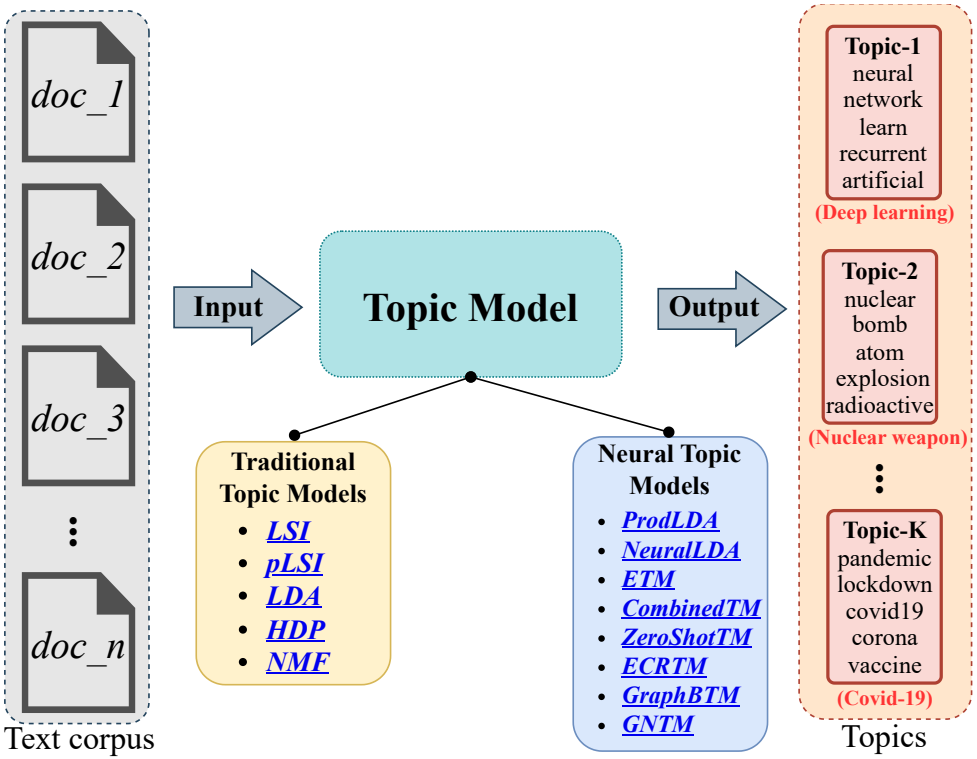
GINopic: Topic Modeling with Graph Isomorphism Network

Suman Adhya, Debarshi Kumar Sanyal

School of Mathematical & Computational Sciences
Indian Association for the Cultivation of Science

NAACL 2024

What is a Topic Model?

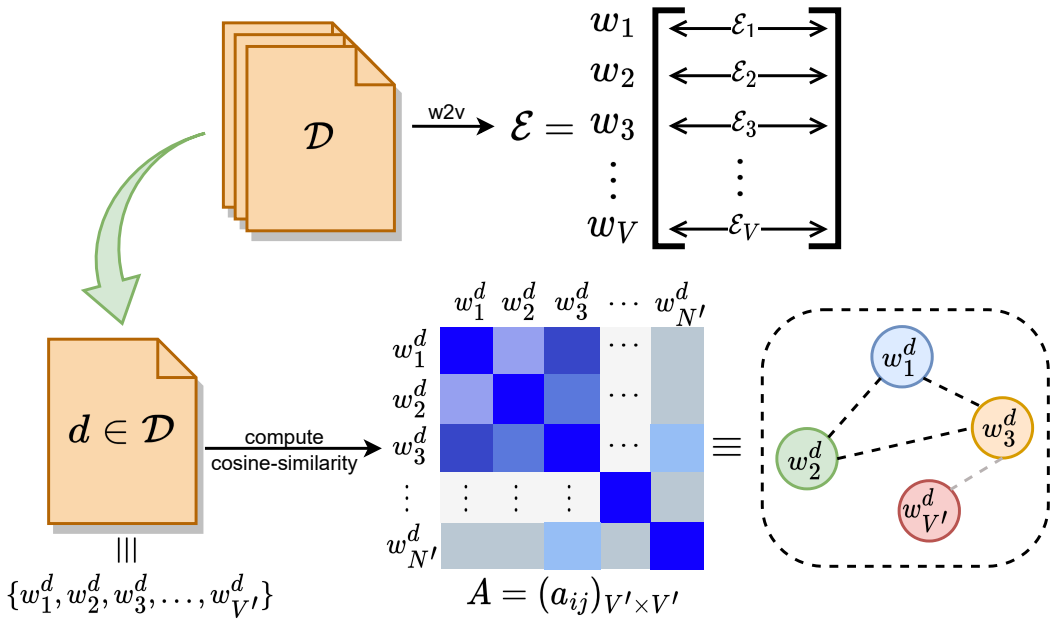


- **Type**: Unsupervised learning;
- **Input**: Set of documents;
  Output: Set of topics;
- **Topic**: Distribution over the words;
- **Use cases**:
  1. Document clustering;
  2. Text classification;
  3. Information retrieval;

Motivation

- How can we incorporate explicit word dependency patterns into topic modeling?
  - The recent neural topic models emphasize the document's contextualized embeddings over word interactions.
- How can we address issues in current graph-based neural topic models?
  - GraphBTM exhibits poor performance.
  - Graph construction methodology for both the GraphBTM and GNTM are computationally expensive.

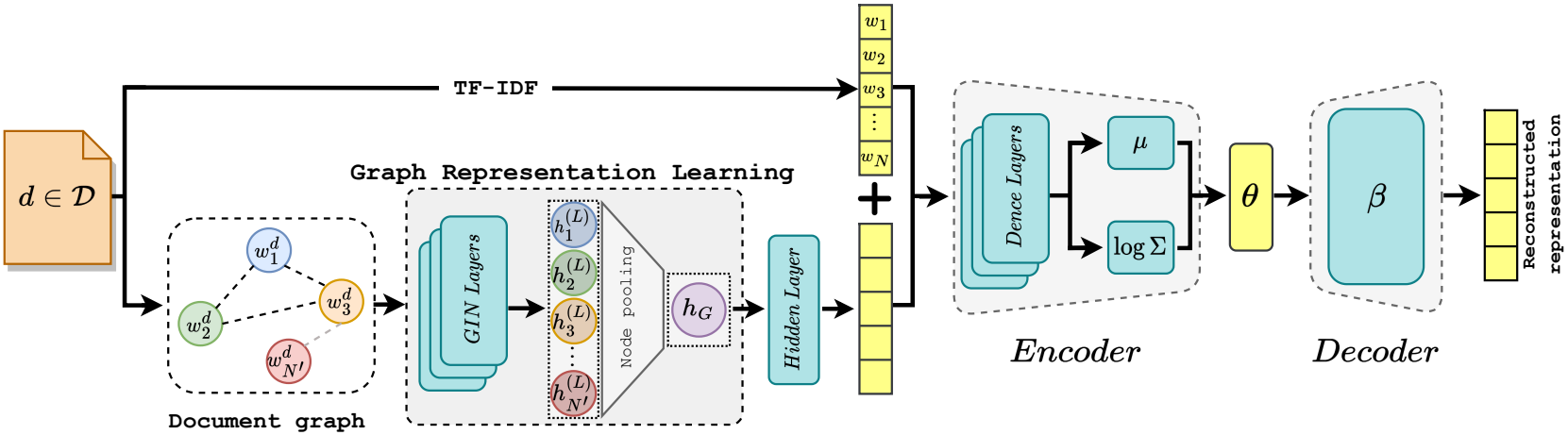
Graph Construction



Graph construction methodology.

$$A = (a_{ij})_{V' \times V'}, \text{ such that: } a_{ij} = \begin{cases} 0 & \text{if } \mathrm{Sim}(\mathcal{E}_i, \mathcal{E}_j) < \delta \\ \mathrm{Sim}(\mathcal{E}_i, \mathcal{E}_j) & \text{otherwise} \end{cases}$$

Proposed Framework



Proposed framework for GINopic model.

- **Graph Representation Learning**;
- **VAE framework**;

Experimental Setup

All experiments are performed using **OCTIS**.

- **Datasets**:

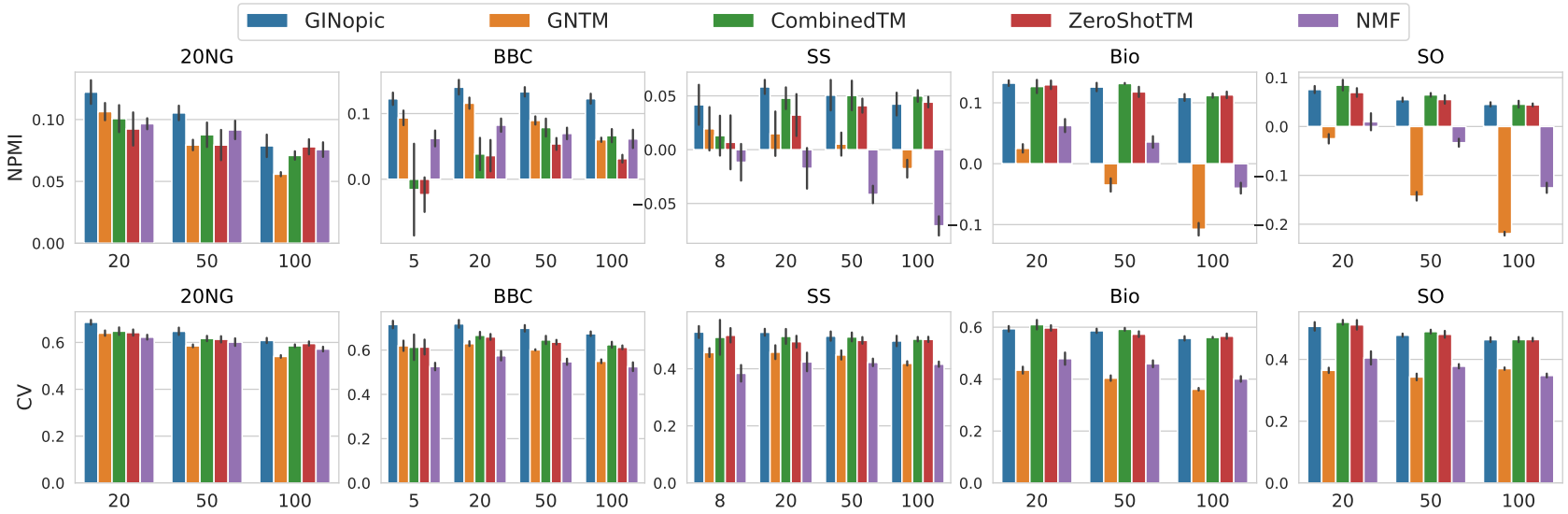
Dataset	#Total Docs	#Tr Docs	#Ts/Va Docs	Avg. Doc. length	Labels (k_{gold})
20NG	16309	11415	2447	48.020	20
BBC	2225	1557	334	120.116	5
SS	12270	8588	1841	13.104	8
Bio	18686	13080	2803	7.022	20
SO	15696	10986	2355	5.106	20

Table 1: Statistics of the used datasets.

- **Baselines**:

1 **ECRTM**; 2 **CombinedTM**; 3 **ZeroShotTM**;
4 **ProdLDA**; 5 **NeuralLDA**; 6 **ETM**;
7 **LDA**; 8 **LSI**; 9 **NMF**;
10 **GraphBTM**; 11 **GNTM**;

Quantitative Analysis



Topic coherence (NPMI and CV) scores for each topic count for top-5 topic models on five datasets. The mean and standard deviation over 5 random runs are shown.

Quantitative Analysis - Diversity Scores

Model	20NG			BBC			SS			Bio			SO		
	IRBO	wl-M	wl-C	IRBO	wl-M	wl-C	IRBO	wl-M	wl-C	IRBO	wl-M	wl-C	IRBO	wl-M	wl-C
ECRTM	**0.998**	**0.473**	0.852	**0.999**	0.454	0.848	**1.000**	0.442	0.839	**1.000**	0.433	0.838	**1.000**	0.382	0.825
GraphBTM	0.971	0.462	0.852	0.986	0.448	0.846	0.947	0.421	0.836	0.924	0.427	0.837	0.958	0.374	0.821
GNTM	0.984	0.461	0.852	0.983	0.444	0.845	0.995	**0.454**	0.846	0.999	0.455	0.845	0.949	0.406	0.831
GINopic	0.989	0.468	**0.895**	0.992	**0.457**	**0.893**	0.998	**0.454**	**0.889**	0.983	**0.462**	**0.888**	0.986	**0.497**	**0.879**

Comparison of topic models on five datasets. For each metric and each topic model, we mention the mean scores over topic counts $\{20, 50, 100\} \cup \{k_{gold}\}$.

Qualitative Analysis

Model	Topics
GraphBTM	**armenian**, **afraid**, neighbor, clock, **soldier**, **turkish**, floor, **soviet**, **beat**, arrive
	game, **score**, car, engine, **play**, **goal**, **season**, **playoff**, **shot**, **player**
	tire, **bike**, connector, ide, **brake**, scsi, cable, **car**, **rear**, **engine**
GNTM	**israeli**, **arab**, **jewish**, policy, **land**, **territory**, **area**, peace, **human**, **population**
	team, **game**, **play**, **player**, **win**, year, good, call, point, time
	tire, **oil**, **brake**, **bike**, **paint**, weight, corner, air, **lock**, **motorcycle**
GINopic	**genocide**, **muslim**, **armenian**, **massacre**, **turkish**, **population**, **kill**, **government**, **troop**, **war**
	team, **win**, **score**, **baseball**, **game**, **player**, **hockey**, **playoff**, **goal**, **play**
	car, **bike**, **ride**, **brake**, **light**, **tire**, **engine**, **lock**, side, **mile**

Three topics "Armenian genocide", "Sports", and "Automobile" presented from the 20NG dataset.

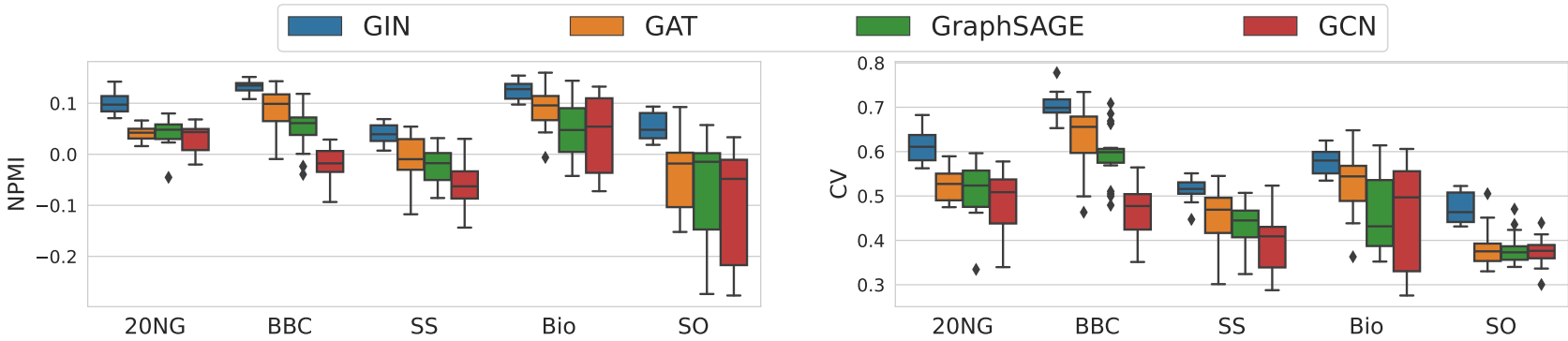
Extrinsic Evaluation – Document Classification

Model	20NG	BBC	SS	Bio	SO
ECRTM	0.411	0.816	0.492	0.361	0.457
GraphBTM	0.052	0.231	0.224	0.060	0.050
GNTM	**0.449**	0.806	0.222	0.049	0.053
GINopic	0.441	**0.888**	**0.713**	**0.566**	**0.785**

Average accuracy scores in the document classification task for all the models trained with topic count k_{gold} for all five datasets.

GINopic has the best accuracy on most datasets, except for the 20NG dataset where it comes in second place in terms of accuracy.

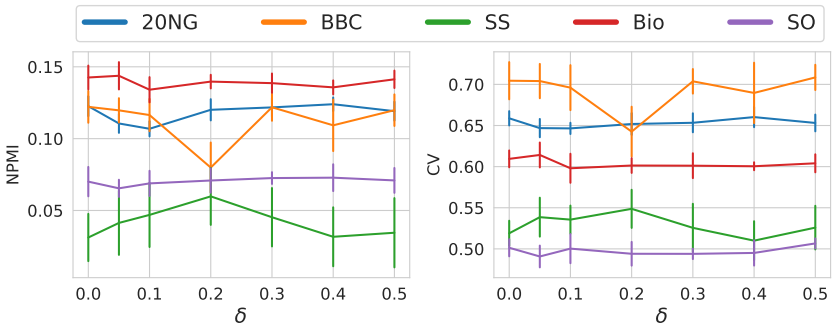
Sensitivity Analysis – Choice of GNN



Box plot of topic coherence (NPMI and CV) scores incorporating GIN, GAT, GraphSAGE, and GCN in GINopic on five datasets.

The results highlight the effectiveness of GIN over other graph neural networks in our topic model.

Sensitivity Analysis – Choice of δ



Coherence (NPMI and CV) scores for each dataset by varying the threshold (δ) value in $\{0.0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$.

Datasets	Optimal threshold (δ)	Train Time Reduction (%)
20NG	0.4	154.27%
BBC	0.3	266.72%
SS	0.2	16.71%
Bio	0.05	0.29%
SO	0.1	1.06%

Optimal threshold (δ) value along with the percentage of training time reduction for all five datasets.

By $\uparrow \delta$, the document graphs become sparser and consequently training time $\downarrow$.

Conclusion

Introducing GINopic, a neural topic model using graph isomorphism networks. It outperforms existing models across datasets. Sensitivity analysis shows how graph thresholds affect performance and training time. GIN proves superior to other GNNs in topic modeling.

Future Work

- Alternative methods for constructing document graphs.
- Explore capturing diverse word dependencies and integrating them to construct a multifaceted document graph.

Thank you all for listening...

Code: `https://github.com/AdhyaSuman/GINopic`