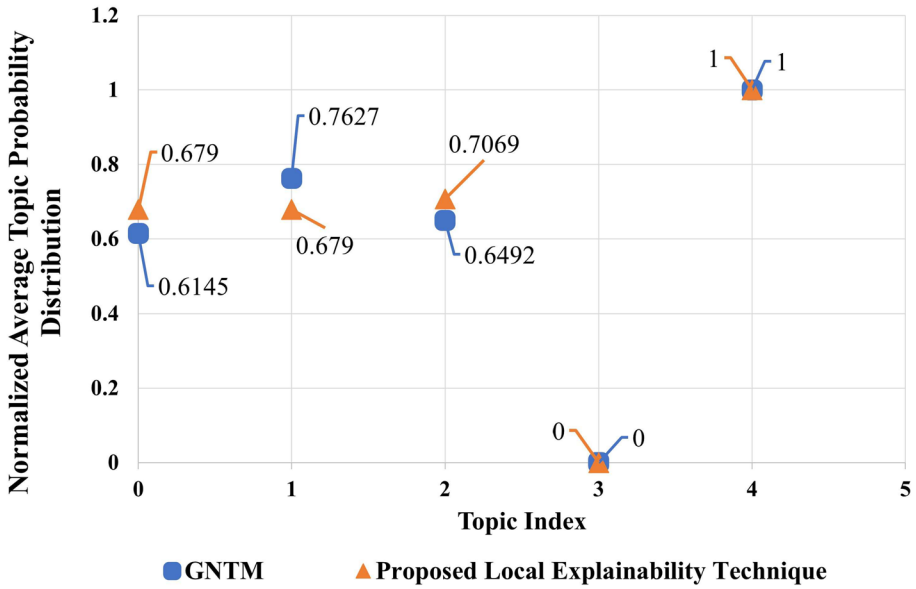



**Fig. 5** Normalized average topic mix obtained using graph neural topic model and proposed local explainability technique for graph neural topic model of five topics (Tag My News dataset) in two-dimensional Euclidean space

**Table 3** Euclidean distance between normalized average topic mix obtained using local explainability technique for graph neural topic model and graph neural topic model for different datasets [Best results bolded]

| Dataset | Topic index | | | | |
|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 |
| Tag my news | 0.0644 | 0.0837 | 0.0577 | **0.0** | **0.0** |
| 20 news groups | 0.6134 | 0.0588 | 0.0553 | **0.0** | 0.1067 |
| Reuters-21578 | 0.1930 | 0.0738 | **0.0** | 0.0605 | 0.2407 |

modelling techniques is also carried out to better position the performance of the proposed technique.

#### 4.1.1 Corpus Level Similarity Evaluation in Terms of Euclidean Distance Among Normalized Average Topic Mix

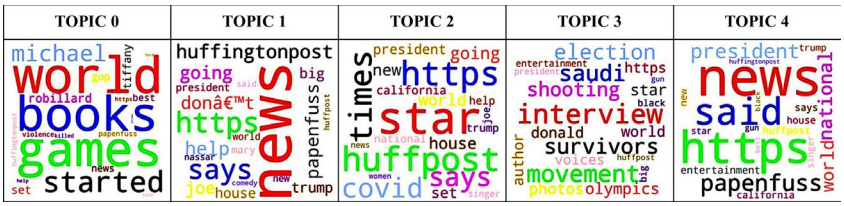
This Section evaluates similarity by calculating the Euclidean distance between the normalized average topic mix obtained using the GNTM and the proposed local explainability technique.

The document-topics probability distribution for each document is evaluated first and averaged over the entire corpus to obtain the average topic mix. The average topic mix is then normalized using min–max normalization. The similarity is evaluated in terms of the Euclidean distance [51, 52]. Equation (6) gives the mathematical formula to find the Euclidean distance between two points in Euclidean space.

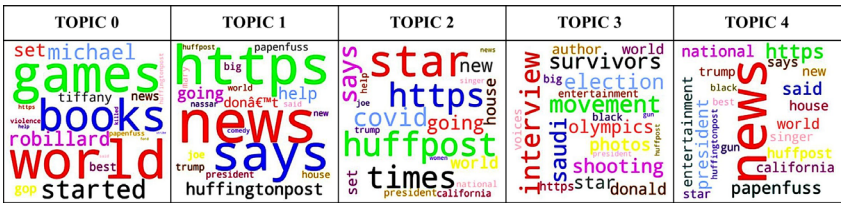
$$E_d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{6}$$

where $E_d$ represents the Euclidean distance, $(x_1, y_1)$ and $(x_2, y_2)$ represents the coordinates of two points.The normalized average topic mix obtained using the GNTM and local explainability technique for the GNTM for five model topics represented in two-dimensional Euclidean space using the Tag My News dataset is shown in Fig. 5.

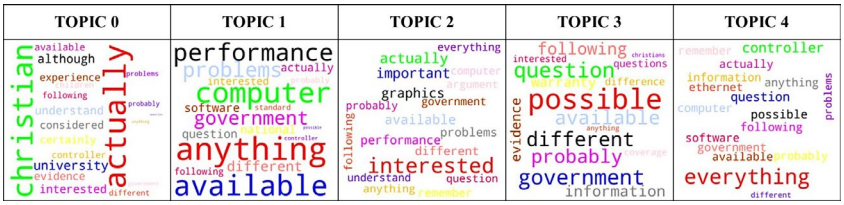
In Fig. 5, the x and y-coordinates denote the topic index and normalized average topic probability distribution, respectively. It can be observed that the x-coordinates are the same. Hence, the Euclidean distance given in Eq. (6) is the difference between the y-coordinates, i.e., the normalized average topic probability distribution. Thus, Eq. (6) becomes,



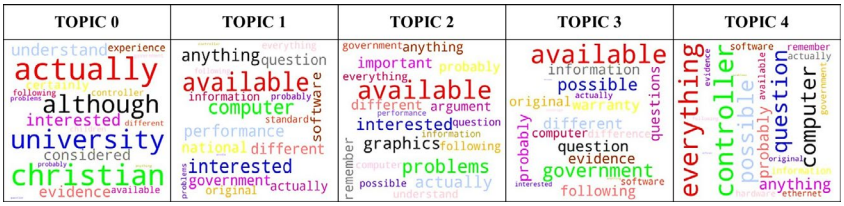

(a) Graph Neural Topic Model - Tag My News Dataset



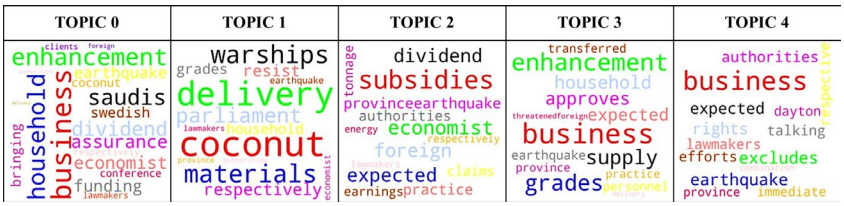

(b) Proposed Local Explainability Technique for Graph Neural Topic Model - Tag My News Dataset



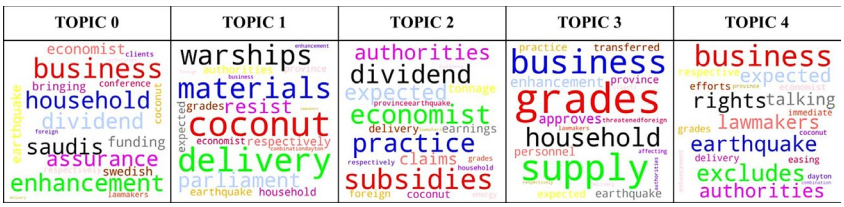

(c) Graph Neural Topic Model - 20 News Groups




(d) Proposed Local Explainability Technique for Graph Neural Topic Model - 20 News Groups




(e) Graph Neural Topic Model - Reuters-21578




(f) Proposed Local Explainability Technique for Graph Neural Topic Model - Reuters-21578

**Fig. 6** Topic-wise word cloud of graph neural topic model and proposed local explainability technique for graph neural topic model for different datasets with five model topics, **a** Tag My News, **b** 20 News Groups, **c** Reuters-21578 (Red, Green, and Blue indicates first, second and third top words)

$$E_d(t) = |y_{nb}(t) - y_{gntm}(t)| \tag{7}$$

where $t$ denotes topic index, $E_d(t)$ is the $t$th topic Euclidean distance, $y_{nb}(t)$ and $y_{gntm}(t)$ represents the normalized average topic probability of the $t$th topic obtained using local explainability technique for GNTM and GNTM respectively.

The Euclidean distance evaluated for five topics of three datasets using Eq. (7) is given in Table 3. From the Table, the Euclidean distance obtained is very close to zero, indicating that the local explainability technique for the GNTM can produce outputs close to that of the GNTM. In fact, topics 3 and 4 in Tag My News, topic 3 in 20 News Groups, and topic 2 in Reuters-21578 are zero (indicated in bold), indicating that both techniques are identical in modelling the respective topics.

#### 4.1.2 Corpus Level Similarity Evaluation in Terms of Topic-Wise Word Cloud

In this Section, the similarity between the topic models obtained using GNTM and the proposed local explainable model is showcased by visualizing it using topic-wise word clouds. Analyzing these topic-wise word clouds makes it possible to identify the prominent words present in each topic obtained using both models. Furthermore, the frequency of words within a particular topic can also be visualized by the size of words in the word cloud. The prominent three words in the topic-wise word cloud prepared over the entire corpus are given the same color format for a straightforward interpretation of similarity. Figure 6 illustrates the topic-wise word cloud for five model topics evaluated using three datasets obtained with the local explainable model and GNTM. Figure 6a, c, e represent the topic-wise word cloud over corpus obtained using GNTM with Tag My News, 20 News Groups, and Reuters datasets, respectively. Similarly, the topic-wise word cloud obtained using the local explainability technique for the GNTM is illustrated in Fig. 6b, d, f.

Considering Fig. 6a, it can be observed that the top three prominent words that constitute topic 0 of the topic model obtained using GNTM are 'world,' 'games,' and 'books.' On comparing these with the prominent words of topic 0 obtained using the proposed local explainability method, as shown in Fig. 6b, it can be noted that they are the same. This indicates that the proposed local explainability method can produce outputs similar to GNTM. This behaviour is also notable in the case of other topics and with all datasets. This indicates that the local explainability technique for GNTM can produce prediction outputs similar to that obtained using GNTM.

#### 4.1.3 Document Level Similarity Evaluation in Terms of Percentage of Topics Matched

The average document-wise percentage of topics matched is calculated by comparing how the topics are distributed with respect to their probabilities in the documents as obtained using both techniques.

The average document-wise percentage of topics matched is evaluated in two scenarios. The first scenario evaluates the matching by considering whether the top n percent of model topics are present in both outputs. The second scenario considers the presence and the order in which these model topics are present in the results.

The percentage of topics matched is determined for the top 80, 60, and 40 percent of model topics. Initially, the total number of topics in the output of topic modelling techniques is fixed at five. Later, the number of topics increases in steps up to 15, and its effect on matching percentage is studied. The results obtained are shown in Figs. 7 and 8. The red, green, and blue lines indicate the percentage matching obtained using 20NG, Reuters-21578, and TMN datasets.

**4.1.3.1 Percentage of Topics Matched Considering only The Presence of The Top n Percent of Model Topics** Here, the top n percent of model topics from the GNTM outputs and the proposed technique are compared without considering their order. Figure 7 depicts the average document-wise percentage of topics matched for different top n percent of model topics. The total number of model topics is varied in the case of different datasets, and its effect on the matching percentage is also illustrated in the figure. Figure 7a–c give the percentage of topics matched for the top 80, 60, and 40 percent model topics, respectively. From the Figure, the following observations can be made. For the top 80 percent of model topics shown in Fig. 7a, it can be noted that the overall percentage of topics matched is above 80 percent. A similar result is also observed for the top 60 percent of model topics. In the case of the top 40 percent of model topics, the average document-wise percentage of topics matched is around 80 percent, except for 12 topics. As a concluding note, in most cases, the portion of topic matched stayed above 80 percent.

**4.1.3.2 Percentage of Topics Matched Considering The Presence as well as The Order of Top n Percent of Model Topics** In addition to the presence of the top n percent of model topics in outputs obtained using GNTM and the proposed technique, the order in which these topics are present in the output documents is considered. The average document-wise percentage of topics matched applied in different data-

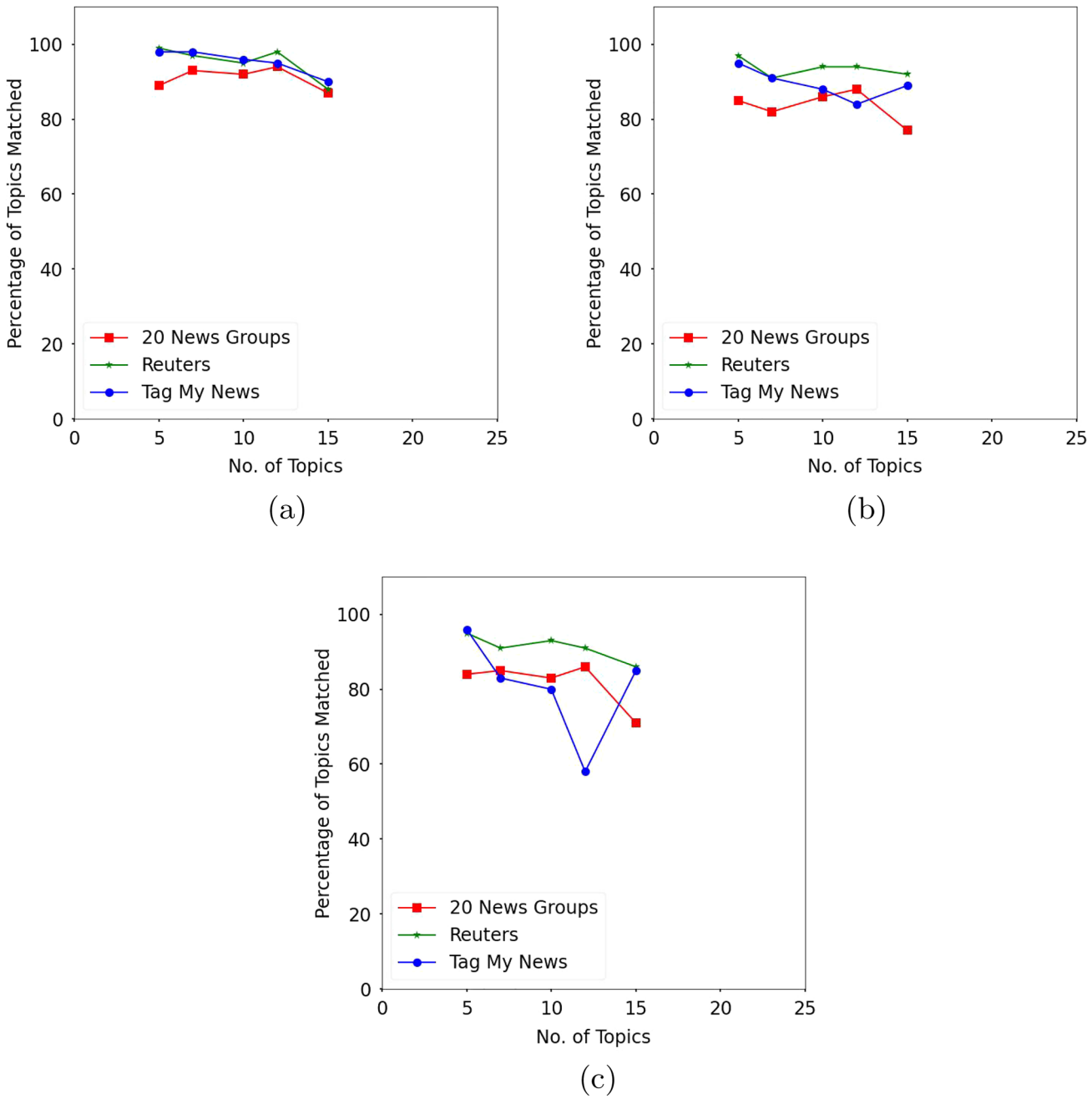



**Fig. 7** Percentage of topics matched (without order) for different top n percent model topics in the case of different benchmark datasets, **a** top 80 percent model topics, **b** top 60 percent model topics, **c** top 40 percent model topics

sets is depicted in Fig. 8. Figure 8a–c illustrate the results obtained for various top n percent model topics matched. When the order of model topics is also considered, the overall percentage of topics matched for five topics is above 80 percent in the case of the top 80, 60, and 40 model topics. As the number of model topics increased, a similarity of around 60 percent is observed.

#### 4.1.4 Document Level Similarity Evaluation in Terms of Fidelity

In this Section, the overall performance of the local explainability technique is evaluated in terms of fidelity using the entire test set of the three datasets. Fidelity, as described in [53, 54], makes sure that the proposed local explainability technique captures how GNTM makes decisions so that

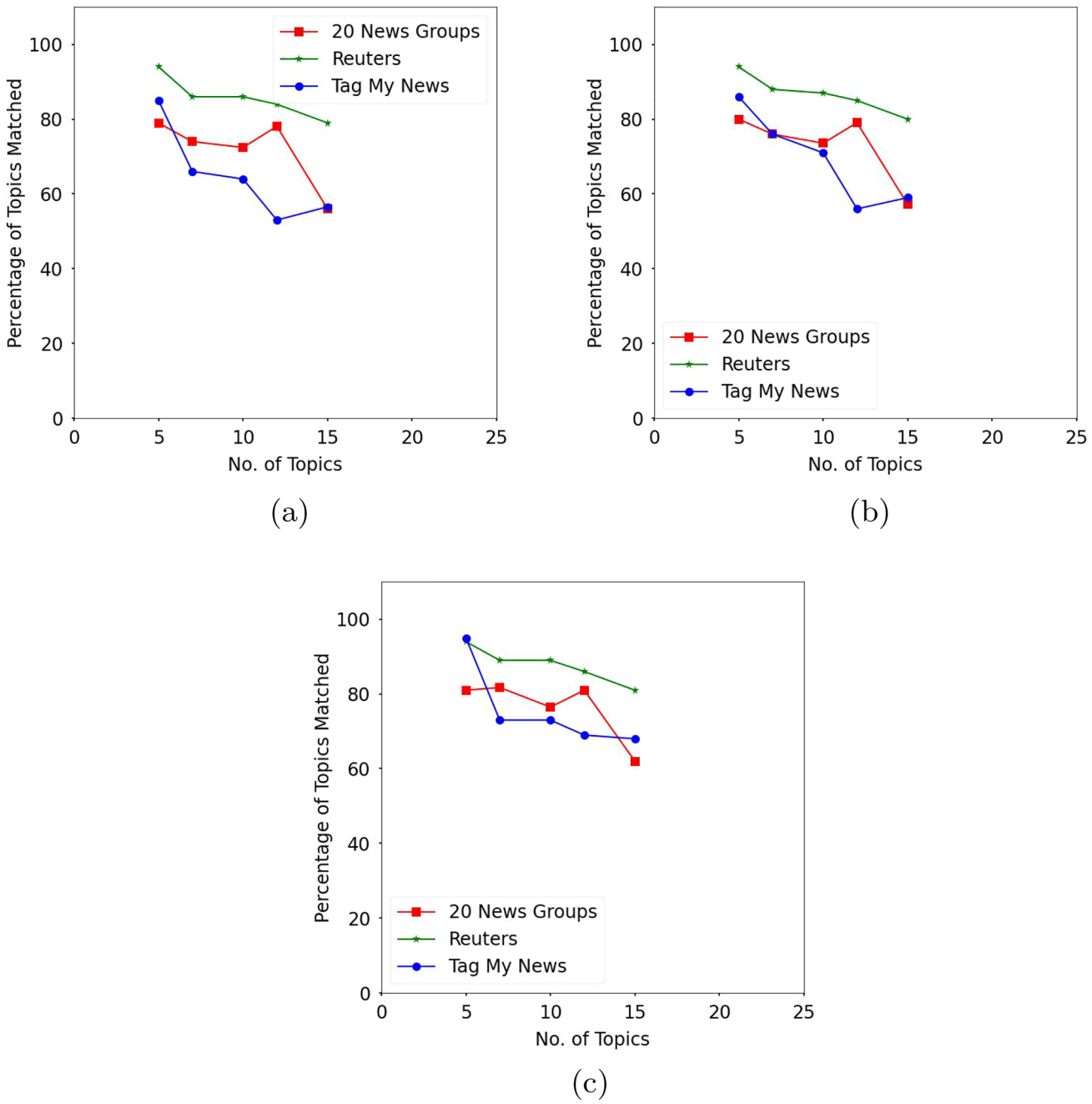



**Fig. 8** Percentage of topics matched (with order) for different top n percent model topics in the case of different benchmark datasets, **a** top 80 percent model topics, **b** top 60 percent model topics, **c** top 40 percent model topics

the explanations that can be used to understand GNTM are accurate.

The fidelity of the local explainability technique obtained using three datasets is given in Fig. 9. From the Figure, it can be noted that the fidelity of the proposed local explainability technique for GNTM is 88.39%, 78.9% and 83.4% with TMN, 20NG and Reuters datasets. This shows that the proposed technique can capture the decision process of GNTM with respectable accuracy.

The following conclusions can be made from all the experimental analyses conducted in this Section.

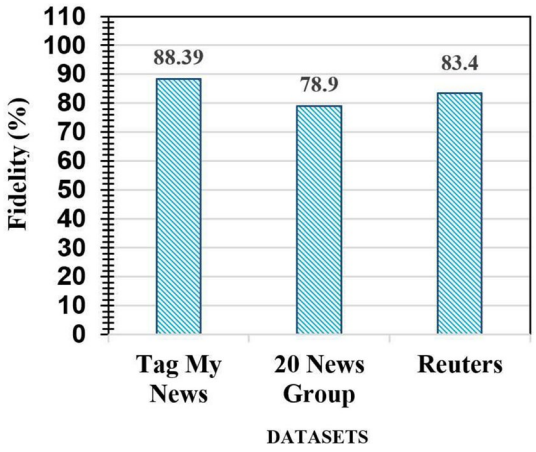



**Fig. 9** Fidelity of Proposed Local Explainability Technique for Graph Neural Topic Model using Tag My News, 20 News Group and Reuters-2157 Datasets

**Table 4** Baseline comparison of proposed local explainability technique for GNTM in terms of topic diversity and topic coherence across five topics with 20NG

| Parameters | Baseline methods | | | | Proposed explainability technique |
|---|---|---|---|---|---|
| | LDA | LSA | BERTopic | GNTM | |
| Topic diversity | 0.2080 | 0.2512 | 0.0988 | 0.4169 | 0.4169 |
| Topic coherence | 0.6926 | 0.7114 | 0.7279 | 0.7238 | 0.7238 |

- The Euclidean distance between the normalized average topic mix over the corpus obtained using the GNTM and local explainability technique for the GNTM was found close to zero. This indicated that both techniques produced similar document topic probability distributions.
- The topic-wise word cloud over the entire corpus illustrated the resemblance between the top frequently occurring words in the output obtained using both techniques.
- Document-wise percentage of topics matched calculated by comparing the topic distribution obtained using both methods indicated a matching percentage of nearly 80 percent when the presence of the top n percent of model topics was considered. Similarly, the proposed technique produced more than 60 percent similarity when the presence and order of top n percent model topics were considered.
- Document-wise similarity evaluation in terms of fidelity showed that the proposed explainability technique has a fidelity measure of 88.39%, 78.9% and 83.4% using TMN, 20NG and Reuters datasets respectively.

The above conclusions indicate that the proposed local explainability technique for the GNTM could mimic the GNTM at respectable levels.

#### 4.1.5 Baseline Comparison of Proposed Local Explainability Technique

In order to better position the performance of the proposed local explainability technique for GNTM, a comparison with baseline topic modelling techniques is carried out. Baseline algorithms include Latent Dirichlet Allocation (LDA) [55, 56], Latent Semantic Analysis (LSA) [57], BERTopic [58], and GNTM. Being an unsupervised technique, the performance of topic modelling techniques is assessed in terms of topic diversity and topic coherence.

Table 4 gives the performance comparison for the proposed explainability technique across five topics using 20NG dataset. The Table shows that the proposed technique's topic diversity and coherence align with the baseline topic modelling techniques.

An illustration demonstrating the performance of the proposed explainability technique for GNTM across five and fifteen topics using 20NG, TMN and Reuters datasets are also given in Fig. 10. The Figure shows that, as seen in Table 4, the topic diversity and topic coherence of the proposed technique for five and fifteen topics also align with the baseline topic modelling techniques.

Thus, from Table 4 and Fig. 10, it can be observed that,

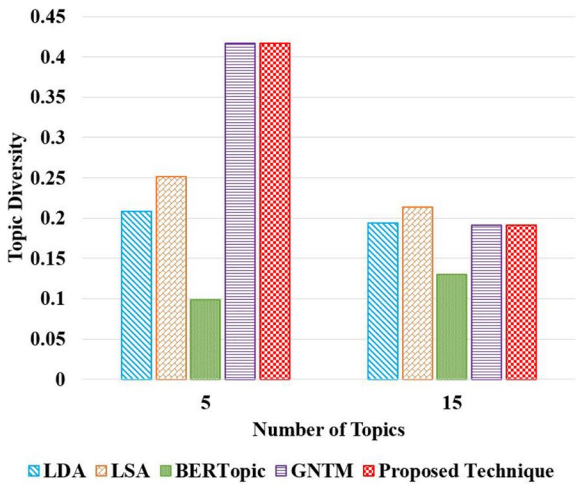
1. The performance of the proposed explainability technique for GNTM is comparable with the performance of the baseline modelling techniques.
2. Explainability (Demonstrated in Sect. 4.2) is achieved without compromising the performance.

#### 4.1.6 Evaluation of Graph Neural Topic Model and Proposed Local Explainability Technique Using Diverse Test Samples

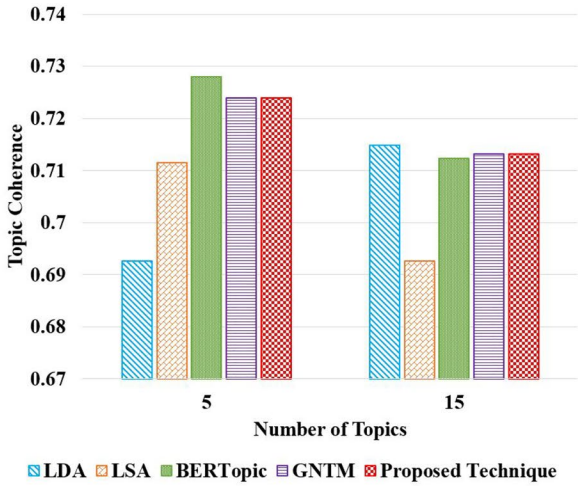
In this Section, three test samples from 20NG,TMN, and Reuters-2157 are selected and evaluated using GNTM and local explainability technique for the GNTM.

Topic modelling is carried out with 5 topics, each comprising 15 words. The topic modelling obtained using GNTM trained using the three datasets are given in Table 5.

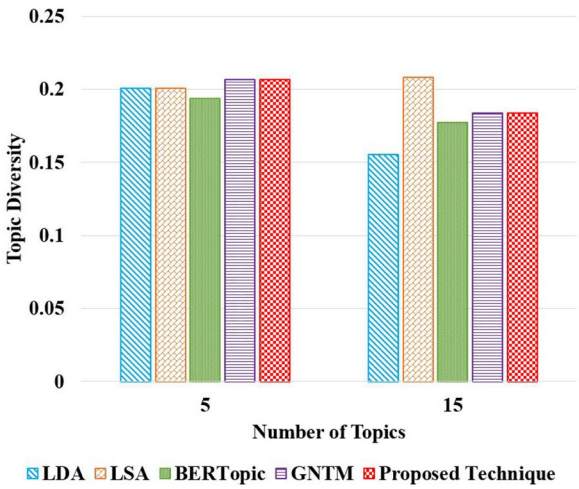
The three sample test documents selected are presented in Fig. 11a–c. Table 6 presents the probability distribution of topics across the individual selected test samples from three datasets, using the GNTM and the local explainability technique with the GNTM. Based on the data presented in the Table, it is


0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Topic Diversity
5
15
Number of Topics
LDA LSA BERTopic GNTM Proposed Technique

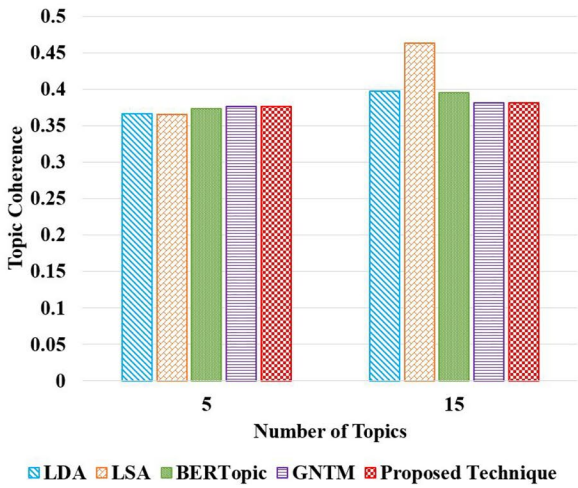
(a)


0.74
0.73
0.72
0.71
0.7
0.69
0.68
0.67
Topic Coherence
5
15
Number of Topics
LDA LSA BERTopic GNTM Proposed Technique

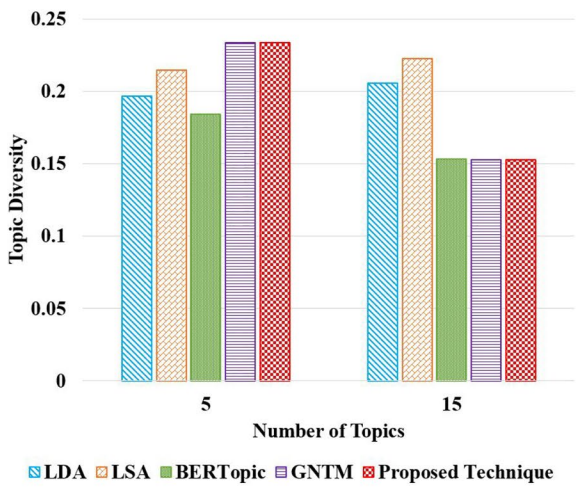
(b)


0.25
0.2
0.15
0.1
0.05
0
Topic Diversity
5
15
Number of Topics
LDA LSA BERTopic GNTM Proposed Technique

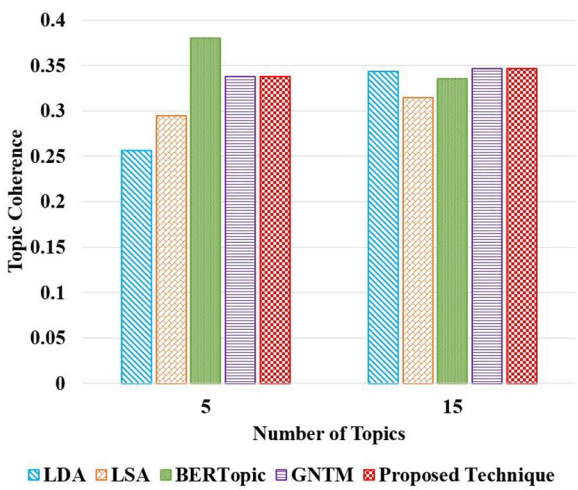
(c)


0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Topic Coherence
5
15
Number of Topics
LDA LSA BERTopic GNTM Proposed Technique

(d)


0.25
0.2
0.15
0.1
0.05
0
Topic Diversity
5
15
Number of Topics
LDA LSA BERTopic GNTM Proposed Technique

(e)


0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Topic Coherence
5
15
Number of Topics
LDA LSA BERTopic GNTM Proposed Technique

(f)

◂**Fig. 10** Baseline comparison of the proposed GNTM explainability technique across various topic numbers using different benchmark datasets, **a** Topic Diversity [20NG], **b** Topic Coherence [20NG], **c** Topic Diversity [TMN], **d** Topic Coherence [TMN], **e** Topic Diversity [Reuters-2157], **f** Topic Coherence [Reuters-2157]

apparent that, according to the probability distribution of topics derived from applying GNTM to the 20NG dataset, Topic 2 is the most prominent, followed by Topic 4. Conversely, Topics 0,1 and 3 demonstrate the least prominence. Upon performing an in-depth analysis, it is evident that the probability distribution of topics derived from the proposed local explainability technique exhibits similar behavior, with Topic 2 emerging as the most prevalent topic and the rest of the topics also following the same order as seen with GNTM. This phenomenon is also seen in the TMN and Reuters-2157 datasets.

## 4.2 Explainability of Graph Neural Topic Model: A Demonstration

The explainability of the proposed technique is demonstrated using a sample test document taken from Tag My News dataset, as shown in Fig. 11a. A GNTM configured with five model topics is trained using the Tag My News dataset. The topic models generated using the GNTM for the Tag My News dataset with a maximum of 5 topic words are shown in Table 5.

A labelled vectorized form of topic models obtained using a GNTM is used to train the Naïve Bayes classifier. The trained Naive Bayes classifier generates a class conditional probability table, and the classifier uses this Table to make predictions on any unknown documents.

The class condition probability table obtained is given in Table 7, with a value of 0.036 if a particular word is present in a topic and a probability value of 0.018 if the word is absent. The class conditional probability table is limited to the first five words (for illustration purposes), as shown in Table 7. It is possible to deduce from Tables 5 and 7 that the word "actor" appears only in Topic 1 and Topic 4. As a result, the probability value for these two topics for the word is 0.036, and the probability value for other topics is 0.018. The formation of the class conditional probability table of the Naïve Bayes classifier concludes the training.

After the formation of topic model, testing of the sample test document is carried out. Let the test document after pre-processing be "tweeters school amazingly dumb Donald Trump over special council typo politics the special council is a unit of the space force Lee Moran". The vectorized test document after pre-processing is given in Table 8.

Table 8 indicates the presence/absence of words in the test document with respect to the words present in the topic model given in Table 5. If a particular word is present, the value is 1; otherwise, it is treated as 0.

The individual prediction, i.e., the probability distribution of topics over the test document obtained using the trained Naïve Bayes classifier for Tag My News dataset [Extracted from Table 6 for ease of interpretation] is given in Table 9.

The mathematical formulation of the Naïve Bayes classifier given in Eq. (5) can be rewritten to obtain the prediction probability for the sample test document in the topic modelling context as given in Eq. (8).

$$P(T_k|W) = \prod_{i=1}^{N} P(W_i|T_k) \times P(T_k) \tag{8}$$

where $T_k$ represents the $k$th topic, $W$ represents the set of words from the test document which is present in the topic model, $i$ is the index of the words in $W$, $N$ is the total number of words in $W$, $W_i$ gives the $i$th word in $W$, $P(T_k|W)$ represents the probability of the $k$th topic given all the words in $W$, $P(W_i|T_k)$ represents the probability of $i$th word in $W$ given $k$th topic, $P(T_k)$ is the probability of $k$th topic among all topics.

As Naïve Bayes is an interpretable model, it is possible to calculate the probability distribution of topics over the test document as given in Table 9 using the basic mathematical formulation of Naïve Bayes given in Eq. (8). $P(W_i|T_k)$ which is the probability of $i$th word in the pre-processing test document given $k$th topic is obtained with the help of the class conditional probability table given in Table 7 and the vectorized test document represented in Table 8.

Suppose the probability of zeroth topic in the document topic probability matrix is calculated then, k = 0;

Therefore, Eq. (8) implies

$$P(T_0|W) = \prod_{i=1}^{N} P(W_i|T_0) \times P(T_0) = 0.25 \tag{9}$$

The value obtained in Eq. (9) is the same as obtained by the Naïve Bayes classifier, as shown in Table 9. The details of the calculation of $P(T_0|W)$ are given in Fig. 12. Similarly, all other probabilities for topics 1, 2, 3, and 4 can be calculated. On observing Table 9, it can be noted that the test document is confined to topics 0, 2, and 4. The test document probability distribution obtained using the GNTM is given in Table 10 (Extracted from Table 6 for ease of interpretation.), which also has the same relationship among topics as in Table 9.

The demonstration presented in this Section shows how explanations of individual predictions of GNTM can be made using the proposed technique.

Tweeters School ‘Amazingly Dumb’ Donald Trump Over ‘Special Council’ Typo POLITICS “” The Special Council is a unit of the Space Force.”” Lee Moran”

(a) Tag My News

B
BK>Is it possible to plug in 70ns or 60ns SIMMs into a motherboard saying
BK>wants 80ns simms?
You shouldn't have troubles. I have heard of machines having problems with slower than recommended memory speeds, but never faster.
BK>Also, is it possible to plug in SIMMs of different
BK>speeds into the same motherboard? ie - 2 megs of 70ns and 2 megs of 6
BK>or something like that? Sure. I have 4 70ns SIMMs in one bank and 4 60ns SIMMS in the other ( I have a 486 ). I wouldn't recommend mixing speeds within a bank, just to
be on the safe side.
-rdd
rdesonia@erim.org
---
. WinQwk 2.0b#0 |. Unregistered Evaluation Copy
* KMail 2.95d W-NET HQ, hal9k.ann-arbor.mi.us, +1 313 663 4173 or 3959

(b) 20 News Group

b'U.S. CONSUMER PRICES ROSE 0.4 PCT IN FEBRUARY\n' b' U.S. consumer prices, as measured by\n' b' the Consumer Price Index for all urban consumers (CPI-U), rose\n' b' a seasonally adjusted 0.4 pct in February after a 0.7 pct\n' b' January gain, the Labor Department said.\n' b' The CPI for urban wage earners and clerical workers (CPI-W)\n' b' rose to 329.0 in February, the department said.\n' b' Prices for petroleum-based energy rose sharply for a second\n' b' consecutive month during February but by less than in January,\n' b' the department said.\n' b' Energy prices rose 1.9 pct last month after a 3.0 pct rise\n' b' in January, accounting for one-third of the overall CPI rise.\n' b' For the 12 months ended in February, the CPI rose an\n' b' unadjusted 2.1 pct.\n' b' Transportation prices rose 0.5 pct in February after a 1.5\n' b' pct increase in January. Smaller price rises for motor fuels\n' b' and declines in new car prices and finance charges were\n' b' responsible for the moderation.\n' b' Gasoline prices rose 4.2 pct last month after increasing\n' b' 6.6 pct in January, but were still 18 pct below levels of a\n' b' year ago, the department said.\n' b' Housing prices rose 0.4 pct in February after a 0.5 pct\n' b' January increase, largely due to a rise in fuel oil prices.\n' b' Fuel oil prices were up 4.4 pct in February after\n' b' increasing 9.8 pct in January, but were still 15 pct below\n' b' price levels of February 1986.\n' b' Food prices rose 0.2 pct last month after a 0.5 pct January\n' b' increase. Grocery store food prices were up 0.4 pct, the same\n' b' as in January, but meat, poultry, fish and eggs cost less for a\n' b' third consecutive month, the department said.\n' b' Medical care rose 0.3 pct in February to a level 7.1 pct\n' b' above one year ago, because of higher costs for prescription\n' b' and non-prescription drugs and medical supplies, the department\n' b' said.\n' b' The index for apparel and upkeep rose 0.7 pct in February\n' b' after a 0.4 pct increase in January. The department said the\n' b' introduction of higher priced spring merchandise, particularly\n' b" men's clothing, was responsible for the advance.\n" b' Prices for other goods and services rose 0.7 pct in\n' b' February following a 1.1 pct increase in January.

(c) Reuters-2157

**Fig. 11** Sample Test Document from Tag My News, 20 News Group and Reuters-2157 Datasets

**Table 5** Topic modelling obtained using graph neural topic model for five topic words with Tag My News, 20 News Group, and Reuters-2157 Datasets

| Dataset | Topics | Words |
|---|---|---|
| Tag my news | Topic 0 | https, world, Trump, house, white, politics, election, Jordan, magazine, Donald, report, bologna, caroline, says, like |
| | Topic 1 | Entertainment, huffingtonpost, says, https, house, pruitt, dicker, election, politics, scott, world, said, human, reportedly, actor |
| | Topic 2 | Entertainment, https, world, Donald, Moran, politics, actors, Trump, house, married, black, said, huffpost, Putin, Jordan |
| | Topic 3 | https, coming, world, elections, politics, power, entertainment, book, John, Delbyck, cole, republicans, race, huffpost, election |
| | Topic 4 | huffingtonpost, https, politics, coming, white, says, entertainment, world, house, said, Trump, actor, John, report, Chris |
| 20 news group | Topic 0 | Believe, launch, matter, parents, world, books, point, players, bring, information, light, boards, right, windows, development |
| | Topic 1 | Armenian, round, court, drives, building, going, source, problem, information, second, thought, right, asking, looking, tried |
| | Topic 2 | Images, years, interested, little, price, think, possible, single, windows, government, point, right, problem, local, group |
| | Topic 3 | Email, windows, problem, information, running, question, years, following, power, probably, works, right, thanks, think, point |
| | Topic 4 | Windows, software, information, program, source, right, point, address, things, problem, thing, years, available, second, going |
| Reuters-2157 | Topic 0 | Authorities, foreign, clients, lawmakers, respectively, conference, bringing, Swedish, funding, economist, assurance, dividend, Saudis, household, enhancement |
| | Topic 1 | Household, materials, province, coconut, delivery, parliament, enhancement, grades, respectively, foreign, lawmakers, combination, Dayton, resist, warships |
| | Topic 2 | earthquake, energy, respectively, authorities, business, claims, dividend, province, economist, earnings, grades, tonnage, practice, subsidies, foreign |
| | Topic 3 | Foreign, personnel, expected, supply, grades, approves, practice, enhancement, threatened, affecting, transferred, clients, household, delivery, province |
| | Topic 4 | Lawmakers, rights, enhancement, expected, business, delivery, excludes, immediate, Dayton, province, earthquake, talking, easing, respective, combination |

**Table 6** Probability distribution of topics over test documents of Tag My News, 20 News Group, and Reuters-2157 Datasets using Graph Neural Topic Model and Proposed Local Explainability Technique of Graph Neural Topic Model

| Dataset | Model | Topic 0 | Topic 1 | Topic 2 | Topic 3 | Topic 4 |
|---|---|---|---|---|---|---|
| Tag my news | GNTM | 0.2608 | 0.1304 | 0.3043 | 0.0869 | 0.2173 |
| | Proposed explainability technique | 0.2500 | 0.0625 | 0.5000 | 0.0625 | 0.125 |
| 20 news group | GNTM | 0.0555 | 0.1111 | 0.5555 | 0.1111 | 0.1666 |
| | Proposed explainability technique | 0.00191 | 0.0038 | 0.9827 | 0.0038 | 0.0076 |
| Reuters-2157 | GNTM | 0.1818 | 0.2727 | 0.2727 | 0.1818 | 0.0909 |
| | Proposed explainability technique | 0.1538 | 0.3076 | 0.3076 | 0.1538 | 0.0769 |

**Table 7** Class conditional probability table of first five words in the topic model of the proposed local explainability technique using Tag My News dataset

| Topics | Words | | | | |
|---|---|---|---|---|---|
| | Actor | Actors | Black | Bologna | Book |
| Topic 0 | 0.018 | 0.018 | 0.018 | 0.018 | 0.018 |
| Topic 1 | 0.036 | 0.018 | 0.018 | 0.018 | 0.018 |
| Topic 2 | 0.018 | 0.036 | 0.036 | 0.018 | 0.018 |
| Topic 3 | 0.018 | 0.018 | 0.018 | 0.018 | 0.036 |
| Topic 4 | 0.036 | 0.018 | 0.018 | 0.018 | 0.018 |

## 4.3 Discussion

### 4.3.1 Theoretical Implications

Development of the proposed local explainability technique as a means of explanation for GNTM has significant theoretical ramifications within the field of XAI. Theoretical significance comes from its potential to convert the opaque character of GNTM into a more easily understood framework. The ability of the proposed technique to comprehend and clarify the relationships between topics and documents

**Table 8** Vectorized test document

| Words | Vectorized Value |
|---|---|
| Actor, actors, black, bologna, book, caroline, Chris, cole, coming, Delbyck, dicker, election, elections, entertainment, house, https, huffingtonpost, huffpost, human, John, Jordan, like, magazine, married, power, Pruitt, Putin, race, report, reportedly, republicans, said, says, Scott, white, world | 0 |
| Donald, Moran, politics, Trump | 1 |

**Table 9** Probability distribution of topics over test document from Tag My News dataset using local explainability technique for graph neural topic model

| Document | Topic 0 | Topic 1 | Topic 2 | Topic 3 | Topic 4 |
|---|---|---|---|---|---|
| Test Document | 0.25 | 0.0625 | 0.5 | 0.0625 | 0.125 |

within the GNTM framework presents novel opportunities for theoretical investigation in graph-based topic modeling. These findings can potentially be applied across diverse domains, including natural language processing, social network analysis, and recommendation systems.

Another theoretical implication of the proposed technique is that all the theories built on explainable models can be applied to black-box models, and their validity can be evaluated. This opens up a wide array of opportunities for understanding and harnessing the potential of black box models in various applications while maintaining a certain level of interpretability and transparency.

#### 4.3.2 Practical Implications

The proposed explainability technique allows researchers to understand the factors that influence predictions. The achieved level of transparency provides the ability to make well-informed choices and effectively address any issues related to model behavior. Consequently, this generates more dependable and practical insights in various applications.

In the context of health care applications such as disease prediction, the proposed explainability technique enables the model to reveal the essential traits and circumstances present in a patient's medical history that contributed to the prediction of a given illness or medical condition. The information systematically emphasizes relevant patient data, including symptoms, laboratory test results, medical records, and underlying risk factors. The degree of clarity at hand provides healthcare providers with diverse skills. First and foremost, this enables healthcare specialists to verify the model's suggestions by comparing them with their clinical experience and the patient's real health condition. Furthermore, equipped with a comprehensive understanding of the factors that influence the predictions made by the model, healthcare practitioners can develop customized treatment strategies to meet the particular requirements of individual patients, including distinct risk factors and medical backgrounds taken into account. Finally, within the field of medical research, this practice not only facilitates the discovery of novel ideas but also enhances the refinement of research procedures, thus propelling the progress of the healthcare and medical science domains.

When it comes to e-commerce applications such as loan dispersal or credit card approval also, explainability of the model is crucial. This ensures that applicants are provided with transparent and understandable justifications for the results of their applications, whether they are approved or denied. Proper justifications enable financial institutions to ensure that their decisions are based on objective criteria, reducing the potential for bias or discriminatory practices. This will also enable the applicant to understand the factors that influence their application status, such as credit score, income, or outstanding debts, and also promote the applicant to maintain good financial discipline.

#### 4.3.3 Limitations

When it comes to providing explanations in a more human-friendly symbolic manner, the proposed technique faces significant constraints mainly due to its probabilistic nature. The proposed method just tries to mimic and cannot contribute towards interpretability.

**Fig. 12** Calculation of posterior probability of topic 0

Here,
W= [donald, moran, politics, trump] (from Table 8)
N=4
No. of Topics = 5
On product expansion of posterior probability

$$P(T_0|W) = [P(W_1|T_0) \times P(W_2|T_0) \times P(W_3|T_0) \times P(W_4|T_0)] \times P(T_0)$$

From Table 7 (Class Conditional Probability Table)

$$P(W_1|T_0) = P(donald|T_0) = 0.036$$
$$P(W_2|T_0) = P(moran|T_0) = 0.018$$
$$P(W_3|T_0) = P(politics|T_0) = 0.036$$
$$P(W_4|T_0) = P(trump|T_0) = 0.036$$

and

$$P(T_0) = \frac{1}{TotalNo.ofTopics} = \frac{1}{5} = 0.2$$

On substituting values,

$$P(T_0|W) = [0.036 \times 0.018 \times 0.036 \times 0.036] \times 0.2 = 1.7486e^{-07}$$

Similarly,

$$P(T_1|W) = 4.3712e^{-08}$$
$$P(T_2|W) = 3.4970e^{-07}$$
$$P(T_3|W) = 4.7128e^{-08}$$
$$P(T_4|W) = 8.7425e^{-08}$$

Normalize $P(T_0|W)$ as given below.

$$P(T_0|W) = \frac{P(T_0|W)}{P(T_0|W)+P(T_1|W)+P(T_2|W)+P(T_3|W)+P(T_4|W)} = 0.25$$

**Table 10** Probability distribution of topics over test document from Tag My News dataset using graph neural topic model

| Document | Topic 0 | Topic 1 | Topic 2 | Topic 3 | Topic 4 |
|---|---|---|---|---|---|
| Test Document | 0.2608 | 0.1304 | 0.3043 | 0.0869 | 0.2173 |

## 5 Conclusions and Future Works

Graph neural topic model is an effective neural topic modelling technique capable of capturing topics' hierarchical structure and modeling the complex relationships between words in a document. However, the complex black-box nature of the GNTM makes it less applicable to some applications. This paper presented a local explainability technique for explaining the documents topics probability distributions output of the GNTM. The explanation was provided by learning a local interpretable model such as a Naive Bayes classifier. Extensive experimental analyses using three datasets clearly indicated the close resemblance between the individual predictions obtained using the GNTM and the local explainable model. The performance of the proposed technique was also compared with a few baseline topic modelling techniques. The paper also presented a demonstration of how explanations are formed.

In future endeavours, the attainment of explainability can be tried using other explainable models such as random forests, decision trees, etc.

**Acknowledgements** The authors would like to extend their gratitude to the reviewers and editor for their insightful remarks that helped improve the overall quality of this work.

**Author Contributions** All authors contributed to the study, conception, and design. Material preparation, data collection, and analysis were performed by BR, CGV, JS, and SA. The first draft of the manuscript was written by JS and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

**Funding** No funding was received for conducting this study.

**Availability of Data and Materials** All data generated or analyzed during this study are included in this published article.

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest.

**Consent for Publication** Not applicable.

**Ethics Approval and Consent to Participate** Not applicable.



## References

1. Abdelrazek A, Eid Y, Gawish E, Medhat W, Hassan A. Topic modeling algorithms and applications: a survey. Inform Syst. 2023;112: 102131. https://doi.org/10.1016/j.is.2022.102131.
2. Churchill R, Singh L. The evolution of topic modeling. ACM Comput Surv. 2022;54(10):1–35. https://doi.org/10.1145/3507900.
3. Rüdiger M, Antons D, Joshi AM, Torsten-Oliver S. Topic modeling revisited: new evidence on algorithm performance and quality metrics. PLoS ONE. 2022. https://doi.org/10.1371/journal.pone.0266325.
4. Kherwa P, Bansal P. Topic modeling: a comprehensive review. EAI Endors Trans Scalable Inf Syst. 2019;7(24):16. https://doi.org/10.4108/eai.13-7-2018.159623.
5. Anoop VS, Deepak P, Asharaf S. A distributional semantics-based information retrieval framework for online social networks. Intell Decis Technol. 2021;15(2):189–99. https://doi.org/10.3233/IDT-200001.
6. Qi J, Ohsawa Y. Matrix-like visualization based on topic modeling for discovering connections between disjoint disciplines. Intell Decis Technol. 2016;10(3):273–83. https://doi.org/10.3233/IDT-150252.
7. Asmussen CB, Møller C. Smart literature review: a practical topic modelling approach to exploratory literature review. J Big Data. 2019. https://doi.org/10.1186/s40537-019-0255-7.
8. Silva CC, Galster M, Gilson F. Topic modeling in software engineering research. Empir Softw Eng. 2021. https://doi.org/10.1007/s10664-021-10026-0.
9. Egger R, Yu J. A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify twitter posts. Front Sociol. 2022. https://doi.org/10.3389/fsoc.2022.886498.
10. Hagerer G, Leung WS, Liu Q, Danner H, Groh G. A case study and qualitative analysis of simple cross-lingual opinion mining. In: Proceedings of the 13th international joint conference on knowledge discovery, knowledge engineering and knowledge management—KDIR. 2021; pp. 17–26. SciTePress, Portugal. https://doi.org/10.5220/0010649500003064. INSTICC
11. Liu W, Pang J, Li N, Zhou X, Yue F. Research on multi-label text classification method based on tALBERT-CNN. Int J Comput Intell Syst. 2021. https://doi.org/10.1007/s44196-021-00055-4.
12. Chauhan U, Shah A. Topic modeling using latent dirichlet allocation: A survey. ACM Comput Surv. 2022;54(7):1–35. https://doi.org/10.1145/3462478.
13. Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, Zhao L. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimed Tools Appl. 2019;78:15169–211. https://doi.org/10.1007/s11042-018-6894-4.
14. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res. 2003;3:993–1022.
15. Shakeel K, Tahir GR, Tehseen I, Ali M. A framework of URDU topic modeling using Latent Dirichlet Allocation (LDA). In: 2018 IEEE 8th annual computing and communication workshop and conference (CCWC), Las Vegas, NV, USA; 2018. https://doi.org/10.1109/CCWC.2018.8301655.
16. van der Velden BHM, Kuijf HJ, Gilhuijs KGA, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med Image Anal. 2022;79:102470. https://doi.org/10.1016/j.media.2022.102470.
17. Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang G-Z. XAI-explainable artificial intelligence. Sci Robot. 2019. https://doi.org/10.1126/scirobotics.aay7120.
18. Samek W, Wiegand T, Müller KR. Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. ITU J ICT Discover. https://doi.org/10.48550/arXiv.1708.08296.
19. Angelov PP, Soares EA, Jiang R, Arnold NI, Atkinson PM. Explainable artificial intelligence: an analytical review. WIREs Data Min Knowl Disc. 2021;11(5):1424. https://doi.org/10.1002/widm.1424.
20. Samek W, Montavon G, Vedaldi A, Hansen LK, Müller KR. Explainable AI: interpreting, explaining and visualizing deep learning, vol. 11700. Lecture Notes in Artificial Intelligence. Switzerland: Springer; 2019.
21. Saeed W, Omlin C. Explainable AI (XAI): a systematic meta-survey of current challenges and future opportunities. Knowl-Based Syst. 2023;263: 110273. https://doi.org/10.1016/j.knosys.2023.110273.
22. Shen D, Qin C, Wang C, Dong Z, Zhu H, Xiong H. Topic modeling revisited: a document graph-based neural network perspective. In: Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Vaughan JW (eds.) Advances in neural information processing systems, vol 34. Curran Associates, Inc., Virtual Mode; 2021. p. 14681–93. https://openreview.net/pdf?id=yewqeLly5D8.
23. Zhu B, Cai Y, Ren H. Graph neural topic model with commonsense knowledge. Inf Process Manag. 2023;60(2): 103215. https://doi.org/10.1016/j.ipm.2022.103215.

24. Murakami R, Chakraborty B. Investigating the efficient use of word embedding with neural-topic models for interpretable topics from short texts. Sensors. 2022. https://doi.org/10.3390/s22030852.
25. Kang X, Xiaoqiu L, Yuan-fang L, Tongtong W, Guilin Q, Ning Y, Dong W, Zheng Z. Neural topic modeling with deep mutual information estimation. Big Data Res. 2022;30: 100344. https://doi.org/10.1016/j.bdr.2022.100344.
26. Garg R, Kiwelekar AW, Netak LD, Bhate SS. In: Gunjan, V.K., Zurada, J.M. (eds.) Personalization of news for a logistics organisation by finding relevancy using NLP. Cham: Springer; 2021. p. 215–226. https://doi.org/10.1007/978-3-030-68291-0_16.
27. Garg R, Kiwelekar AW, Netak LD, Bhate SS. In: Gunjan VK, Zurada JM (eds) Potential use-cases of natural language processing for a logistics organization. Cham: Springer; 2021. p. 157–191. https://doi.org/10.1007/978-3-030-68291-0_13.
28. Sammut C. In: Sammut C, Webb GI (eds) Markov Chain Monte Carlo. Encyclopedia of machine learning Boston: Springer;. 2011. p. 639–42. https://doi.org/10.1007/978-0-387-30164-8_511.
29. Haugh MB. A tutorial on Markov chain Monte-Carlo and Bayesian modeling. Report; 2021. https://doi.org/10.2139/ssrn.3759243. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3759243.
30. Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK. An introduction to variational methods for graphical models. Mach Learn. 1999;37:183–233. https://doi.org/10.1023/A:1007665907178.
31. Kingma DP, Welling M. Auto-encoding variational bayes. In: 2nd international conference on learning representations (ICLR2014). Ithaca, NY. arXiv.org. Rimrock Resort, Canada. 2014; https://arxiv.org/abs/1312.6114.
32. Miao Y, Grefenstette E, Blunsom P. Discovering discrete latent topics with neural variational inference. In: Precup D, Teh YW (eds) Proceedings of the 34th international conference on machine learning. Proceedings of machine learning research, vol. 70. PMLR, Sydney, Australia; 2017. p. 2410–19. https://proceedings.mlr.press/v70/miao17a.html.
33. Zhao H, Phung D, Huynh V, Jin Y, Du L, Buntine W. Topic modelling meets deep neural networks: a survey. In: Proceedings of the thirtieth international joint conference on artificial intelligence (IJCAI-21) survey track; 2021. p. 4713–20. https://doi.org/10.24963/ijcai.2021/638.
34. Sun H, Tu Q, Li J, Yan R. Convntm: conversational neural topic model. Proc AAAI Conf Artif Intell. 2023;37(11):13609–17. https://doi.org/10.1609/aaai.v37i11.26595.
35. Zhao H, Phung D, Huynh V, Jin Y, Du L, Buntine W. Topic modelling meets deep neural networks: a survey. In: Zhou Z-H (ed) Proceedings of the thirtieth international joint conference on artificial intelligence. Association for the Advancement of Artificial Intelligence (AAAI), United States of America; 2021. p. 4713–20. https://doi.org/10.24963/ijcai.2021/638. https://www.ijcai.org/proceedings/2021/. https://ijcai-21.org.
36. Wu Z, Pan S, Chen F, Long G, Zhang C, Yu PS. A comprehensive survey on graph neural networks. IEEE Trans Neural Netw Learn Syst. 2021;32(1):4–24. https://doi.org/10.1109/tnnls.2020.2978386.
37. Zhou D, Hu X, Wang R. Neural topic modeling by incorporating document relationship graph. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). Association for Computational Linguistics, Online; 2020. p. 3790–6. https://doi.org/10.18653/v1/2020.emnlp-main.310.
38. Ying R, Bourgeois D, You J, Zitnik M, Leskovec J. GNNexplainer: generating explanations for graph neural networks; 2019. arXiv:1903.03894. https://doi.org/10.48550/arXiv.1903.03894.
39. Yuan H, Tang J, Hu X, Ji S. XGNN: towards model-level explanations of graph neural networks. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery and data mining. KDD'20. Association for Computing Machinery, New York, NY, USA; 2020. p. 430–38. https://doi.org/10.1145/3394486.3403085.
40. Yuan H, Yu H, Wang J, Li K, Ji S. On explainability of graph neural networks via subgraph explorations. In: Meila M, Zhang T (eds) Proceedings of the 38th international conference on machine learning. Proceedings of machine learning research, vol. 139. p. 12241–52. PMLR, Virtual Mode; 2021. https://proceedings.mlr.press/v139/yuan21c.html.
41. Vu MN, Thai MT. PGM-explainer: probabilistic graphical model explanations for graph neural networks; 2020. arXiv:2010.05788. https://doi.org/10.48550/arXiv.2010.05788.
42. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier; 2016. arXiv:1602.04938. https://doi.org/10.48550/arXiv.1602.04938.
43. Huang Q, Yamada M, Yuan Tian DS, Yin D, Chang Y. GraphLIME: local interpretable model explanations for graph neural networks; 2020. arXiv:2001.06216. https://doi.org/10.48550/arXiv.2001.06216.
44. Yuan H, Yu H, Gui S, Ji S. Explainability in graph neural networks: a taxonomic survey. IEEE Trans Pattern Anal Mach Intell. 2023;45(5):5782–99. https://doi.org/10.1109/TPAMI.2022.3204236.
45. Wu L, Zhao H, Li Z, Huang Z, Liu Q, Chen E. Learning the explainable semantic relations via unified graph topic-disentangled neural networks. ACM Trans Knowl Discov Data. 2023. https://doi.org/10.1145/3589964.
46. Holzinger A, Malle B, Saranti A, Pfeifer B. Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI. Inform Fus. 2021;71:28–37. https://doi.org/10.1016/j.inffus.2021.01.008.
47. Xie Q, Tiwari P, Gupta D, Huang J, Peng M. Neural variational sparse topic model for sparse explainable text representation. Inf Process Manag. 2021. https://doi.org/10.1016/j.ipm.2021.102614.
48. Berrar D. Bayes' theorem and Naive Bayes classifier. In: Ranganathan S, Gribskov M, Nakai K, Schönbach C (eds) Encyclopedia of bioinformatics and computational biology. Academic Press, Oxford; 2019. p. 403–12. https://doi.org/10.1016/B978-0-12-809633-8.20473-1. https://www.sciencedirect.com/science/article/pii/B9780128096338204731.
49. Chang V, Ali MA, Hossain A. Chapter 2-Investigation of Covid-19 and scientific analysis big data analytics with the help of machine learning. In: Chang V, Abdel-Basset M, Ramachandran M, Green NG, Wills G (eds) Novel AI and data science advancements for sustainability in the era of COVID-19. Academic Press, Oxford; 2022. p. 21–66. https://doi.org/10.1016/B978-0-323-90054-6.00007-6. https://www.sciencedirect.com/science/article/pii/B9780323900546000076.
50. Theodoridis S. Chapter 2-Probability and stochastic processes. In: Theodoridis S (ed) Machine learning (Second Edition), Second edition. Academic Press, Oxford; 2020. p. 19–65. https://doi.org/10.1016/B978-0-12-818803-3.00011-8. https://www.sciencedirect.com/science/article/pii/B9780128188033000118.
51. D'Agostino M, Dardanoni V. What's so special about Euclidean distance? Soc Choice Welf. 2009;33:211–33. https://doi.org/10.1007/s00355-008-0353-5.
52. Suwanda R, Syahputra Z, Zamzami EM. Analysis of Euclidean distance and Manhattan distance in the K-means algorithm for variations number of centroid K. J Phys Conf Ser. 2020;1566(1):012058. https://doi.org/10.1088/1742-6596/1566/1/012058.
53. Alangari N, El Bachir MM, Mathkour H, Almosallam I. Exploring evaluation methods for interpretable machine learning: a survey. Information. 2023. https://doi.org/10.3390/info14080469.
54. Craven MW, Shavlik JW. Extracting tree-structured representations of trained networks. In: Proceedings of the 8th international

conference on neural information processing systems. MIT Press, Cambridge, MA, USA; 1995. p. 24–30. https://dl.acm.org/doi/10.5555/2998828.2998832.
55. Blei DM, Andrew MIJ, Ng Y. Latent Dirichlet allocation. J Mach Learn Res. 2003;3:993–1022.
56. Asmussen CB, Møller C. Smart literature review: a practical topic modelling approach to exploratory literature review. J Big Data. 2019;6(1):93. https://doi.org/10.1186/s40537-019-0255-7.
57. Albalawi R, Yeap TH, Benyoucef M. Using topic modeling methods for short-text data: a comparative analysis. Front Artif Intell. 2020. https://doi.org/10.3389/frai.2020.00042.
58. Grootendorst M. Bertopic: neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794 [cs.CL], 10; 2022. https://doi.org/10.48550/arXiv.2203.05794.