Topic Modeling for Noisy Short Texts with Multiple Relations

Chiyu Liu, Zheng Liu, Tao Li and Bin Xia
Jiangsu Key Laboratory of Big Data Security & Intelligent Processing
School of Computer Science, Nanjing University of Posts and Telecommunications
Nanjing 210023, People's Republic of China
{1016041229, zliu, towerlee,bxia}@njupt.edu.cn

***Abstract*—Understanding contents in social networks by inferring high-quality latent topics from short texts is a significant task in social analysis, which is challenging because social network contents are usually extremely short, noisy and full of informal vocabularies. Due to the lack of sufficient word co-occurrence instances, well-known topic modeling methods such as LDA and LSA cannot uncover high-quality topic structures. Existing research works seek to pool short texts from social networks into pseudo documents or utilize the explicit relations among these short texts such as hashtags in tweets to make classic topic modeling methods work. In this paper, we explore this problem by proposing a topic model for noisy short texts with multiple relations called MRTM (Multiple Relational Topic Modeling). MRTM exploits both explicit and implicit relations by introducing a *document-attribute distribution* and a *two-step random sampling strategy*. Extensive experiments, compared with state-of-the-art topic modeling approaches, demonstrate that MRTM can alleviate the word co-occurrence sparsity and uncover high-quality latent topics from noisy short texts.**

***Index Terms*—Topic Modeling · Multiple Relations · Short Texts · Social Analysis**

I. Introduction

Topic modeling based on probabilistic graphical models with latent variables for uncovering hidden thematic structures is widely used in various applications including content recommendation [1], user profiling [2], trend detection [3], etc. Short texts, especially ones from social networks, tweets, feature short length, inordinate structure, and colloquialism. As a result, uncovering the potential topics from these short texts is not an easy task [4], [5]. Take the short texts from popular social networks such as Twitter as an example. Tweets are short, informal, and lack regular patterns, which leads to poor performance of the classic topic model Latent Dirichlet Allocation (LDA) [6], as well as many other LDA-like topic models. The underlying reason is that there are not enough word co-occurrence instances due to term sparsity in short texts.

Pooling tweets or other short texts [7] by aggregating them into pseudo documents based on their attributes has been proved to be a promising way to improve the quality of topics found by LDA-like methods. Possible attributes could be the authors of short texts [8], [9], or time periods. For tweets from social networks, hashtags [4] or burst scores [10] could also server as the aggregation cornerstones. These pseudo documents by aggregating short texts enrich the word co-occurrence, but on the other hand, they not only bring duplicates of short texts containing multiple attributes as well as new co-occurrence instances which do not exist, but also ignore the relations among these attributes. For example, each tweet could contain more than one hashtag, so some hashtags are likely to appear together than other hashtags. The hashtag correlations are the bridges of words in different short texts, and could help to improve the quality of topics. Wang et al. [11] proposed a Hashtag Graph based Topic Model (HGTM) for tweets, in which user-contributed hashtags are considered in the generative process of tweets. Experimental results show that it is more reliable than the simple pooling strategy.

It is not difficult to see that there are multiple attributes in short texts. Besides the explicit attributes like hashtags or users, there are many other implicit attributes such as various kinds of entities and temporal attributes. In this paper, we use labels to denote the possible values of attributes. For example, if the attribute is actors, then the corresponding labels are actor names. Tweets discussing movies could contain entities like actors, directors, as well as movie genres, movie released date, etc. Each of these attributes could be represented by a relational graph, where a vertex is a possible label of the attribute and an edge indicates the co-occurrence relations between two labels, which we will explain in detail in Section III. These relational graphs of attributes could reveal the semantic associations between labels. On the other hand, unnatural co-occurrence words about the background of short texts are noises, which impact the topic quality. For example, if the short texts are about some movies, then words like movie, film or cinema are not discriminative, while they exist in informal oral presentation [9]. Traditional solutions like TF-IDF can not handle well in short texts. Noisy words are highly related to the domain knowledge of the short texts, and not helpful in understanding the corpus.

With above observations, in this paper, we propose a topic model for noisy short texts with multiple relations, called MRTM, which can uncover meaningful topics. The main idea is to incorporate multiple relations into the generative process of short texts to produce high-quality topics measured both subjectively and objectively. Gibbs Sampling [12] is adopted to estimate the parameters in MRTM. The main contributions of this paper are summarized below.

* Corresponding author: Zheng Liu

DOI reference number: No.10.18293/SEKE2018-022

- The proposed MRTM alleviates the sparsity of word co-occurrence in short texts by incorporating multiple relations into the generative process, resulting in a coherent generative topic model.
- MRTM further improves the quality of the uncovered topics by removing unnatural word co-occurrence instances caused by considering weakly-supervised relations.
- Extensive experiments are conducted on real data sets crawled from Microblog. The experimental results are carefully analyzed, showing that MRTM can uncover coherent topics of higher quality than the start-of-the-art approaches.

The rest of this paper is organized as follows. Section II discusses the related work. Section III describes the proposed multiple relational topic model in detail, as well as the inference process of its parameters. Section IV presents the experimental results and finally, Section V concludes the paper with possible future directions.

II. Related Work

Classic topic models such as Latent Dirichlet Allocation (LDA) [6] suffers term sparsity and noises when applied to short texts, resulting in low-quality topics due to insufficient word co-occurrence instances. Many researchers studied this problem, where the approaches can be categorized as follows.

Pooling based strategy

Pooling short texts by aggregating them into pseudo documents based on certain attributes [8], [9], [10] has been proved to be a promising way to make LDA-like approaches work. Zhao et al. [9] analyzed the internal characteristic of short texts by introducing topic category and background words. In particular, they found users' topics are concentrated and consist of only a few words. Mehrotra et al. [4] analyzed the extensive experimental results of various pooling attributes and found that not all pooling attributes are helpful in capturing high-quality topics.

Semantic based strategy

Using word semantics as the prior knowledge could benefit topic modeling for short texts, where the prior knowledge is pre-trained word embedding based on the large corpus. Li et al. [13] proposed a topic model based on the Dirichlet Multinomial Mixture (DMM), which is able to find more semantically related word pairs under the same topic during the sampling process. Similarly, Nguyen et al. [14] incorporated vector representations of words during topic modeling to improve the word-topic mapping. The vector representations of words are learned by using a large corpus.

Relation based strategy

Recently, many efforts [11], [15], [16] were put on constraining LDA with semi-structured relations. By integrating this kind of prior knowledge, the generative processes in these models for short texts are more reasonable. Rosen et al. [8] proposed an author-topic model (ATM) for documents to include authorship information. Each author is associated with a topic distribution. However, it is not appropriate for short texts. Daniel et al. [15] proposed Labeled LDA to learn one-to-one mappings between topics and labels, but they ignored the correlations between labels. Labeled-LDA is a strong supervised model that labels have equal impact on short texts. Wang et al. [11] improved the graphical models in LDA family with hashtag graph based topic model (HGTM). Unlike pooling based strategy in which tweets are aggregated into pseudo documents, it incorporated a hashtag graph into topic modeling. The key differences between HGTM and MRTM in this paper could be explained in terms of both complexity and robustness. HGTM only considers a single explicit relation, i.e., hashtags, while MRTM incorporates multiple relations into topic modeling. Moreover, MRTM integrates a new hidden variable which makes it more robust than HGTM. Note that multiple relations represented by information networks have various inherent textual information. Their rich semantics could enhance the inherent coherence among texts, and as a result, MRTM could uncover topics with better quality which is shown in Section IV.

III. The Proposed Approach

A. Notations

Let $D = \{d_1, d_2, \ldots, d_m\}$ denote the corpus of short texts where corpus and d_i is a short text. Let $W = \{w_1, w_2, \ldots, w_n\}$ denote the vocabulary set of D, where w_i is a word. Let $C = \{c_1, c_2, \ldots, c_k\}$ denote the attribute sets where c_i is an attribute. In the following of this paper, we also use c to denote a certain attribute.

Take the tweets of movie reviews as an example. Fig. 1 shows the overall concept of multiple relational graphs. Recall that labels denote values of a certain attribute. A tweet could contain hashtags, as well as many other entities, such as title, actors, released time, etc., as shown in Fig. 1. Then we can construct a relation graph $g^c = (V_c, E_c)$ for each attribute c, where V_c is the vertex set in which each vertex represents a label belong to attribute c, and E_c is the edge set in which each edge represents the relation between two vertices. For each attribute c, let $L^c = \{l_1^c, l_2^c, \ldots, l_o^c\}$ represent the label set of attribute c. g_{ij}^c is a weighted relation between label l_i^c and l_j^c vertices in the relation graph g^c of attribute c.

Let a denote the relation of actors and h denote the relation of hashtags, respectively. Then we can construct the relation graph $g^a = (V_a, E_a)$ of actors and the relation graph $g^h = (V_h, E_h)$ of hashtags. In the relation graph g^a, each edge indicates the relation between two actors who are likely to co-occur in one tweet. In the relation graph g^h, each edge indicates relations between two hashtags which are likely to co-occur in one tweet. The weights of the edges in both graphs are the number of co-occurrence instances of the adjacent vertices in the corpus of short texts. As shown in Fig. 1, usually there are multiple relation graphs.

B. Multiple Relational Topic Modeling

θ_1^K is the distribution over topics with dirichlet prior parameter α. ϕ_k is the topic-word distribution with dirichlet prior parameter β. z represents the topic assignment matrix and z_d represents the topic assignment for a short text d. w_d is the word sequence of short text d and w_{di} represents the word at position i in d. Then the parameters of MRTM are as follows.

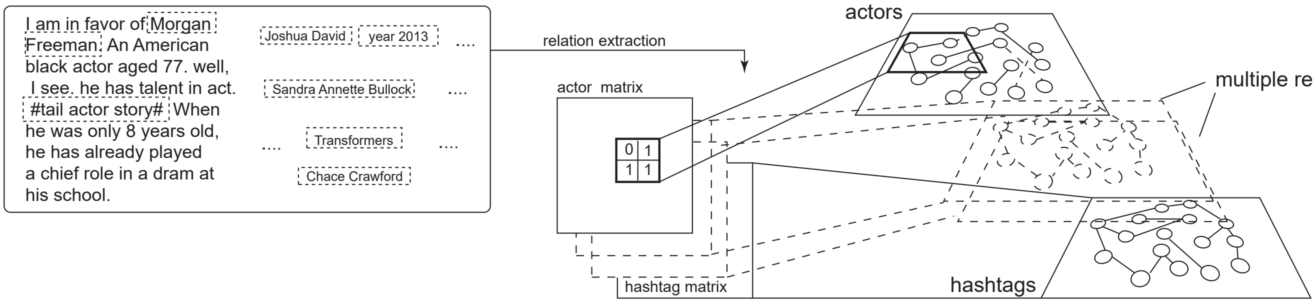
I am in favor of Morgan Freeman. An American black actor aged 77. well, I see. he has talent in act. #tail actor story# When he was only 8 years old, he has already played a chief role in a dram at his school.
Joshua David
year 2013
....
Sandra Annette Bullock
....
....
Transformers
....
Chace Crawford
relation extraction
actor matrix
0 1
1 1
hashtag matrix
actors
multiple relations
hashtags

Fig. 1: Multi relations extracted from short texts

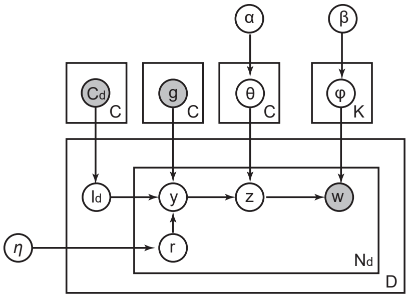

Fig. 2: The model structure of MRTM

$$\theta_c|\alpha \sim Dirichlet(\alpha) \quad (1)$$

$$\phi_k|\beta \sim Dirichlet(\beta) \quad (2)$$

$$z_{di} \sim Multinomial(\theta_{y_{di}}) \quad (3)$$

$$w_{di} \sim Multinomial(\phi_{z_{di}}) \quad (4)$$

Fig. 2 presents the structure of our proposed multiple relation topic model. A unique topic distribution is associated with each label in each attribute. Each topic is represented as a multinomial distribution over words. A short text could include two or more attributes, and attributes can serve as bridges between short texts. In order to incorporate multiple relations into our topic model, we first introduce the following concepts.

Document-attribute distribution C_d. Documents could prefer some attributes. For example, users talking about a movie are likely to discuss actors in the movie rather than hashtag concepts. Vice verse, users might be willing to talk about hashtag concepts than actors. We use document-attribute matrix to model this kind of preference of attributes, and each attribute has its unique distribution of topics.

Observed attribute labels l_d. Observed attribute labels refer to the labels existing in a short text. As prior knowledge, it is better than the simple pooling strategy [11].

Potential attribute labels y^p. Potential attribute labels refer to the labels inferred from the whole corpus based on the graph based attribute relation. Related labels of the observed labels are also be utilized as prior knowledge.

Let l denote the attribute assignment and l_d be assignment vector of short text d. y_{di} represents the attribute assignment for position i in short text d. Different short texts have different attribute preference, represented by C_d, the document-attribute distribution. By introducing an indicator parameter r to decide whether we choose the observed attribute label or potential attribute label(s), the label assignment at position i of short text d, y_{di} is defined in Eq. 5.

$$y_{di}|C_d, g^c, r \sim Bernoulli(\eta) \quad (5)$$

where η controls the randomness of attribute labels. Large η means more randomness.

The overall generative process for MRTM is described below.

1) for each of the label l in attribute C, $l \in \{1, \ldots, L\}$, sample $\theta_l \sim Dir(\alpha)$;
2) for each of the topic $k \in \{1, \ldots, K\}$, draw words $\phi_k \sim Dir(\beta)$;
3) for each of the documents $d \in \{1, \ldots, D\}$, give all document-attribute distribution C_d and all prior knowledge g^c;
4) for each word $w_{di} \in d$, $i \in \{1, \ldots, N_d\}$
5)   sample $c_{di} \sim Multinomial(C_d)$;
6)   sample observed attribute label $l_{di} \sim Uniform(c_{di})$
7)   sample potential attribute label related with observed label in graph g^c, $y^p \sim Multinomial(norm(g^{c_{di}}_{l_{di}}))$;
8)   sample $r \sim Bernoulli(\eta)$;
9)   if r is 0, sample $y_{di} = l_{di}$;
10)   else if r is 1, sample $y_{di} = y^p$;
11)   draw topic $z_{di} \sim Multinomial(\theta_{y_{di}})$;
12)   draw word $w_{di} \sim Multinomial(\phi_{z_{di}})$.

Firstly we sample current label based on C_d. Secondly, we sample a value of r from Bernoulli distribution and decide whether this label is related with current word. If not, we sample from highly related labels from the relation graph. Through the two-step sampling from line 6 to 10, we can obtain the document-attribute assignment. $norm(g^{c_{di}}_{l_{di}})$ in line 7 is a normalized L-dimension possibility vector, where L is the number of labels in attribute c and each element in the vector is calculated by using the following equation.

$$p(l_j|l^{c_{di}}_{di}) = \frac{g^{c_{di}}_{l_{di},l_j}}{\sum_{l'=1}^{L}(g^{c_{di}}_{l_{di},l_{l'}})} \quad (6)$$

This model adds word co-occurrence under graph structure when latent relationships between short texts are found, and we filter unnatural word co-occurrence caused by merely aggregating short texts. In our experiment after introducing actor networks or hash-tag network, we enhance the semantic information in movie reviews. More exciting is that we can add more relations to this framework in general tasks based on short text topic modeling.

C. Model Parameter Inference

The joint distribution of the latent variables is

$$p(\boldsymbol{w}|\theta,\phi,r,\boldsymbol{l},G)=\prod_{d=1}^{D}p(\boldsymbol{w_d}|\theta,\phi,r,\boldsymbol{l_d},\boldsymbol{g^c}). \tag{7}$$

Assuming that attribute-topic distribution $\boldsymbol{l_d}$ and topic-word distribution ϕ are independent, we have

$$\begin{aligned}p(\boldsymbol{w_d}|\theta,\phi,r,\boldsymbol{l_d},g^c)&=\prod_{i=1}^{N_d}p(w_{di}|\theta,\phi,r,\boldsymbol{l_d},\boldsymbol{g^c})\\&=\prod_{i=1}^{N_d}\sum_{c=1}^{C}\sum_{k=1}^{K}p(w_{di},z_{di}=k,y_{di}=c|\\&\qquad\qquad\theta,\phi,r,\boldsymbol{l_d},\boldsymbol{g^c})\\&=\prod_{i=1}^{N_d}\sum_{c=1}^{C}\sum_{k=1}^{K}\phi_{w_{di}}\theta_{kc}p_{cy_{di}}\end{aligned} \tag{8}$$

where $p_{cy_{di}}=p(y_{di}=c|r,\boldsymbol{l_d},g^c)$ represents the assignment possibility of attribute c on attribute-topic distribution $\boldsymbol{l_d}$ and potential labels.

According to the two-step sampling mentioned in Section III-B, the assignment of c is associated with relation graph $\boldsymbol{g^c}$. Then we have

$$\begin{aligned}p(y_{di}=&c|r,\boldsymbol{l_d},\boldsymbol{g^c})\\=&\Big\{p(l^c_{di}=c|l_d)p(y_{di}=c|l^c_{di})\Big\}^{r}\times\\&\Bigg\{\sum_{l'=1}^{L}p(l^c_{di}=l_{l'}|C_d)p(y_{di}=c|l^c_{di}=l_{l'},\boldsymbol{g^c})\Bigg\}^{1-r}\end{aligned} \tag{9}$$

It is computational infeasible to estimate directly the conditional probability distribution $p(\boldsymbol{w_d}|\theta,\phi,r,\boldsymbol{l_d},g^c)$, like many other topic modeling approaches, we adopt Gibbs sampling [12] to approximate it as follows.

$$\begin{aligned}&p(z_{di}=k,y_{di}=c,r_{di}=u|\\&\qquad w_{di}=v,\boldsymbol{z_{-di}},\boldsymbol{y_{-di}},\boldsymbol{w_{-di}},C,G,\alpha,\beta,\eta)\\&\propto\frac{N^{VK}_{vk,-di}+\beta}{\sum_{v'}N^{VK}_{v'k,-di}+V\beta}\cdot\frac{N^{KC}_{kc,-di}+\alpha}{\sum_{k'}N^{KC}_{k'c,-di}+K\alpha}\cdot p_{cy_{di}}\end{aligned} \tag{10}$$

Let N^{KC} be the matrix recording the number of times that a topic is assigned to some attribute. Let N^{VK} be the matrix recording the number of times that a real word is assigned to some topic. After iterative sampling, the final θ_c and ϕ_k are as follows.

$$\begin{aligned}\hat{\theta_c}&\propto\frac{N^{KC}_{kc}+\alpha}{\sum_{k'}N^{KC}_{k'c}+K\alpha}\\\hat{\phi_k}&\propto\frac{N^{VK}_{vk}+\beta}{\sum_{v'}N^{VK}_{v'k}+V\beta}\end{aligned} \tag{11}$$

IV. Experimental Evaluation

In this section, we report our experimental results. The quality of uncovered topics of short texts is evaluated using both subjective and objective metrics. All experiments are done on a PC with Intel i5 CPU at 2.3 GHz and 8GB memory, running Windows 10. All algorithms are implemented in Python.

A. Datasets and Settings

We collected more than 150,000 tweets from Microblog[1], which is a Chinese social network site similar to twitter. All tweets are in Chinese, and related to Chinese movies released in 2017 in order to narrow down the domain of the potential topics for analysis. Unlike English, sentences in Chinese do not contain spaces between words. We applied JieBa[2] (an open-source NLP tool for Chinese) to segment sentences into words and remove stop words.

It is worth noting that the relations utilized in the proposed topic model is widely available in short texts in many applications. In particular, these attributes, labels, and relations could be manually defined, automatically learned, or extracted from existing knowledge bases. In the experiments, we extract relations by using both knowledge based matching and RegExp based matching.

B. Subjective Quality Evaluation

We report the uncovered topics of each method and evaluate their quality in a subjective view in this section. We conducted the experiments on tweets about five most popular Chinese movies released in 2017. The characteristics of the movies and the tweets are presented in Table II. For readers not speaking Chinese, we translated all the Chinese words into English in Table II, as well as ones in the following table, where phrases in italic are movie titles, and phrases with underline are actor names.

We compared the proposed MRTM with the state-of-the-art topic model for short texts, i.e., HGTM [11], as well as the classic LDA topic model [6] as a baseline. Recall that HGTM is the topic model for short texts based on the hashtag graph. In all models, the topic number $K=150$, the latent variable $\alpha=0.5$, and $\beta=0.1$. We set $\eta=0.5$ in HGTM as indicated in [11]. The number of iteration times for Gibbs sampling is set to be 1000 in all experiments.

Table I shows the top 10 words from top 1 topic uncovered by each models ranking based on probabilities. The irrelevant words found by LDA are marked with label irrelevant in parentheses. New words found by HGTM and MRTM are in bold. We also summarized the new words of HGTM and MRTM in Table III, by categorizing all new words into two groups, major actors and relevant words consistent with movie genres.

By careful analysis of the discovered topics together with the original tweets, we have the following observations. For Movie #1 and #2, the topic found by LDA contains many

[1] http://www.weibo.com

[2] https://github.com/fxsjy/jieba

TABLE I: Top 1 topic dicovered by LDA, HGTM and MRTM

	Movie #1	Movie #2	Movie #3	Movie #4	Movie #5
LDA	*Taohua*, *Sansheng*, film, *Sanshi*, *Shili*, happiness,summer-vacation, interesting, movie season, not bad	film, sacrifice, suspect, **_duo_(irrelevant)**, **_meng_(irrelevant)**, superise, not bad, **both(irrelevant)**, propaganda, acting skill	support, *Zhanlang*, movie season, happiness, interesting, film, good, summer vacation, dream, Chinese	memory, film, master, acting skill, plot, reversal, interesting, murderer, really, Huang Bo	new year movie, movie-season, not good, strongest, Jackie Chan, more and more, interesting, film, *Yoga*, not bad
HGTM	movie season, summer-vacation, happiness, *Sansheng*, *Sanshi*, *Shili*, *Taohua*, film, **Yang Yang**, **special effects**	film, suspect, sacrifice, like, support, **awesome**, fighting, superise, acting skill, propaganda	Chinese, *Zhanlang*, **Wu Jing**, film, box-office, **pride**, support, **strike**, **poke**, enjoy	memory, master, Huang Bo, **Duan Yi Hong**, film, expect, eyesight, plot, reversal, murderer	***Kung Fu***, *Yoga*, strongest, interesting , Jackie Chan, **Zhang Yi Xin**, brother, excellent, **laugh**, like
MRTM	summer vacation, **Yang Yang**, interesting, **special effects**, *Sansheng*, *Sanshi*, *Shili*, *Taohua*, film, movie season	suspect, film, sacrifice, awesome, like, superise, acting-skill, love, enjoy, **fear**	*Zhanlang*, **Wu Jing**, patriotic, support, strong, **strike**, Chinese, film, scene, enjoy	memory, master, Huang Bo, **Duan Yi Hong**, murderer, plot, reversal, film, **killer**, except	***Kung Fu***, Jackie Chan, interesting ,*Yoga*, **laugh** , **zhang yi xin**, love, excellent, **indian**, **funny**

TABLE II: The characteristics of movies and tweets

ID	Movie Title	Movie Genre	# of Tweets
#1	Sansheng Sanshi Shili Taohua	romance, fantasy	9,932
#2	The Devotion Of Suspect X	feature, crime	9,215
#3	Wolf Warriors II	action, military	19,230
#4	Battle of Memories	suspense, crime	9,355
#5	Kung-Fu Yoga	comedy, adventure	7,375

TABLE III: Words discovered by HGTM and MRTM

ID	Words of leading actor	Words related to movie genres
#1	Yang Yang	special effects
#2	-	awesome, fear
#3	Wu Jing	pride, strike, poke
#4	Huang Bo, Duan Yi Hong	killer
#5	Zhang Yi Xin	*Kung Fu*, laugh, Indian, funny

irrelevant words. For Movie #3 ,#4, and #5, the topic contains major actors names such as Huang Bo (A Chinese actor) and Jackie Chan. In general, it seems that words in LDA's topics provide an overview but also bring some unnatural words such as 'not bad' or 'film'(background noise, [9]) which caused by unnatural co-occurrence.

For HGTM and MRTM, in Movie #1, they both found out the leading actor 'Yang Yang' with different rankings. More substantive words have higher probabilities than words in the movie title, such as *Sansheng*, *Sanshi*, *Shili*, *Taohua*. In Movie #2 and #3, both HGTM and MRTM can get rid of irreverent words and find new words . In Movie #4, movie genre related word like 'killer' is within the top 10-word list, substituting the irrelevant word 'really'. In Movie #5, word 'laugh' shows this movie is a comedy. The word 'Indian' appearing in MRTM's topics is because the story of the movie took place in India.

C. Objective Semantic Coherence

In this section, we report the uncovered topics of each method and evaluate their quality in an objective view. We employed Pointwise Mutual Information (PMI) [17] to measure the topic coherence, which has been proved to be an effective measure for topic quality [18]. Given a topic t and its top K words $W^t = (w_1^t, w_2^t, \dots, w_K^t)$ (The top K words with highest probabilities.), $P(w)$ denotes the document frequency of word w and $P(w_l, w_l')$ is the probability w_l and w_l' co-occur. The metric is defined for a specific topic t as

$$PMI(w_l^t, w_{l'}^t) = \log(\frac{P(w_l^t, w_{l'}^t)}{P(w_l^t)P(w_{l'}^t)}). \qquad (12)$$

Then the coherence of a top-K models is the summarization of the PMI scores of each topic as follows. Larger coherence score shows better topic partition.

$$Coherence(t, W^t) = \sum_{i=2}^{K}\sum_{j=1}^{i} \log(\frac{P(w_i^t, w_j^t) + \epsilon}{P(w_i^t)P(w_j^t)}) \qquad (13)$$

We filtered the tweet corpus by removing tweets of unpopular movies and split the remaining tweets into two data sets, MR1 and MR2. MR1 has 2,3452 tweets with the average length 6.96, wile MR2 has 7,329 tweets with the average length 7.97. We compared MRTM with the classic topic model Latent Dirichlet Allocation **(LDA)** [6], the author topic model **(ATM)** [8], and the hashtag graph based topic model **(HGTM)** [11].

We extract top 10 words for each topic generated by each model for computing the coherence. The number of topics is 100, $\alpha = 0.3$ and $\beta = 0.1$. The number of iterations of Gibbs sampling is 200 which is enough for uncovering topics. We conduct all the experiments repeatedly for 5 times and report the mean value of each measure in Table IV. For each topic,

TABLE IV: Average PMI and coherence scores of in MR1 and MR2

Model	dataset	MR1					MR2				
	K	5	10	15	20	30	5	10	15	20	30
LDA [6]	Average PMI	-6.3264	-7.9251	-8.3323	-8.6869	-9.1779	-3.8294	-5.6971	-6.9512	-8.0397	-8.9713
	Coherence	-31.632	-79.251	-124.9845	-173.738	-275.337	-19.147	-56.971	-104.268	-160.794	-269.139
ATM [8]	Average PMI	-7.8479	-8.9783	-9.6743	-9.5824	-9.546	-6.3908	-7.5662	-8.0043	-8.1444	-8.6616
	Coherence	-39.2395	-89.783	-145.1145	-191.648	-286.38	-31.954	-75.662	-120.0645	-162.888	-259.848
HGTM [11]	Average PMI	**-1.9698**	-3.9027	-5.0757	-5.4254	-6.243	-3.0478	-3.6876	-4.2665	-4.8512	-5.1454
	Coherence	**-9.849**	-39.027	-76.1355	-108.508	-187.29	-15.239	-36.876	-63.9975	-97.024	-154.362
MRTM	Average PMI	-2.3854	**-2.7416**	**-3.0652**	**-3.3337**	**-3.6853**	**-2.7831**	**-3.2135**	**-3.9521**	**-4.3683**	**-4.8573**
	Coherence	-11.927	**-27.416**	**-45.978**	**-66.674**	**-110.559**	**-13.9155**	**-32.135**	**-59.2815**	**-87.366**	**-145.719**

we only keep K words with the largest probabilities, and the average PMI indicates the average PMI scores of all topics.

MRTM achieves lowest scores in all settings except when topic length is 5 in MR1. Classic LDA and ATM show lower coherence because they cannot handle the sparsity of short texts. For all approaches, the coherence becomes unstable when K goes larger generally. However, an interesting fact we found is that using hashtags is more stable than using other relations, and with multiple relations, we can further lower the trend of increasing. Another observation is related to the range of average PMI and coherence along with the change of K. The range in MRTM is much smaller than the one in HGTM, i.e., 60% approximately. With smaller range of average PMI and coherence, MRTM is more robust than HGTM with respect to K.

It is essential to incorporate multiple relations appropriately. Otherwise, it might deteriorate the quality of discovered topics. In the proposed MRTM, we introduce new latent document attribute layer and incorporate multiple relations to obtain high coherent topics.

V. Conclusion

A Multiple Relation Topic Model (MRTM) is proposed in this paper with the aim of overcoming the difficulties of sparsity and informality caused by noisy short texts. By incorporating explicit and implicit relations among short texts into the generative process of short texts, MRTM can uncover high-quality topics. Extensive experiments demonstrate that MRTM can achieve better performance than both the classic topic model approach LDA, and the state-of-the-art topic modeling approaches, i.e., ATM and HGTM. Possible future directions include accelerating the sampling speed and trading off between explicit and implicit relations of short texts.

Acknowledgements

This work is supported in part by Jiangsu Provincial Natural Science Foundation of China under Grant BK20171447, Jiangsu Provincial University Natural Science Research of China under Grant 17KJB520024, Jiangsu Key Laboratory of Big Data Security & Intelligent Processing under Grant BDSIP1803 and Nanjing University of Posts and Telecommunications under Grant No. NY215045.

References

[1] R. Krestel, P. Fankhauser, and W. Nejdl, "Latent dirichlet allocation for tag recommendation," in *Proceedings of the third ACM conference on Recommender systems*. ACM, 2009, pp. 61–68.

[2] A. McCallum, X. Wang, and A. Corrada-Emmanuel, "Topic and role discovery in social networks with experiments on enron and academic email," *Journal of Artificial Intelligence Research*, vol. 30, pp. 249–272, 2007.

[3] J. H. Lau, N. Collier, and T. Baldwin, "Online trend analysis with topic models: twitter trends detection topic model online." in *International Conference on Computational Linguistics*, 2012, pp. 1519–1534.

[4] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie, "Improving lda topic models for microblogs via tweet pooling and automatic labeling," in *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2013, pp. 889–892.

[5] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in *Proceedings of the 22nd international conference on World Wide Web*. ACM, 2013, pp. 1445–1456.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3.

[7] Y. Zuo, J. Wu, H. Zhang, H. Lin, F. Wang, K. Xu, and H. Xiong, "Topic modeling of short texts: A pseudo-document view," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 2105–2114.

[8] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *Proceedings of the 20th conference on Uncertainty in artificial intelligence*. AUAI Press, 2004, pp. 487–494.

[9] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," in *European Conference on Information Retrieval*. Springer, 2011, pp. 338–349.

[10] M. Naaman, H. Becker, and L. Gravano, "Hip and trendy: Characterizing emerging trends on twitter," *Journal of the Association for Information Science and Technology*, vol. 62, no. 5, pp. 902–918, 2011.

[11] Y. Wang, J. Liu, J. Qu, Y. Huang, J. Chen, and X. Feng, "Hashtag graph based topic model for tweet mining," pp. 1025–1030, 2014.

[12] T. Griffiths, "Gibbs sampling in the generative model of latent dirichlet allocation," 2002.

[13] C. Li, H. Wang, Z. Zhang, A. Sun, and Z. Ma, "Topic modeling for short texts with auxiliary word embeddings," in *International Acm Sigir Conference on Research & Development in Information Retrieval*, 2016, pp. 165–174.

[14] D. Q. Nguyen, R. Billingsley, L. Du, and M. Johnson, "Improving topic models with latent feature word representations," *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 299–313, 2015.

[15] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, "Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora," in *Conference on Empirical Methods in Natural Language Processing: Volume*, 2009, pp. 248–256.

[16] S. Li, J. Li, and R. Pan, "Tag-weighted topic model for mining semi-structured documents," in *International Joint Conference on Artificial Intelligence*, 2013, pp. 2855–2861.

[17] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2-4, 2010, Los Angeles, California, USA*, 2010, pp. 100–108.

[18] A. Fang, C. Macdonald, I. Ounis, and P. Habel, *Topics in tweets: A user study of topic coherence metrics for Twitter data*, 2016.