Continuous Relational Diffusion driven Topic Model with Multi-grained Text for Microblog

Chenhao Wu[1,3], Ruifang He[2,3,1*], Chang Liu[2,3] and Bo Wang[2,3]
[1]School of New Media and Communication, Tianjin University, Tianjin, China
[2]College of Intelligence and Computing, Tianjin University, Tianjin, China
[3]Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin, China
{ abby_wu1999,rfhe,changliu,bowang}@tju.edu.cn

Abstract

Topic model is a statistical model that leverages unsupervised learning to mine hidden topics in document collections. The data sparsity and colloquialism of social texts make it difficult to accurately mine the topics. Traditional methods assume that there are only 0/1-state relationships between the two parties in the social networks, but the relationship status in real life is more complicated, such as continuously changing relationships with different degrees of intimacy. This paper proposes a continuous relational diffusion driven topic model (CRTM) with multi-grained text for Microblog to realize the continuous representation of the relationship state and make up for the context and structural information lost by previous representation methods. Multi-grained text representation learning distinguishes the impact of formal and informal expression on the topics further and alleviates colloquialism problems. Specifically, based on the original social network, the reconstructed social network with continuous relationship status is obtained by using information diffusion technology. The graph convolution model is utilized to learn node embeddings through the new social network. Finally, the neural variational inference is applied to generate topics according to continuous relationships. We validate CRTM on three real datasets, and the experimental results show the effectiveness of the scheme.

Keywords: Topic Model, Microblog, Social Media, Graph Diffusion, Multi-grained Text

1. Introduction

In recent years, social media platforms, such as Twitter and Sina Weibo, flourish and generate large-scale posts every day. These posts contain rich semantic information. Automatically detecting topics in social media can reveal hidden semantic information that can be applied in downstream tasks such as short text classification (Inoue et al., 2021), extract text summarization (Joshi et al., 2023), machine translation (Maier et al., 2022) and so on.

The conventional topic models, such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003), infer topics based on word co-occurrence. Numerous studies prove the effectiveness of LDA on long documents. However, they can not be directly applied in Microblog due to the severe data sparsity and colloquialism of Microblog posts (Abdelrazek et al., 2023). The existing researches on social media topic modeling can be mainly categorized into (1) Models aggregate short text with different heuristic aggregation strategies or directly generate biterms (Mehrotra et al., 2013; Alvarez-Melis and Saveski, 2016; Yan et al., 2013). (2) Models use representation learning to produce semantic information (Hu and Tsujii, 2016; Shi et al., 2018; Li et al., 2016; Wu et al., 2020). (3) Models take into account the content and social network structures to deduce topics (Guo et al., 2015; Li et al., 2018; He et al., 2018; Liu et al., 2020; Wang et al., 2022). However, their social relationship is modeled as two states “yes" and “no".

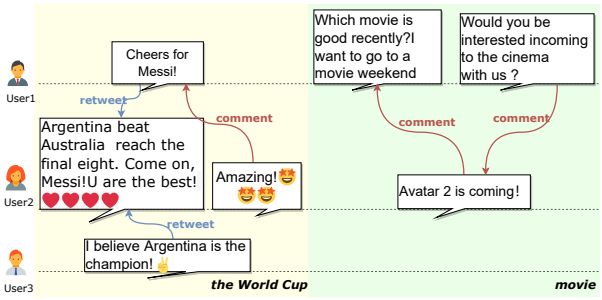


Figure 1: Relationships include intimacy between users. Different background colors represent different topics. Different colored arrows represent different types of interaction. The dashed line is the timeline of the post.

The relationships that include different intimacy in real life are far more complex than in artificially constructed social networks. Features that reflect intimacy such as interaction type, number of interactions, and topic of interaction are ignored in previous work as depicted in Fig. 1. It is observed that there is a connection between **[U1]** and **[U2]**, and the same connection between **[U2]** and **[U3]**. No other information is contained. However, considering features mentioned above, **[U1]** has a closer relationship with **[U2]** than **[U3]**. It can even be inferred that **[U1]** and **[U2]** are also friends in real life, while **[U2]** and **[U3]** may be just the netizens who discussed “the World Cup" together.

We can conclude from the above example, that

LREC-COLING 2024, pages 3897–3906

there are different intimacy in relationships. Our work tries to model continuous relationships, indicating the intimacy between users. We design information diffusion with intimacy to extend the discrete message-passing process in Graph Convolutional Networks (GCNs) (Kipf and Welling, 2017) to a continuous diffusion process. It enables nodes to aggregate information closely related to them through determination and dissemination of attention. What's more, we relieve colloquialism problems at the word level and determine the contributions of users to the topic at user level in multi-grained text representation (Liu et al., 2022). The multi-grained text representation improves data interoperability through two-level attention. Especially the word level distinguishes the influence of formal from informal tokens on text representation. Self-attention mechanisms are adopted at the word level and user level respectively. The purpose of word-level attention is to alleviate the impact of colloquial words. The user-level attention is used to determine the contributions of users to the topic (Liu et al., 2022).

To this end, we propose a **C**ontinuous **R**elational Diffusion driven **T**opic **M**odel (CRTM) with multi-grained text. The relationship Continuity (RC) module is designed to reconstruct relationships and multi-grained is used to learn hierarchical text. Two-layer GCNs module is utilized to integrate new structure embedding and hierarchical text embedding. We introduce Variational AutoEncoder (VAE) (Kingma and Welling, 2013) to infer topics. Contributions can be summarized as follows:

- We propose a CRTM for Microblog, which applies information diffusion with intimacy to continue the relationships. To the best of our knowledge, this is the first to consider the continuous relationship between users.
- Multi-grained text representation learning by two-level attention enhances the interpretability of the data and mitigates the impact of colloquialism.
- Comprehensive experiments show the effectiveness of our proposed model. The visualization of social network demonstrates a supplement of intimacy.

2. Related Work

Previous researches and relevant work for topic detection can be mainly classified into the following aspects.

2.1. Focusing on Content Information

These methods depend on pure content to generate document-topic distribution and topic-word distribution. **Aggregation Strategy Based.** To alleviate data sparsity, some methods aggregate posts based on heuristic strategies, such as hashtags (Mehrotra et al., 2013), dialogue relationship (Alvarez-Melis and Saveski, 2016) and then apply the traditional topic model on the long pseudo documents. **Biterm Based.** BTM-like methods (Yan et al., 2013; Chen et al., 2015; Lu et al., 2017) model the generation of biterms in the corpus to alleviate data sparsity. **Embedding Based.** LCTM (Hu and Tsujii, 2016) introduces latent concepts, which are represented by word embeddings to capture the conceptual similarity of words. DWGTM (Wang et al., 2021) learns word features from global word co-occurrence. TSCTM (Wu et al., 2022) applies contrastive learning to distinguish topics in semantic space. However, only content is insufficient.

2.2. Incorporating Content and Social Contexts

This kind of research considers the post content and social contexts together. **Static social characteristic.** AdjEnc (Zhang and Lauw, 2020) incorporates the network structure implicitly. LeadLDA (Li et al., 2018) differentiates messages from leader and follower. Followers' contribution to the topic information is minimal. Social relationships based on static attributes cannot dynamically change over time and are easily influenced by zombie fans. **Dynamic user interaction.** On the basis of VAE, IATM (He et al., 2018) mines the dynamic user behaviors by integrating network embedding and user attention. PCFTM (Liu et al., 2020) seamlessly fuses the parallel social contexts in nonlinear correlation. DGTM (Wang et al., 2022) considers both wide dispersion and deep propagation spread characteristics in social media. However, they ignore intimacy.

2.3. Graph Diffusion Convolution and Neural Variational Inference

Network Representation Learning (NRL) represents nodes in social networks in a low-dimensional vector space. The embeddings obtained can be employed for various graph-based tasks, such as node classification and link prediction. GCNs (Kipf and Welling, 2017) is widely employed in network representation learning due to its superior ability to blend structure and attribute information. Diffusion-convolutional neural networks (Atwood and Towsley, 2016) build a latent representation by scanning a diffusion process across each node in a graph-structured input. **Graph diffusion convolution** (Gasteiger et al., 2019) improves GCNs by spreading attention among neighbors. It removes the restriction of using direct neighbors and alleviates the problem of edges arbitrarily defined in

the social network. Inspired by this idea, we propose information diffusion with intimacy to achieve topic propagation. (Feng et al., 2022) point out that thanks to the flexibility of the neural networks, VAE has the ability to learn complex nonlinear distribution. It replaces the arduous inference work of probabilistic models through stochastic back propagation. The above attract us to learn context-enhanced representations by GCN and induce more coherent topics by VAE.

3. Model

Text on social media is characterized by data sparsity and colloquialism, which poses challenges for topic modeling in social media scenarios. Existing methods for aggregating social context do not take into account the difference in relationships among users, but simply consider it as a yes or no situation. Our method breaks this limitation and achieves continuous representation involving the closeness of the relationship. In addition, word-level and user-level attention mechanisms are adopted to learn text representation for noise reduction. The proposed CRTM framework is shown in Fig. 2. It mainly includes (1) Relationship continuity module, (2) GCNs module, and (3) VAE-based topic generation module.

3.1. User-level Social Network

We build a user-level social network $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{T})$. $\mathcal{V} = \{v_i | 1 \leq i \leq L\}$ is the set of nodes. v_i represents users i in social network. L is the size of the node set. $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is the edge set. $e(i, j) \in \mathcal{E}$ stands for the forwarding and comment between v_i and v_j. $\mathcal{T} = \{tx_1, tx_2, ..., tx_L\}$ is the set of posts. Each tx_i contains M words. According to the posts set $\mathcal{T}$, we obtain the attribute matrix $\boldsymbol{X} \in \mathbb{R}^{L \times C}$. C is the dimension of the user embedding. We replace each word with its corresponding low dimensional, continuous, and real-valued embedding (Mikolov et al., 2013) and weighted average.

The adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{L \times L}$ contains $e(i, j)$. Continuous relation adjacency matrix $\boldsymbol{A}_d$ obtained by information diffusion. The reconstructed relationship is stored in $\boldsymbol{A}_r$, which is sparsized.

3.2. Model Description

RC module is employed to learn a new adjacency matrix. GCNs module represents nodes with continuous relationships and hierarchical text. Then VAE based topic inference module infers topics.

3.2.1. Limitations of MPNN

GCN can be regarded as Message Passing Neural Networks (MPNN) frameworks. Graph neural networks iteratively aggregate neighborhood information through message passing to update the feature of their own node. Message passing can be decomposed into two steps of message aggregation and updating. While GCN does take advantage of higher-order neighborhoods in deeper layers, it seems arbitrary and impractical to restrict messages at the same neighbor size. Because edges in real graphs are often noisy and discrepant. This general message-passing approach simply models the presence or absence of node relationships and considers all nodes to be equally important. Therefore, we need to reflect the relationship difference to achieve the effect of noise reduction and complement intimacy information.

3.2.2. Relationship Continuity

We assume that topic is passed to the neighbors along the starting node and gradually dispersed as Eq. 1. Diffusion after several iterations, the information distribution at this time represents the intimacy from the starting node to other nodes. By doing this for each node, we get a new adjacency matrix that contains continuous relationships.

$$\boldsymbol{A}_d = \sum_{k=1}^{\infty} \theta^k \boldsymbol{S}^k \tag{1}$$

$$\boldsymbol{S} = \tilde{\boldsymbol{D}}^{-1/2} \tilde{\boldsymbol{A}} \tilde{\boldsymbol{D}}^{-1/2} \tag{2}$$

$$\tilde{\boldsymbol{D}} = \boldsymbol{I}_N + \boldsymbol{D} \qquad \tilde{\boldsymbol{A}} = \boldsymbol{I}_N + \boldsymbol{A} \tag{3}$$

$\boldsymbol{S}$ is a symmetric transition matrix and simulates random walk. $\boldsymbol{D}$ is the diagonal matrix, i.e. $\boldsymbol{D}_{ii} = \sum_{j=1}^{L} \boldsymbol{A}_{ij}$. $\tilde{\boldsymbol{D}}$ and $\tilde{\boldsymbol{A}}$ are diagonal matrix and adjacency matrix added self-loop respectively. They symbolize nodes' delay and stagnation in information dissemination. We learn the continuous edge representation instead of the vanilla binary form. In order to make Eq.1 converge, the weighting coefficient θ_k should satisfy $\sum_{k=0}^{\infty} \theta_k = 1$ and $\theta_k \in [0, 1]$. The eigenvalues of $\boldsymbol{S}$ are bounded by $\lambda_i \in [0, 1]$ (Gasteiger et al., 2019). Then we filter out the relationship with low closeness.

3.2.3. Fusion of Continuous Relationships and Hierarchical Text

For each v_i, the word embedding matrix $\boldsymbol{W}^i \in \mathbb{R}^{M \times L}$. $\boldsymbol{W}_q$, $\boldsymbol{W}_k$, and $\boldsymbol{W}_v$ are weight matrices. The word-level attention layer eases the impact of colloquial words in a fine-grained manner.

$$\boldsymbol{Q}^{wi} = \boldsymbol{W}_q \cdot \boldsymbol{W}^i \qquad \boldsymbol{K}^{wi} = \boldsymbol{W}_k \cdot \boldsymbol{W}^i \qquad \boldsymbol{V}^{wi} = \boldsymbol{W}_v \cdot \boldsymbol{W}^i \tag{4}$$

$$\boldsymbol{X}^{wi} = softmax(\frac{\boldsymbol{Q}^{wi} \cdot \boldsymbol{K}^{wi^T}}{\sqrt{C}}) \boldsymbol{V}^{wi} \tag{5}$$

Then user-level attention is used to learn the contribution of users to the topic distribution. We refer

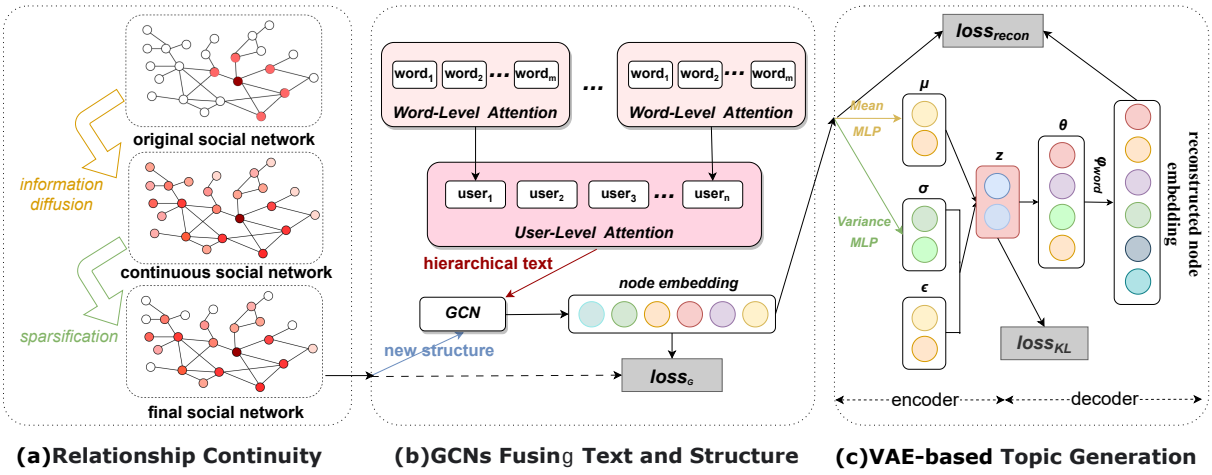


Figure 2: The framework of CRTM.

to $\boldsymbol{X}^s = \{\boldsymbol{X}^{w1}, \boldsymbol{X}^{w2}, ..., \boldsymbol{X}^{wL}\}$ as input and the user-level attention layer is designed as follows:

$$\boldsymbol{Q}^s = \boldsymbol{W}_q \cdot \boldsymbol{X}^s \qquad \boldsymbol{K}^s = \boldsymbol{W}_k \cdot \boldsymbol{X}^s \qquad \boldsymbol{V}^s = \boldsymbol{W}_v \cdot \boldsymbol{X}^s \tag{6}$$

$$\boldsymbol{X}^r = softmax(\frac{\boldsymbol{Q}^s \cdot \boldsymbol{K}^{sT}}{\sqrt{C}})\boldsymbol{V}^s \tag{7}$$

Taking $\boldsymbol{A}_r$ and $\boldsymbol{X}^r$as input. Two-layer GCNs are utilized to learn the node embedding.

$$\widehat{\boldsymbol{A}_r} = \tilde{\boldsymbol{D}}^{-1/2}\boldsymbol{A}_r\tilde{\boldsymbol{D}}^{-1/2} \tag{8}$$

$$\boldsymbol{H}_1 = f(\widehat{\boldsymbol{A}_r}\boldsymbol{X}^r\boldsymbol{W}_1) \tag{9}$$

$$\boldsymbol{H}_2 = f(\widehat{\boldsymbol{A}_r}\boldsymbol{H}_1\boldsymbol{W}_2) \tag{10}$$

$\boldsymbol{W}_1$and $\boldsymbol{W}_2$ are the weight matrices. We adopt $ReLU(\cdot)$ as the activation function $f(\cdot)$. Neighbors at this time are updated after considering the closeness between nodes. Under the effect of GCNs, each node receives the information sent from its new neighbors and aggregates them. $\boldsymbol{H}_2$ is the representation of nodes, which fuses hierarchical text and new structure.

The loss function based on the original first-order neighbor similarity is no longer suitable. So we design a random walk-like optimization function to encourage users with high intimacy to have similar representations:

$$Loss_G = -\sum_{v_i \in \mathcal{V}} \sum_{v_j \in N_i'} logP(v_j|v_i) \tag{11}$$

$$P(v_j|v_i) = \frac{exp(cos(\boldsymbol{h}_j^T, \boldsymbol{h}_i))}{\sum_{v_k \in N_i'} exp(cos(\boldsymbol{h}_k^T, \boldsymbol{h}_i))} \tag{12}$$

$\boldsymbol{h_i}$ is intercepted from $\boldsymbol{H}_2$. v_k is v_i's new neighbor in N_i'. $cos(\cdot,\cdot)$represents the cosine similarity. $Loss_G$ computes the similarity of high-order neighbors.

3.2.4. Topic Inference

We feed $\boldsymbol{h}_i$ into the encoder part of VAE. The mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma}$ of Gaussian distribution are obtained through the encoder.

$$\boldsymbol{e} = ReLU(\boldsymbol{W}_e\boldsymbol{h}_i + \boldsymbol{b}_e) \tag{13}$$

$$\boldsymbol{\mu} = \boldsymbol{W}_\mu\boldsymbol{e} + \boldsymbol{b}_\mu \quad log\boldsymbol{\sigma}^2 = \boldsymbol{W}_\sigma\boldsymbol{e} + \boldsymbol{b}_\sigma \tag{14}$$

where $\boldsymbol{W}_\mu$, $\boldsymbol{W}_\sigma$, $\boldsymbol{b}_\mu$ and $\boldsymbol{b}_\sigma$ are the parameters. The latent topic vector $\boldsymbol{z}$ can be calculated using the reparameterization trick as $\boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{\epsilon} \times \boldsymbol{\sigma}$, where $\boldsymbol{\epsilon} \in N(0,1)$. User-topic distribution $\boldsymbol{\theta} = (p(t_1|\boldsymbol{e}), ..., p(t_K|\boldsymbol{e}))$ is parameterized by $softmax$ function as Eq. 15:

$$\boldsymbol{\theta} = softmax(\boldsymbol{W}_\theta z) \tag{15}$$

Topic-word distribution $\boldsymbol{\phi}_{word} = (p(w|t_1), p(w|t_2), ..., p(w|t_K))$ is the parameter of decoder. The reconstructed user representation $\boldsymbol{h_i'}$ is generated through the full connection layer thereafter:

$$\boldsymbol{d} = softmax(\boldsymbol{\phi}_{word} \times \boldsymbol{\theta}) \tag{16}$$

$$\boldsymbol{h_i'} = ReLU(\boldsymbol{W}_d\boldsymbol{d} + \boldsymbol{b}_d) \tag{17}$$

The loss function of this module is defined as:

$$Loss_V = KL(p(\boldsymbol{z}|\boldsymbol{h_i'})\|q(\boldsymbol{z})) - E_{\boldsymbol{z}\sim p(\boldsymbol{z}|\boldsymbol{h_i'})}q(\boldsymbol{h_i'}|\boldsymbol{z}) \tag{18}$$

3.3. Model Training

To jointly train the GCNs module and the VAE-based topic inference module, we design a joint loss as Eq. 19. Where λ_l is a coefficient to balance the losses.

$$L = Loss_G + \lambda_l * Loss_V \tag{19}$$

Table 1: Statistics of datasets.#Users and #Interactions represent the number of users and interactions respectively.

Month	#Users	#Interactions	Vocabulary size
May	8907	10435	5914
June	19293	35962	9368
July	16990	20971	9663

4. Experiments

4.1. Datasets

We utilize the datasets (Li et al., 2018) shown in Table 1 based on the original Microblog corpus. The corpus is made up of the posts on Sina Weibo during May 1 - July 31, 2014, through hashtag-search API [1], which are publicly available and widely employed in social media topic modeling work (Li et al., 2018; He et al., 2018; Wang et al., 2022).

We further deal with the original datasets as follows: (i) remove the posts whose length is less than 3 words or that have no poster username; (ii) remove independent users who have no interactions; (iii) aggregate all the original and the reposting posts from the same user.

4.2. Evaluation Metric

The perplexity (Blei et al., 2003) of topic model as the evaluation metric does not necessarily correspond to semantically coherent topics (Chang et al., 2009). Therefore, we calculate the coherence score (Wang et al., 2022) of all topic models as Eq. 20:

$$C = \frac{1}{K} \cdot \sum_{k=1}^{K} \sum_{i=2}^{N} \sum_{j=1}^{i-1} log \frac{D(w_i^k, w_j^k) + 1}{D(w_j^k)} \quad (20)$$

K is the number of topics and N is the number of top words under each topic. w_i^k represents the i-th word in topic k ranked by topic-word distribution ϕ_{word}. $D(w_i^k, w_j^k)$ stands for the count of posts in which word w_i^k and word w_j^k co-occur, and $D(w_j^k)$ refers to the count of posts that contain word w_j^k.

4.3. Comparison Methods

To validate whether our method in Section 3 is useful for topic extraction, we compare the proposed CRTM and its variants with the following state-of-the-art baselines and conduct extensive experiments:

Focusing on Content Information:

- **LCTM**(Hu and Tsujii, 2016): introduce latent concepts to capture the conceptual similarity between words for tackling the data sparsity.
- **NQTM**(Wu et al., 2020): propose topic distribution quantization generating peakier distributions that are more appropriate for modeling short texts and negative sampling decoder.
- **TSCTM**(Wu et al., 2022): adopt contrastive learning based on topic semantics mitigates the sparsity issue.

Integrating Content and Social Context:

- **LeadLDA**(Li et al., 2018): distinguish “Leader Messages" from “Follower Messages" and predict the likelihood that leader and follower messages may contain the key topic words.
- **AdjEnc**(Zhang and Lauw, 2020): incorporate the network structure for topic inference on structured long documents, such as web pages.
- **PCFTM**(Liu et al., 2020): obtain parallel sequence through the random walk, and seamlessly integrate content and structure embedding for better representation.
- **IATM**(He et al., 2018): learn interaction-aware edge embedding by dynamic interactions and user attention to alleviate data sparsity.
- **DGTM**(Wang et al., 2022): consider both wide dispersion and deep propagation spread characteristics of topics in social media.

4.4. Experiment Settings

For all the baselines, the hyperparameters settings refer to their original papers and run Gibbs samplings (in LCTM and LeadLDA) with 1,000 iterations to ensure convergence. Employing pretrained word embeddings (PWE) for topic modeling achieves limited performance (Zhang et al., 2022). This may be because PWE is based on word order that the text clustering task does not require. So we randomly initialize the node embedding and set the node embedding dimension C as 200. The dimension of hidden layers in GCNs and the dimension of the first encoder layer in VAE are set to 200. We empirically set the $\lambda_l = 1.0$ for balancing GCNs and VAE. Adam is utilized to optimize the objective function with the learning rate of 0.01. We choose different numbers of topics K and top words in one topic N to validate the performance of models. K is set to 50 and 100, which means that we will detect 50 or 100 topics for each dataset, and N is set to 10, 15, and 20. Top words sorted by topic-word distribution ϕ_{word}. For the different continue methods and corresponding diffusion parameters in the RC module, we will discuss them in detail in **Section 4.7**. We utilize Bayesian optimization to tune the hyperparameters. In order to ensure

[1] http://open.weibo.com/wiki/2/search/topics

Table 2: The topic coherence scores on the three datasets.

Datasets	Model	K50			K100		
		N=10	N=15	N=20	N=10	N=15	N=20
May	LCTM	-70.91	-165.37	-296.36	-58.65	-140.10	-261.40
	NQTM	-93.04	-214.33	-384.82	-90.98	-207.64	-376.32
	TSCTM	-75.21	-176.06	-323.98	-76.50	-180.15	-327.31
	LeadLDA	-53.91	-138.53	-258.38	-58.15	-141.34	-261.65
	AdjEnc	-67.57	-159.66	-290.10	-72.02	-165.87	-303.37
	DGTM	-74.67	-175.37	-324.28	-79.63	-190.06	-327.65
	IATM	-43.34	-112.64	**-228.27**	-47.32	-121.46	**-219.96**
	CRTM	**-39.45**	**-108.65**	-244.69	**-46.61**	**-117.46**	-249.37
	CRTM (- information diffusion)	-60.05	-152.50	-298.18	-75.23	-174.32	-306.65
	CRTM (- multi-grained text)	-51.27	-132.60	-276.12	-59.60	-145.46	-292.93
June	LCTM	-91.72	-208.75	-367.76	-81.88	-181.57	-323.16
	NQTM	-102.93	-239.49	-431.17	-102.23	-239.41	-432.95
	TSCTM	-74.04	-179.93	-334.06	-80.09	-187.52	-342.38
	LeadLDA	-63.54	-150.18	-278.19	-72.07	-169.80	-309.40
	AdjEnc	-72.38	-174.55	-375.29	-77.63	-192.85	-321.86
	DGTM	-81.25	-194.17	-353.55	-74.15	-186.40	-339.76
	IATM	-46.69	-113.09	**-213.61**	-59.11	-133.96	**-225.48**
	CRTM	**-43.84**	**-111.75**	-239.40	**-48.12**	**-129.20**	-268.14
	CRTM (- information diffusion)	-63.83	-150.85	-351.16	-74.43	-218.12	-336.05
	CRTM (- multi-grained text)	-57.28	-144.94	-296.12	-61.20	-162.55	-311.74
July	LCTM	-72.78	-160.08	-275.58	-63.56	-137.36	-238.31
	NQTM	-78.35	-185.85	-339.82	-73.31	-178.51	-331.89
	TSCTM	-57.50	-138.01	-258.90	-61.07	-140.32	-261.43
	LeadLDA	-70.40	-157.83	-268.23	-59.75	-130.83	-226.62
	AdjEnc	-51.72	-123.78	-225.29	-55.73	-140.63	-250.75
	DGTM	-75.94	157.86	-248.81	-60.33	-153.80	-297.34
	IATM	-50.75	-119.48	-212.26	-46.80	-110.27	-204.35
	CRTM	**-46.20**	**-117.82**	**-209.92**	**-46.57**	**-101.83**	**-191.73**
	CRTM (- information diffusion)	-69.44	-158.65	-268.12	-70.20	-144.49	-263.90
	CRTM (- multi-grained text)	-57.10	-141.69	-259.93	-61.45	-135.23	-246.65

stability, we conduct 30 groups of experiments under each dataset and combination of diffusion and sparsification approaches and take the average result.

4.5. Performance Evaluation

Table 2 shows the topic coherence scores of all baselines and our model on the three datasets. From the results, we have the following observations:

- Topic models incorporating content and social context perform better than focusing on content in most cases. It indicates that the former is necessary due to reciprocal influence. Furthermore, social context alleviates data sparsity to some extent.
- The topic coherence score of $K50$ is higher than that of $K100$ in most cases. This is probably due to the original corpus composed of 50 hashtags. 100 topics are too detailed. Topic coherence score decreases as N increases when K fixed. As N increases, more irrelevant information tends to appear in the generated topic keywords.
- Our model outperforms most of baselines except for May and June where $N20$. The reasons are three-fold: RC module processes the original network structure by distinguishing the importance of nodes. It guides the path selection of topic propagation. Multi-grained text representation relieves the impact of colloquial words and determines the contributions of users to the topic. When the N grows, some noise words are added and CRTM is sensitive to noise words. Eq. 20 of coherence score can also reflect this phenomenon.
- TSCTM outperforms several methods combining structural information on July. We believe this may be due to insufficient interaction behavior in the July dataset, and TSCTM's use of contrastive learning has widened the distance between words with different semantics, resulting in an excellent performance in methods considering only text. The performances of DGTM are not as high as expected. This is because the user embeddings in both methods are based on random walks, which are affected by network size and social network topology, and have a certain degree of randomness.

Table 3: The improvement percentage of two variants in May dataset.

Test Unit	K50 N=10	K50 N=15	K50 N=20	K100 N=10	K100 N=15	K100 N=20
information diffusion	34.30%	28.75%	17.94%	38.04%	32.62%	23.16%
multi-grained text	23.05%	18.06%	11.38%	21.79%	19.25%	14.87%

4.6. Ablation Study

CRTM (-information diffusion) gets poor coherence scores in almost all ablation experiments. It demonstrates the effectiveness of the RC module is stronger than multi-grained text. In order to further explore the specific impact of two variants on performance, we compare variants in Table 3. With the rise of N, the performance improvements of both variants drop. It indicates further information diffusion and multi-grained text are both sensitive to noise words. The diffusion mechanism is more affected may be because it is more focused on words with a higher probability of topic words. And the improvement effect when $K50$ is lower than that when $K100$, indicating that the RC module is more effective when the topic is more detailed. When $K50$ and $N = 20$, the performance improvement percentage of multi-grained text representation is the lowest, only 11.38% of the original model, indicating that the combination of multi-grained text and information diffusion can better complement each other.

4.7. Parameter Analysis

Heat Kernel (HK) (Kondor and Lafferty, 2002) and Personalized PageRank (PPR) (Page et al., 1999) are two classic graph diffusion convolution cases. They are mainly used and compared in our experiment. Weighting coefficient $\theta_k^{HK} = e^{-t}\frac{t^k}{k!}$ with the iteration time t in the HK and $\theta_k^{PPR} = \alpha(1-\alpha)^k$ with the transition probability $\alpha \in (0, 1)$ in PPR. We sparse the matrix from two aspects. One is selecting nodes in the top N^t importance rankings. And the other is setting a threshold eps and deleting the edges whose weight coefficient does not exceed eps. The best values of N^t and eps are affected by the dataset and diffusion approach. We utilise Bayesian optimization to find their best values. We verify the impact of N^t on sparsity in Fig. 4(c). When t is 6, the information diffusion is the most sufficient. If t is greater than 6, overfitting will occur. If α is less than 0.4, the information received by searching instead of spreading in social networks will decrease, and the performance will fall. α too large may bring serious deviation. There is no significant difference in the best topic coherence score of PPR and HK. Under both diffusion methods, the sparsized nodes have the highest coherence score in the top 70%, indicating that the nodes obtained by HK and PPR have roughly the same ranking. In addition, the coherence score curve has a slighter fluctuation range under t than under α, which indicates that HK is more stable.

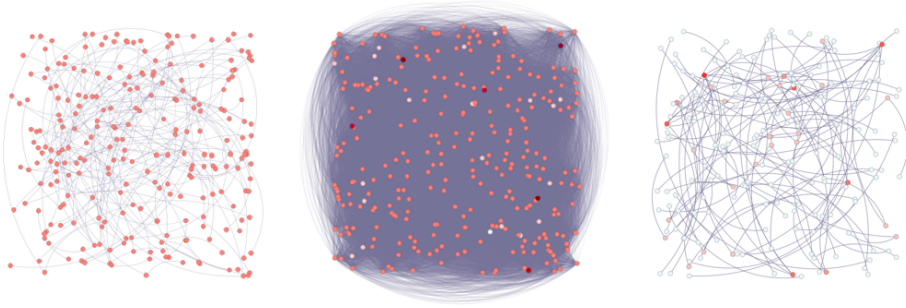
5. Case Study

To get an intuitive understanding of extracted topics, we design an experiment to visualize the top 10 words about “MI press conference" induced by the different models when $K = 50$, depicted in Tab.4. Due to the limited space, we choose representative models that analyze content (LCTM) and integrate social context (IATM) as competitors respectively. Red and italic words are considered to have low relevance to the topic. We have the following observations:

Table 4: TOP 10 topic words for the latent topic “MI PRESS CONFERENCE".

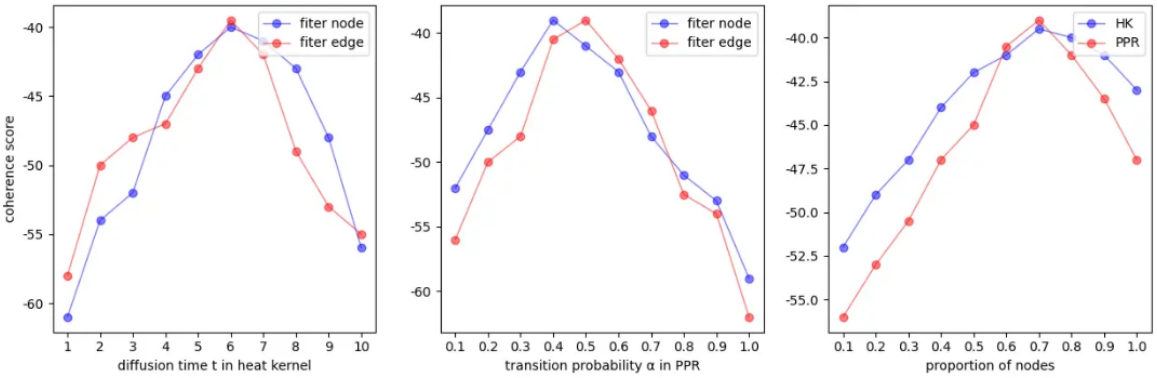
Model	Top 10 Topic Words
LCTM	收件人, 价格, 身体乳, 可能, 香草, 员工, 星期一, 胳膊, 新品, 流量
IATM	手表, Apple(手机),程序, 专家 , 手机, 预定,雷军(老板), 宝宝, 帮忙, 按
CRTM	欢迎, 免费, 太棒了, 勇气, 徕卡, 退钱, 在售, 支持, 高端机, 竞争

- In the given example, the top 10 topic words computed by our method are the best, IATM is slightly worse, and LCTM is the worst. “Recipient", “lotion", “vanilla" and “arm" in LCTM’s inference are less related to the topic. LCTM introduces latent concepts to train word embedding which produces noise. It can be seen that IATM also infers irrelevant words, such as “procedure”, “expert” and “baby". IATM makes full use of context with the aid of the original social structure. It ignores that the relationship includes different intimacy, so it loses many details.
- The top words obtained by our model are the most relevant to the topic. Because we consider the changes in the relationship, we can better simulate the selective propagation process, and capture diverse concerned aspects



(a) original social network (b) continuous social network (c) final social network

Figure 3: Visualization of the continuous social network.



(a)The diffusion algorithm is heat kernel. (b)The diffusion algorithm is personal PageRank. (c)Sparse method of filtering nodes.

Figure 4: The performance trend under the combination of diffusion and sparsification method. Test on the May dataset with the $K50$ and $N10$. N^t takes the first 70%, and eps takes 0.001 in (a)(b). $t = 6$ and $\alpha = 0.4$ in (c).

of the topic. “Popular" and “great" focus on the audience’s reaction. “Leica" and “high-end" focus on the quality of products. “Refund" and “on sale" focus on the purchase and other services.

6. Visualization of Social Network

In order to more intuitively display the structural changes of social network in RC module, we randomly select 300 users from the May dataset. Original social network graph in Fig. 3(a). Fig. 3(b) is network after information diffusion. There are edge connections between any two nodes, and the node color represents the average intimacy. Fig. 3(c) is the sparse social network after removing edges whose weight is less than eps.

7. Conclusion

We propose a **C**ontinuous **R**elational Diffusion driven **T**opic **M**odel (CRTM) with multi-grained text for Microblog. Information diffusion with intimacy supplements context and alleviates data sparsity to a certain extent. Multi-grained text representation reduces the impact of informal expression. GCNs are used to integrate the structure embeddings with intimacy and hierarchical text embeddings. At the same time, VAE generates more coherent topics. Furthermore, we conduct extensive combination experiments and perform parametric analysis. Experimental results demonstrate the effectiveness of CRTM.

8. Limitations

Although our model obtains satisfying results, it also exposes some limitations. **First**, since the difficulty of obtaining data, we mainly carry out relevant experiments and analysis on Microblog datasets. In the future, We plan to comprehensively evaluate our model on more datasets, including datasets from other social platform. **Second**, the method used in this article takes coherence as the only objective evaluation indicator, but the downstream application of the topic model is very extensive. Coherence alone cannot effectively reflect the practicality and diversity of the topic words. In future

work, we will attempt to explore the performance of topic models from the perspective of interpretability to help downstream tasks better utilize topics.

9. Ethics Statement

In this work, we conformed to recognized privacy practices and rigorously followed the data usage policy. We do not introduce social/ethical bias into the model or amplify any bias from the data. Therefore, we do not see any potential risks.

10. Acknowledgement

Our Work is supported by the National Natural Science Foundation of China (No. 62376192, No. 62376188).

11. Bibliographical References

Aly Abdelrazek, Yomna Eid, Eman Gawish, Walaa Medhat, and Ahmed Hassan. 2023. Topic modeling algorithms and applications: A survey. *Information Systems*, 112:102131.

David Alvarez-Melis and Martin Saveski. 2016. Topic modeling in twitter: Aggregating tweets by conversations. In *tenth AAAI on web and social media*, pages 519–522.

James Atwood and Don Towsley. 2016. Diffusion-convolutional neural networks. *Advances in Neural Information Processing Systems*.

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

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, 22:288–296.

Weizheng Chen, Jinpeng Wang, Yan Zhang, Hongfei Yan, and Xiaoming Li. 2015. User based aggregation for biterm topic model. In *Proceedings of the 53rd ACL and the 7th IJCNLP*, pages 489–494.

Jiachun Feng, Zusheng Zhang, Cheng Ding, Yanghui Rao, Haoran Xie, and Fu Lee Wang. 2022. Context reinforced neural topic modeling over short texts. *Information Sciences*, 607:79–91.

Johannes Gasteiger, Stefan Weißenberger, and Stephan Günnemann. 2019. Diffusion improves graph learning. *Advances in neural information processing systems*.

Weiyu Guo, Shu Wu, Liang Wang, and Tieniu Tan. 2015. Social-relational topic model for social networks. In *Proceedings of the 24th CIKM*, pages 1731–1734.

Ruifang He, Xuefei Zhang, Di Jin, Longbiao Wang, Jianwu Dang, and Xiangang Li. 2018. Interaction-aware topic model for microblog conversations through network embedding and user attention. In *Proceedings of the 27th COLING*, pages 1398–1409.

Weihua Hu and Jun'ichi Tsujii. 2016. A latent concept topic model for robust topic inference using word embeddings. In *Proceedings of the 54th ACL*, pages 380–386.

Seiichi Inoue, Taichi Aida, Mamoru Komachi, and Manabu Asai. 2021. Modeling text using the continuous space topic model with pre-trained word embeddings. In *Proceedings of the 59th ACL and the 11th IJCNLP*, pages 138–147.

Akanksha Joshi, Eduardo Fidalgo, Enrique Alegre, and Laura Fernández-Robles. 2023. Deepsumm: Exploiting topic models and sequence to sequence networks for extractive text summarization. *Expert Systems with Applications*, 211:118442.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Risi Imre Kondor and John Lafferty. 2002. Diffusion kernels on graphs and other discrete structures. In *Proceedings of the 19th ICML*, pages 315–322.

Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. 2016. Topic modeling for short texts with auxiliary word embeddings. In *Proceedings of the 39th SIGIR*, pages 165–174.

Jing Li, Ming Liao, Wei Gao, Yulan He, and Kam-Fai Wong. 2018. Topic extraction from microblog posts using conversation structures. In *Social Media Content Analysis: Natural Language Processing and Beyond*, pages 419–437.

Dairui Liu, Derek Greene, and Ruihai Dong. 2022. A novel perspective to look at attention: Bi-level attention-based explainable topic modeling for news classification. *arXiv preprint arXiv:2203.07216*.

Huanyu Liu, Ruifang He, Haocheng Wang, and Bo Wang. 2020. Fusing parallel social contexts within flexible-order proximity for microblog topic detection. In *The 29th CIKM*, pages 875–884.

Heng-Yang Lu, Lu-Yao Xie, Ning Kang, Chong-Jun Wang, and Jun-Yuan Xie. 2017. Don't forget the quantifiable relationship between words: Using recurrent neural network for short text topic discovery. In *Thirty-First AAAI Conference on Artificial Intelligence*, pages 868–869.

Daniel Maier, Christian Baden, Daniela Stoltenberg, Maya De Vries-Kedem, and Annie Waldherr. 2022. Machine translation vs. multilingual dictionaries assessing two strategies for the topic modeling of multilingual text collections. *Communication methods and measures*, 16(1):19–38.

Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. 2013. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th SIGIR*, pages 889–892.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of NAACL-HLT*, pages 746–751.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical report.

Tian Shi, Kyeongpil Kang, Jaegul Choo, and Chandan K Reddy. 2018. Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In *Proceedings of the 2018 WWW*, pages 1105–1114.

Haocheng Wang, Ruifang He, Huanyu Liu, Chenhao Wu, and Bo Wang. 2022. Topic model on microblog with dual-streams graph convolution networks. In *Proceedings of the IJCNN*, pages 1–8.

Yiming Wang, Ximing Li, Xiaotang Zhou, and Jihong Ouyang. 2021. Extracting topics with simultaneous word co-occurrence and semantic correlation graphs: Neural topic modeling for short texts. In *Findings of EMNLP 2021*, pages 18–27.

Xiaobao Wu, Chunping Li, Yan Zhu, and Yishu Miao. 2020. Short text topic modeling with topic distribution quantization and negative sampling decoder. In *Proceedings of the 2020 Conference on EMNLP*, pages 1772–1782.

Xiaobao Wu, Anh Tuan Luu, and Xinshuai Dong. 2022. Mitigating data sparsity for short text topic modeling by topic-semantic contrastive learning. In *Proceedings of the 2022 Conference on EMNLP*, pages 2748–2760.

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In *Proceedings of the 22nd WWW*, pages 1445–1456.

Ce Zhang and Hady W Lauw. 2020. Topic modeling on document networks with adjacent-encoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6737–6745.

Linhai Zhang, Xuemeng Hu, Boyu Wang, Deyu Zhou, Qian-Wen Zhang, and Yunbo Cao. 2022. Pre-training and fine-tuning neural topic model: A simple yet effective approach to incorporating external knowledge. In *Proceedings of the 60th ACL*, pages 5980–5989.