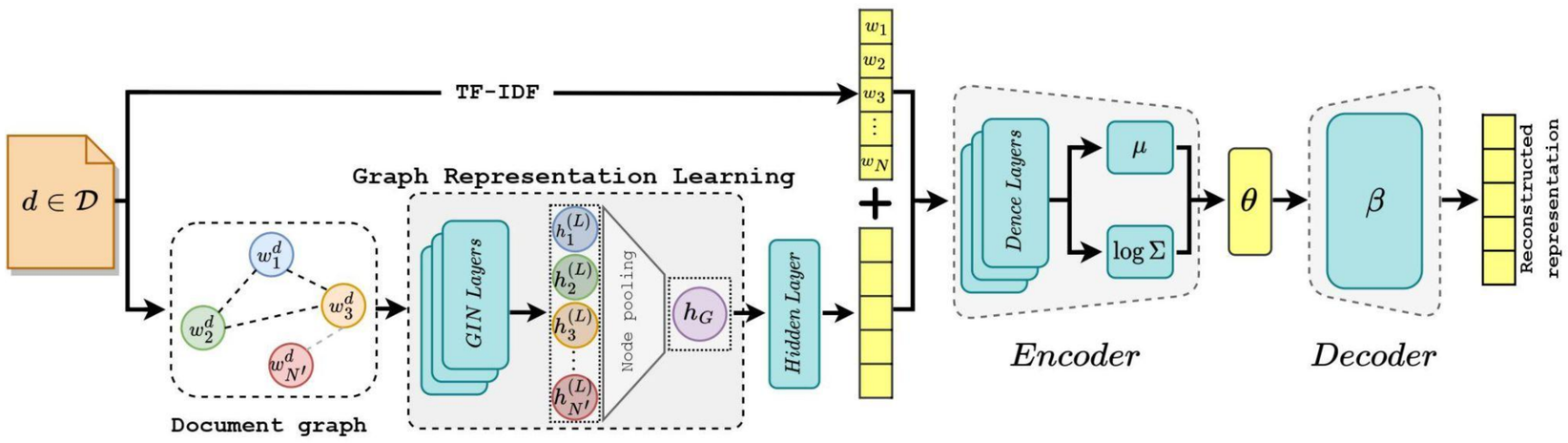
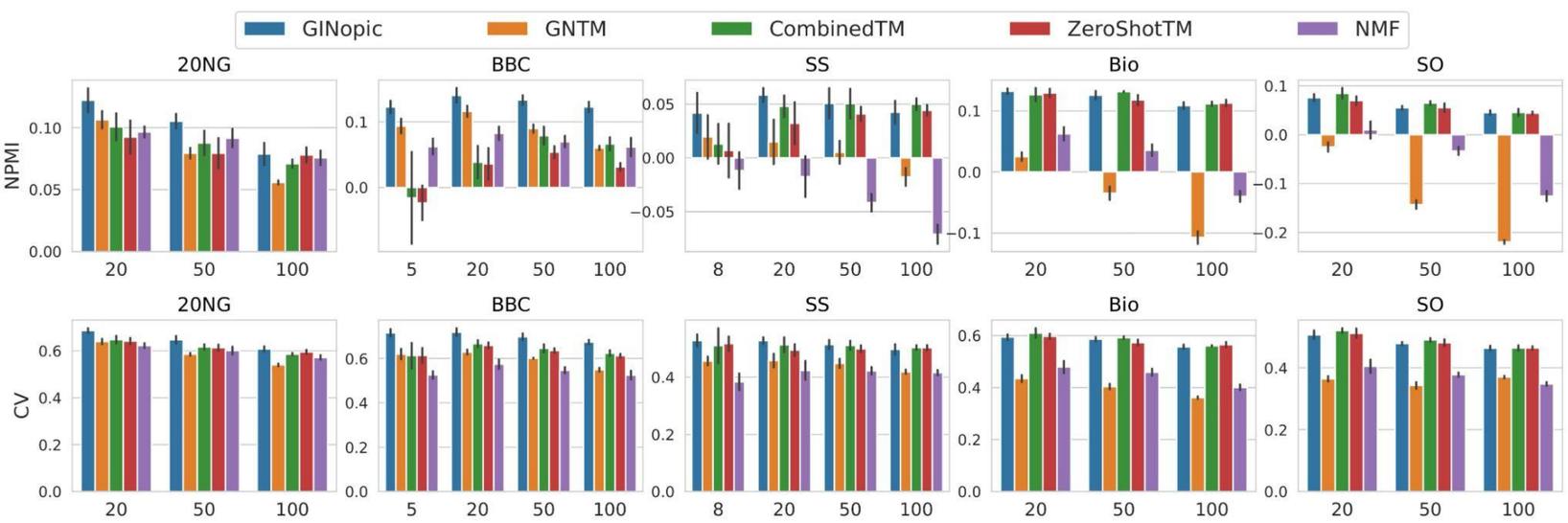
GINopic: Topic Modeling with Graph Isomorphism Network

Proposed Framework



GINopic model

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.

Qualitative Evaluation

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

Model	20NG	BBC	SS	Bio	SO
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 with topic count k_{gold} for all five datasets.

Authors:

1. Suman Adhya
2. Debarshi Kumar Sanyal

GitHub

Acknowledgments:

MSR India

ACM India-IARCS

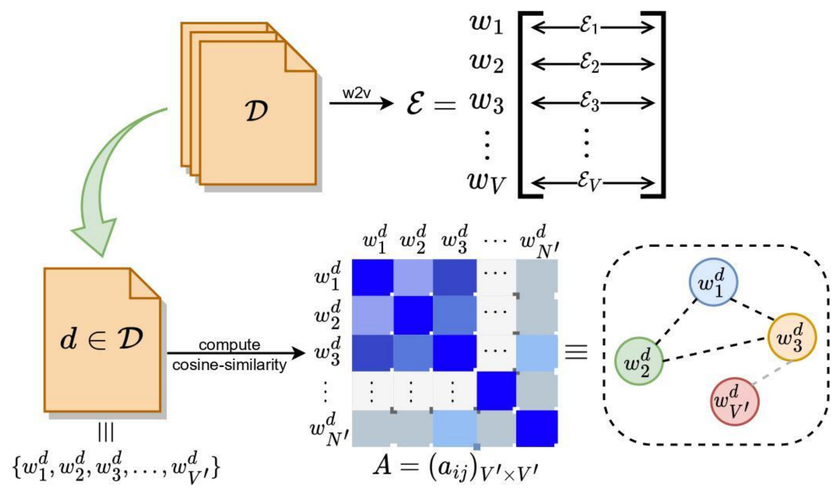
Indian Association for the Cultivation of Science
Jadavpur, Kolkata-700032, India

NAACL 2024

Dataset Overview

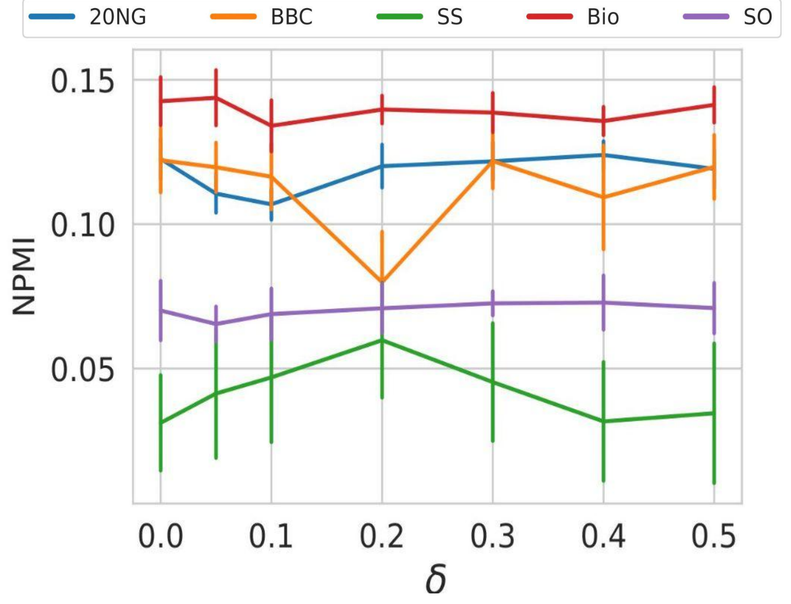
Dataset	#Docs	Avg. Doc. Length	Labels
20NG	16309	48.02	20
BBC	2225	120.12	5
SS	12270	13.10	8
Bio	18686	7.02	20
SO	15696	5.11	20

Graph Construction



$$a_{ij} = \begin{cases} 0 & \text{if } \mathrm{Sim}(\mathcal{E}_i, \mathcal{E}_j) < \delta \\ \mathrm{Sim}(\mathcal{E}_i, \mathcal{E}_j) & \text{otherwise} \end{cases}$$

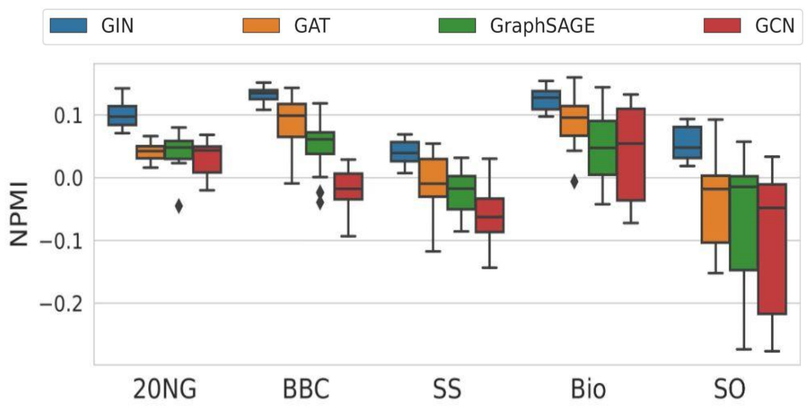
Choice of δ



Dataset	Optimal δ	Train Time ⇩ (%)
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 %

By ⇑δ the document graphs become sparser and consequently training time ⇩

Choice of GNN



GIN-incorporated model consistently outperforms other GNN-based models