Discovering Topics in Long-tailed Corpora with Causal Intervention

Xiaobao Wu[1] **Chunping Li**[1] **Yishu Miao**[2]
[1]Tsinghua University [2]Imperial College London
wxb18@mails.tsinghua.edu.cn, cli@mail.tsinghua.edu.cn,
y.miao20@imperial.ac.uk

Abstract

Topic models are effective in capturing the latent semantics of large-scale textual data while existing methods are normally designed and evaluated on balanced corpora. However, it contradicts the fact that general corpora in our world are naturally long-tailed, and the long-tailed bias can highly impair the topic modeling performance. Therefore, in this paper, we propose a causal inference framework to explain and overcome the issues of topic modeling on long-tailed corpora. In a neat and elegant way, causal intervention is applied in training to take out the influence brought by the long-tailed bias. Extensive experiments on manually constructed and naturally collected datasets demonstrate that our model can mitigate the bias effect, greatly improve topic quality and better discover the hidden semantics on the tail.

1 Introduction

Topic models are proposed to discover the underlying topics and semantic structures from unlabelled text collections. Due to the effectiveness and interpretability, topic models have been applied in various downstream tasks like information retrieval (Wang et al., 2007), content summarization (Ma et al., 2012) and recommendation systems (McAuley and Leskovec, 2013). One of the most widely used topic models is Latent Dirichlet Allocation (LDA) (Blei et al., 2003), a probabilistic graphical model using the conjugate of Dirichlet and Multinomial distribution and inferring the parameters with approximation methods(Griffiths and Steyvers, 2004; Blei et al., 2017). Recently, some popular neural topic models based on Variational AutoEncoder (VAE) (Kingma and Welling, 2014; Rezende et al., 2014) have been introduced, such as Neural Variational Document Model (NVDM) (Miao et al., 2016) and Product of Experts LDA (ProdLDA) (Srivastava and Sutton, 2017). Compared to probabilistic ones, they can easily carry out the inference by gradient backpropagation.

However, these topic models are generally designed and evaluated on balanced corpora, such as the commonly used 20News (Lang, 1995) with evenly distributed labels through which we can infer that the latent topics are also evenly distributed. It hence conflicts with the fact that natural text collections are regularly long-tail distributed following Zipf's law (Reed, 2001), especially the textual data on social network platforms(Zhang and Luo, 2019). More precisely, Figure 1 illustrates that in a collected corpus, documents about some hot topics are numerous (*head topics*), while the documents about most topics are few (*tail topics*). Due to this bias, similar to long-tailed classification tasks where a classifier favors to predict an image as the head classes (Kang et al., 2020; Zhou et al., 2020), topic models on long-tailed corpora tend to reveal the semantics of documents about head topics and ignore the documents about tail topics to a great extent. Namely, the discovered topics are mostly about the latent head ones in the corpus. As a result, their diversities are much impaired and incomplete to represent the whole semantics of a corpus. Thus, it is crucial to explore effective ways for long-tailed topic modeling.

Different from other long-tailed tasks like image classification or relation extraction, the key challenge of this problem lies in that topic modeling is originally designed for unlabelled datasets, so we have no access to classification labels to infer the latent global topic distributions while designing solutions[1]. Owing to this factor, we intend not to introduce complicated modules conditioning

[1]Admittedly, there are supervised topic models (Mcauliffe and Blei, 2007; Card et al., 2018), but the necessity of labels will hugely narrow their application scopes. So, we concentrate on solving the issue without additional labels.

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 175–185

on accessible labels, e.g., re-weighting (Mahajan et al., 2018) or re-sampling (Khan et al., 2017; Lin et al., 2017; Cui et al., 2019) approaches for other long-tail problems.

To overcome this challenge, in this paper, we present a Structural Causal Model (SCM) (Pearl et al., 2016; Pearl and Mackenzie, 2018) to precisely explain how the long-tailed bias undermines the topic modeling performance. Then, to remove the bias effect, we propose an approach via the causal intervention (Pearl et al., 2016) on topic distributions and adopt the backdoor adjustment (Pearl, 1995) to calculate the causality in the condition of no auxiliary information. Furthermore, we introduce a novel neural model named as **Deconfounded Topic Model** (**DecTM**) in the framework of VAE with deconfounded training through an approximation manner. Through comprehensive experiments, we manifest that our new model can mitigate the influence of the long-tailed bias and produce high-quality topics that are more diverse and better disclose the semantics of documents about tail topics. The main contributions of this paper can be concluded as follows:

1. We present a structural causal model to clarify how the problems of topic modeling are incurred by the long-tailed bias in detail;

2. We further propose a neat method to approximate the causal intervention for reducing bias influence, depending on which a novel neural topic is also introduced with deconfounded training;

3. We validate our model on both manually-constructed and extreme multi-label text classification datasets and demonstrate our model is effective to alleviate the impact of bias and greatly improve the topic quality compared to both probabilistic and neural baseline models.

2 Related Work

Topic Modeling Probabilistic topic models can date back to Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999) and LDA (Blei et al., 2003), derving numerous variants(Blei and Lafferty, 2006; Yan et al., 2013; Wu and Li, 2019). Previously, Wang et al. (2015) adapted LDA to discover long-tail semantics from large-scale corpora. Those models usually adopt Gibbs Sampling

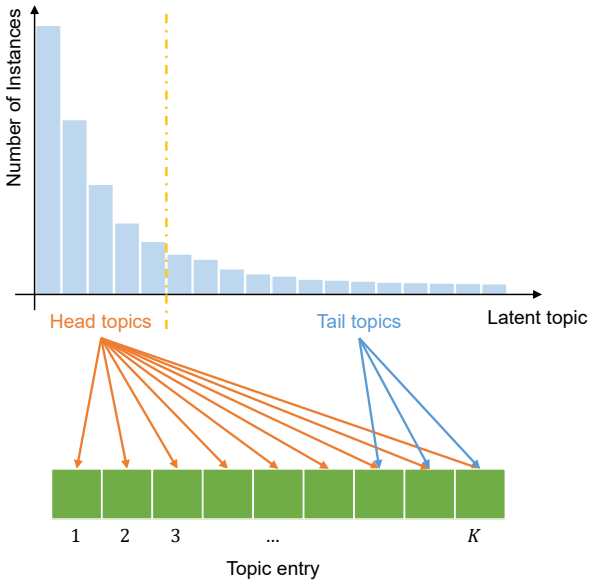

Figure 1: Illustration of topic entries assigned to documents in a long-tailed corpus.

(Griffiths and Steyvers, 2004) or Variational Inference (Blei et al., 2017) for parameter estimations. Based on VAE (Kingma and Welling, 2014; Rezende et al., 2014), neural topic models (Miao et al., 2016; Srivastava and Sutton, 2017; Wu et al., 2020a) are introduced. They are derivation-free and can apply gradient backpropagation directly. Nevertheless, these former works including probabilistic and neural methods are normally evaluated on balanced datasets. Since long-tail distributed data are common in our natural world (Reed, 2001), this inspires us to find out how these topic models perform on long-tailed corpora and propose useful ways to alleviate the long-tailed bias impact.

Causal Inference Causal inference (Pearl et al., 2016) has been widely adopted in various fields for years, like psychology, epidemiology, and medicine (MacKinnon et al., 2007; Richiardi et al., 2013), providing solutions to investigate the causation between research objects. Recently, the causal inference has also increasingly attracted attention in computer vision and NLP society for removing the biases in datasets (Tang et al., 2020; Wu et al., 2020c) or providing counterfactual examples (Zeng et al., 2020) in domain-specific applications. In this paper, we propose to employ the causal inference mechanism to investigate wherefores for the issues of long-tailed topic modeling and propose a solution with deconfounded training via the intervention to alleviate the bias effect.

3 Method

In this section, we first explain how the long-tailed bias affects topic modeling from the perspective of causal inference, and then propose a novel model to overcome this issue with deconfounded training by the causal intervention.

3.1 SCM for Topic Modeling

First of all, we investigate the causal relationship between the latent variables in a topic model with a Structural Causal Model (SCM). SCMs are expressed visually by using directed acyclic graphs. In the graph, vertices are random variables, and directed edges represent direct causation from one variable to another (Pearl et al., 2016). There is a special vertex in the graph: *confounder*, a variable that influences both correlated and independent variables, creating a spurious statistical correlation. For example, considering an interesting study that chocolate consumption is statistically related to the number of Nobel prizes of a country (Dablander, 2020). Is it justified to argue that people can get Nobel prizes if they eat more chocolate? Common sense intuitively tells us this assertion is inaccurate. We can draw a causal graph to detail it: **chocolate consumption** $\leftarrow$ **economy** $\rightarrow$ **number of Nobel prizes**, the chocolate consumption is usually higher in a developed country with good economy, and the number of Nobel prizes is also larger since the citizens' education level is higher in this country. Therefore, the economy acts as a confounder that creates a spurious correlation between chocolate consumption and the number of Nobel prizes.

Similar to the above example, we build a SCM shown in Figure 2a to describe how a biased corpus influences the text generation process of topic modeling. In the graph, $\boldsymbol{C}$ means the unobserved confounding bias in a long-tailed corpus. We note the vocabulary size is V and set K topic entries (the topic number is K) which means the model needs to discover K latent topics. In the setting of topic modeling, a topic entry k is interpreted as the related words and represented with a word distribution $\boldsymbol{\beta}_k \in \mathbb{R}^V$. Then, the word distributions of all topic entries (topic-word distribution matrix) is $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_k, ..., \boldsymbol{\beta}_K) \in \mathbb{R}^{V \times K}$. A document $\boldsymbol{x}$ is assigned with various topic entries with each probability as θ_k, so the distribution over all topic entries (topic distribution of $\boldsymbol{x}$) is $\boldsymbol{\theta} = (\theta_1, ..., \theta_k, ..., \theta_K)^{\mathrm{T}} \in \mathbb{R}^K$. Then, $\boldsymbol{x}$ is generated with its topic distribution $\boldsymbol{\theta}$ and the topic-word distribution matrix $\boldsymbol{\beta}$ of the whole corpus. The paths in Figure 2a can be specifically interpreted as follows:

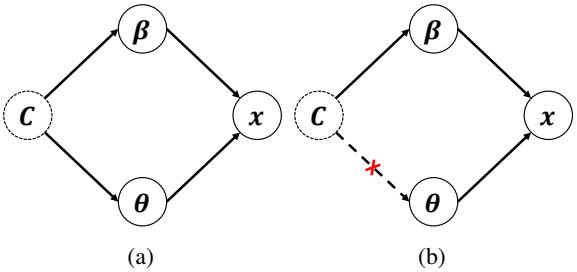

Figure 2: The structural causal model of topic modeling. (a) Complete SCM without interventions. (b) Do intervention on the topic distribution $\boldsymbol{\theta}$.

- $\boldsymbol{C} \rightarrow \boldsymbol{\theta}$: This path says that the topic distributions are trained under bias. If there is no bias, different topic entries are ideally assigned to documents about various topics, and the inferred topic distributions of these documents are also different. However, in a long-tailed corpus with bias, the topic distributions of documents about different topics could be similar. As shown in Figure 1, since documents about the head topics are the absolute majority, most of the topic entries are assigned to them[2]. In this case, for a document about tail topics, its assigned topic entries probably are also assigned to the documents about head topics, as a result of which, its inferred topic distribution becomes similar to the topic distributions of some documents about head topics.

- $\boldsymbol{C} \rightarrow \boldsymbol{\beta} \rightarrow \boldsymbol{x}$: This link denotes the topic-word distribution matrix $\boldsymbol{\beta}$ is trained under the bias and is used to generate the document $\boldsymbol{x}$. Due to the long-tailed bias, the generated $\boldsymbol{x}$ tends to contain words in the documents about head topics.

- $\boldsymbol{\theta} \leftarrow \boldsymbol{C} \rightarrow \boldsymbol{\beta} \rightarrow \boldsymbol{x}$: Because of the confounder $\boldsymbol{C}$, the inferred topic distribution of a document about tail topics could be similar to the topic distributions of some documents about head topics, and the generated documents of these similar topic distributions tend to include words in the documents about head topics instead of tail topics. Therefore, this

[2]If the documents about tail topics are few enough to be ignorable, all the topic entries will be assigned to the documents about head topics.

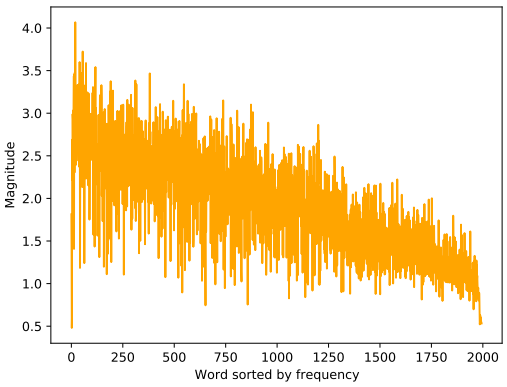

Figure 3: The magnitude of $\boldsymbol{\beta}_{i*}$ obtained from ProdLDA sorted by word frequency.

> backdoor path via $\boldsymbol{C}$ causes the spurious correlation between the topic distribution of a document about tail topics and the words in the documents about head topics.

In consequence, this spurious correlation through the confounder incurs that the discovered topics from documents about tail topics are mixed by the words of latent head topics. When the bias in the corpus gets severer, the discovered topics are even totally occupied by these words. Namely, topic models tend to ignore the semantics of the documents about tail topics and cannot discover the latent tail topics of a corpus.

In the above discussion, we clarify how the bias leads to the problems of long-tailed topic modeling with the presented SCM. In the next section, we propose a neat method to solve this issue without any auxiliary information.

3.2 Intervention on Topic Distribution

To remove the spurious correlation (deconfound), we propose to do causal intervention via do-operator (Pearl et al., 2016). Taking the chocolate and Nobel prizes for example again, intervening on the chocolate consumption means we fix its value through which we curtail the natural tendency of it to vary in response to the economy in nature. This amounts to remove the edges directed into the chocolate consumption. For example, if we were to close all chocolate factories, denoted as $do(\textbf{chocolate consumption} = 0)$, we will find the causality between the chocolate consumption and the number of Nobel prizes.

Similarly, we do intervention on the topic distribution $\boldsymbol{\theta}$ to compute the causality of $\boldsymbol{\theta}$ on $\boldsymbol{x}$, i.e., $p(\boldsymbol{x}|do(\boldsymbol{\theta}))$. As shown in Figure 2b, doing intervention on $\boldsymbol{\theta}$ means cutting off the edge $\boldsymbol{C} \rightarrow \boldsymbol{\theta}$ so that $\boldsymbol{C}$ cannot affect $\boldsymbol{\theta}$. But it is difficult to actually intervene variables (like closing all chocolate factories), so we utilize the backdoor adjustment (Pearl, 1995). The variable $\boldsymbol{\beta}$ meets the backdoor criterion and blocks the backdoor path $\boldsymbol{\theta} \leftarrow \boldsymbol{C} \rightarrow \boldsymbol{\beta} \rightarrow \boldsymbol{x}$. Following the backdoor adjustment, we use Inverse probability Weighting (Pearl et al., 2016) as

$$
\begin{aligned}
& p(\boldsymbol{x}|do(\boldsymbol{\theta})) \\
&= \sum_{\boldsymbol{\beta}} p(\boldsymbol{x}|do(\boldsymbol{\theta}), \boldsymbol{\beta}) p(\boldsymbol{\beta}|do(\boldsymbol{\theta})) && (1) \\
&= \sum_{\boldsymbol{\beta}} p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{\beta}) p(\boldsymbol{\beta}) && (2) \\
&= \sum_{\boldsymbol{\beta}} \frac{p(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\beta})}{p(\boldsymbol{\theta}|\boldsymbol{\beta})} && (3)
\end{aligned}
$$

In Figure 2b, all of $\boldsymbol{\theta}$ and $\boldsymbol{x}$ association flows along the directed path from $\boldsymbol{\theta}$ to $\boldsymbol{x}$ since there cannot be any backdoor paths because $\boldsymbol{\theta}$ has no incoming edges, so $p(\boldsymbol{x}|do(\boldsymbol{\theta}), \boldsymbol{\beta}) = p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{\beta})$. Also, $p(\boldsymbol{\beta}|do(\boldsymbol{\theta})) = p(\boldsymbol{\beta}|\boldsymbol{\theta})$ since there's no other edges from $\boldsymbol{\theta}$ to $\boldsymbol{\beta}$ except through the collider $\boldsymbol{x}$.

But this equation is intractable, we need to approximate it. To find a proper way, we bury in mind that topics are interpreted as word distributions, so long-tail distributed topics can also be seen as long-tail distributed words. If we treat these words as "labels", then the generative process of a document is roughly predicting the probability under each "label". This inspires us to discover the relation between long-tailed topics and long-tailed classification tasks (Kang et al., 2020; Tang et al., 2020). Similar to these tasks, as shown in Figure 3, we observe that the magnitudes of topic distributions of words, i.e., $\boldsymbol{\beta}_{i*}$ for word i, gradually decrease along with the word frequency. Intuitively, the magnitude of $\boldsymbol{\beta}_{i*}$ means the "correlation score" between word i and all topics; therefore, this phenomenon may be because most inferred topics tend to relate to the words in documents about the head topics as mentioned before. Due to this finding, we propose an approximation method following the propensity score modeling (Rosenbaum and Rubin, 1983; Austin, 2011):

$$
p(\boldsymbol{x}|do(\boldsymbol{\theta})) \approx \prod_{i} \frac{\boldsymbol{\beta}_{i*}\boldsymbol{\theta}}{\|\boldsymbol{\beta}_{i*}\|_2 \|\boldsymbol{\theta}\|_2} \qquad (4)
$$

where i refers to a word in $\boldsymbol{x}$ and we also empirically add the magnitude of $\boldsymbol{\theta}$. Here, the denomina-

tor works as a normalizer that balances the magnitude of the variables: β_{i*} and $\boldsymbol{\theta}$ for approximating the intervention probability.

3.3 Proposed Model

In this section, we propose a neural topic model for long-tailed corpora with deconfounded training based on the aforementioned intervention method, named as **Deconfounded Topic Model** (**DecTM**). Our network architecture is under the basic framework of VAE (Kingma and Welling, 2014; Rezende et al., 2014) with an encoder and a deconfounded decoder.

3.3.1 Encoder

The encoder transforms a text $\boldsymbol{x}$ into its topic distribution $\boldsymbol{\theta}$. Following the setting of Miao et al. (2016), we take the bag-of-words (BoW) assumption that ignores the word orders since topic models normally leverage word co-occurrences for topic inference. Inputted the BoW representation of $\boldsymbol{x}$, we first obtain its intermediate representation $\boldsymbol{\pi}$ with a Multi-Layer Perceptron (MLP). Based on $\boldsymbol{\pi}$, we then compute $q(\boldsymbol{r}|\boldsymbol{x})$, the variational distribution of the latent representation $\boldsymbol{r}$. Since the prior distribution $p(\boldsymbol{r})$ is assumed to be a Logistic Normal distribution for approximating the Dirichlet distribution (Srivastava and Sutton, 2017), we model the $q(\boldsymbol{r}|\boldsymbol{x})$ as $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$ are calculated as

$$\boldsymbol{\mu} = \boldsymbol{W}_{\mu}\boldsymbol{\pi} + \boldsymbol{b}_{\mu} \tag{5}$$

$$\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{W}_{\Sigma}\boldsymbol{\pi} + \boldsymbol{b}_{\Sigma}) \tag{6}$$

where $\boldsymbol{W}_{\mu}$, $\boldsymbol{W}_{\Sigma}$, $\boldsymbol{b}_{\mu}$ and $\boldsymbol{b}_{\Sigma}$ are weight matrices and biases respectively , and $\mathrm{diag}(\cdot)$ means converting a vector to a diagonal matrix. Later, to reduce the gradient variance, we adopt the reparameterization trick (Kingma and Welling, 2014) to sample $\boldsymbol{r}$ as

$$\boldsymbol{r} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}). \tag{7}$$

Next, $\boldsymbol{r}$ is normalized with a softmax function to get the topic distribution $\boldsymbol{\theta}$ as

$$\boldsymbol{\theta} = \mathrm{softmax}(\boldsymbol{r}). \tag{8}$$

3.3.2 Deconfounded Decoder

After getting the topic distribution of the input text, we then feed it to the proposed deconfounded decoder for reconstruction. According to the method in Equation (4), the objective function of DecTM can be written as

$$\mathcal{L}(\boldsymbol{x}) = \mathrm{KL}\left(q(\boldsymbol{r}|\boldsymbol{x}) \| p(\boldsymbol{r})\right) - \mathbb{E}_{\epsilon\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left[\sum_{i=1}^{N} \log \frac{\boldsymbol{\beta}_{i*}\boldsymbol{\theta}}{\|\boldsymbol{\beta}_{i*}\|_2 \|\boldsymbol{\theta}\|_2}\right] \tag{9}$$

where the first term is the Kullback-Leibler (KL) divergence between the posterior and prior distribution. It can be computed with the analytical solution for two Normal distributions. The second term is the reconstruction error between the input and output text. Different from normal neural topic models (Miao et al., 2016; Srivastava and Sutton, 2017), the deconfounded decoder in our model employs the approximated probability for causal intervention on $\boldsymbol{\theta}$ to weaken the long-tailed bias. Note that our model can be directly applied to naturally collected corpora since no additional auxiliary information is necessary for our model.

4 Experiment Setup

4.1 Datasets

Unfortunately, common datasets for topic modeling are almost all balanced, so we manually construct the long-tailed variants of them by repeating and deleting documents according to the given labels, making them follow a long-tailed distribution. Through the distribution of labels, we can roughly assume the latent topics are long-tailed distributed. In this way, we form the long-tailed versions (-LT) of 20News (Lang, 1995)[3] and Yahoo Answer[4], called 20News-LT and Yahoo Answer-LT respectively. Moreover, to better evaluate the performance of long-tailed topic modeling, we adopt the datasets for eXtreme Multi-label Text Classification (XMTC) (You et al., 2019), a task to predict the most relevant multiple labels for texts from an extremely large-scale label set. The label set includes hundreds and thousands, even millions of labels, and most are tail labels with very few positive samples. These plentiful labels can be naturally interpreted as the latent topics of documents; thus, we can evaluate the proposed model on these long-tailed distributed datasets. We conducted experiments on the the subsets of standard benchmark XMTC datasets Amazon-670K, Wiki-500K, AmazonCat-13K and Amazon-3M (Bhatia

[3] http://qwone.com/~jason/20Newsgroups
[4] https://answers.yahoo.com

Datasets	#docs	Average length	#labels	Vocabulary size	Average #labels per doc	Average #docs per label
20News-LT	11,314	73.5	20	1,984	-	-
Yahoo Answer-LT	99,806	31.0	10	4,738	-	-
Wiki-500K	100,000	504.0	175,206	5,000	6.0	3.4
Amazon-670K	100,000	90.0	333,863	5,000	2.0	1.6
AmazonCat-13K	200,000	78.8	11,096	5,000	5.0	91.0
Amazon-3M	300,000	40.7	1,917,999	5,000	13.0	5.6

Table 1: Statistics of datasets.

et al., 2016)[5].

For all datasets, we conduct the following steps for preprocessing: (1) tokenize texts and lowercase words; (2) remove stop words and illegal characters; (3) remove low-frequency words. The statistics of preprocessed datasets are reported in Table 1. It is worth noting that although labels are provided in these datasets, they are not used by our model.

4.2 Baseline Models

We take both probabilistic and neural topic models as baselines. For probabilistic models, we consider the widely used LDA (Blei et al., 2003) with python-lda[6] package for topic inference. For neural topic models, we use NVDM (Miao et al., 2016)[7], ProdLDA (Srivastava and Sutton, 2017)[8] and Scholar (Card et al., 2018)[9]. Scholar is an extension of ProdLDA via optionally incorporating metadata of documents like sentiments.

5 Experiment Results

5.1 Topic Quality Analysis

5.1.1 Evaluation Metrics

Following Nan et al. (2019) and Wu et al. (2020b), we evaluate the topic quality concerning two aspects, topic coherence and diversity. Topic coherence means that the words in the discovered topics are supposed to be as coherent as possible instead of irrelevant ones, and topic diversity means that topics should differ from each other instead of being similar ones.

Topic Coherence For topic coherence, we employ C_V (Röder et al., 2015), an improved variant of the Normalized Pointwise Mutual Information (NPMI) (Bouma, 2009; Chang et al., 2009; Newman et al., 2010). Its detailed calculation can be found in Wu et al. (2020b). We need to mention that given a topic z and its top T probable words $(x_1, x_2, ..., x_T)$, the NPMI of (x_i, x_j) used in the C_V computation is defined as

$$\text{NPMI}(x_i, x_j) = \frac{\log \frac{p(x_i,x_j)+\epsilon}{p(x_i)p(x_j)}}{-\log(p(x_i, x_j) + \epsilon)} \quad (10)$$

where $p(x_i)$ is the occurrence probability of word x_i and $p(x_i, x_j)$ the co-occurrence probability of (x_i, x_j). These probabilities are estimated in a reference corpus. To exhaustively assess the topic coherence performance of long-tailed topic modeling, we use three kinds of C_V scores with the probabilities estimated in different reference corpora. First, we adopt the public tool[10] which uses Wikipedia documents as the external reference corpus (-E), so it is named as C_V-E. Then, we directly use the internal training documents (-I) as the reference corpus, named as C_V-I. However, since documents about head topics occupy the main portion of a long-tailed corpus, previous C_V-E and C_V-I probably are insufficient to appraise the performance on the documents about tail topics. To this end, we heuristically introduce C_V-T that employs the documents including the tail labels (-T) provided by the datasets instead of all the training documents as the reference corpus, so it can assess whether the discovered topics can reveal the hidden semantics of documents about tail topics, i.e., discover the tail topics.

Topic Diversity For topic diversity evaluation, we employ the Topic Unique (TU) (Nan et al.,

[5] http://manikvarma.org/downloads/XC/XMLRepository.html

[6] https://github.com/lda-project/lda

[7] https://github.com/ysmiao/nvdm

[8] https://github.com/akashgit/autoencoding_vi_for_topic_models

[9] https://github.com/dallascard/scholar

[10] https://github.com/dice-group/Palmetto

Models	20News-LT					Yahoo Answer-LT					Wiki-500K				
	TU	C_V-E	C_V-I	C_V-T	TQ	TU	C_V-E	C_V-I	C_V-T	TQ	TU	C_V-E	C_V-I	C_V-T	TQ
LDA	0.469	0.318	**0.712**	0.378	0.220	0.531	0.335	0.408	0.384	0.199	0.759	0.426	0.680	0.669	0.449
NVDM	0.849	0.315	0.367	0.492	0.332	0.934	0.352	0.440	0.549	0.417	0.836	0.397	0.429	0.372	0.334
ProdLDA	0.743	0.321	0.521	0.543	0.343	0.830	**0.383**	0.434	0.516	0.369	0.889	**0.438**	0.698	0.653	0.530
Scholar	0.787	0.321	0.537	0.528	0.364	0.876	0.380	0.445	0.519	0.392	0.874	0.437	0.682	0.633	0.510
DecTM	**0.937**	**0.324**	0.543	**0.554**	**0.443**	**0.948**	0.381	**0.516**	**0.573**	**0.464**	**0.979**	0.437	**0.754**	**0.716**	**0.622**

Models	Amazon-670K					AmazonCat-13K					Amazon-3M				
	TU	C_V-E	C_V-I	C_V-T	TQ	TU	C_V-E	C_V-I	C_V-T	TQ	TU	C_V-E	C_V-I	C_V-T	TQ
LDA	0.752	0.387	0.601	0.566	0.390	0.722	0.385	0.616	0.585	0.382	0.746	0.368	0.692	**0.624**	0.418
NVDM	0.873	0.391	0.338	0.366	0.319	0.881	0.403	0.392	0.405	0.352	0.889	0.433	0.425	0.487	0.399
ProdLDA	0.869	0.424	0.606	0.501	0.443	0.893	**0.430**	0.641	0.531	0.477	0.925	0.442	0.636	0.498	0.486
Scholar	0.874	**0.427**	0.607	0.502	0.449	0.886	0.428	0.644	0.540	0.476	0.908	**0.445**	0.614	0.508	0.474
DecTM	**0.987**	0.404	**0.672**	**0.571**	**0.542**	**0.991**	0.406	**0.702**	**0.590**	**0.561**	**0.991**	0.405	**0.701**	0.530	**0.541**

Table 2: Topic quality results concerning topic coherence and diversity. The best in each column is in bold.

2019) defined as

$$TU(z) = \frac{1}{T}\sum_{i=1}^{T}\frac{1}{\text{cnt}(x_i)} \quad (11)$$

where $\text{cnt}(x_i)$ is the total number of times that word x_i appears in the top T words of all topics. Accordingly, TU ranges from $1/K$ to 1, and a higher TU score means topics are more diverse since fewer words are repeated across all.

Comprehensive Evaluation It is necessary to mention that if the topic coherence performance of a model remains about the same and the diversity gets higher, it means the overall topic quality is also improved since it can unearth more various semantics of documents. To provide a forthright and comprehensive evaluation of both coherence and diversity performance, following Dieng et al. (2019), we propose Topic Quality (TQ) that combines C_V and TU as

$$TQ = TU \times \frac{1}{3}(C_V\text{-E} + C_V\text{-I} + C_V\text{-T}) \quad (12)$$

which is the product of TU and the average of three different C_V scores. Thus, TQ can provide a direct comparison of the overall topic quality performance.

5.1.2 Results Analysis

Table 2 reports the topic quality results concerning different metrics of the top 15 words with the topic number $K = 50$. At first, we notice that C_V-E scores of DecTM are the highest on 20News-LT and are very close to the best on Yahoo Answer-LT and Wiki-500K. Although C_V-E scores of some baseline models are higher on other datasets, DecTM stably outperforms them in terms of TU by a large margin, and the C_V-I scores of DecTM are also mostly better. This implies that baseline models are disposed to generate repetitive topics because of the bias of these long-tailed corpora, while the topics of our DecTM are more diverse. Therefore, despite that C_V-E of some baselines are higher, their lower diversity performance indicates that their yielded topics are redundant. More significantly, DecTM commonly surpasses baseline models in terms of C_V-T, which shows the

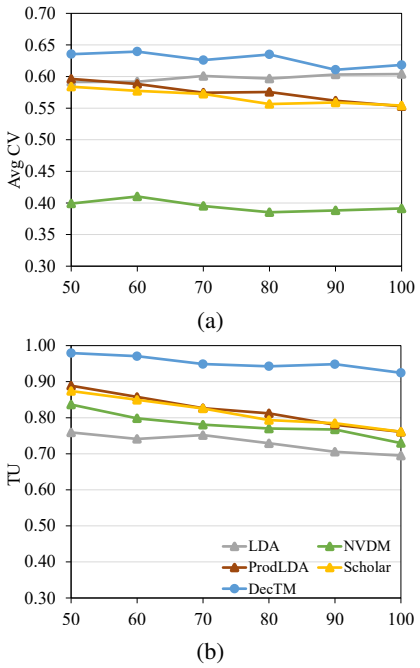

Figure 4: TU and average C_V scores on Wiki-500K under different topic numbers.

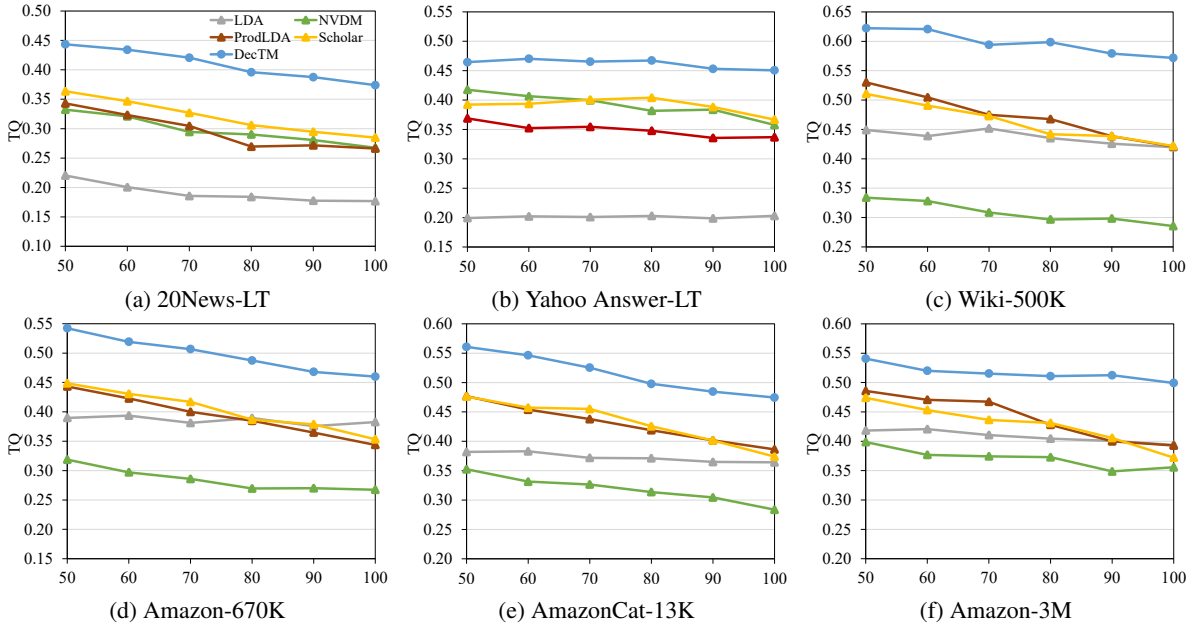

Figure 5: Topic quality performance (TQ) under different topic numbers.

discovered topics can preferably reflect the semantics of documents about tail topics; thus, the produced topics of DecTM are more complete. These arguments are further illustrated with topic examples in Section 5.3. At last, we find that DecTM achieves the highest TQ scores on all datasets, showing the overall performance of our model is fairly better.

From the above experimental results, we observe the problems of the long-tailed topic modeling formerly mentioned in Section 1 and Section 3.1, that the performance of existing topic models, especially topic diversity, is deteriorated on account of the bias. But with the help of the deconfounded decoder, our proposed method can alleviate the effect of the bias and hugely improve the topic diversity while remain good coherence performance with a better ability to expose the semantics of documents about tail topics.

5.2 Impact of Topic Number

To investigate how performance varies concerning the topic number, we report the TU and the average C_V (Avg C_V) scores defined in Equation (12) under topic number K ranging from 50 to 100 on Wiki-500K in Figure 4. We can see that the Avg C_V of DecTM is relatively better in Figure 4a. Besides, as shown in Figure 4b, TU scores of all models gradually decline when the topic number gets bigger, but the performance of DecTM is constantly higher and decreases slower. The reason is that those baseline models tend to focus on documents about head topics which are inadequate to support larger topic numbers, while DecTM can also discover semantics of documents about tail topics. Furthermore, Figure 5 presents the TQ with different topic numbers of all datasets. We can observe that whether on manually constructed or XMTC datasets, our model DecTM outperforms baseline models under different topic numbers. These experiments demonstrate that the performance of our model is relatively stable.

5.3 Discovered Topic Examples

To further illustrate the topic quality performance of different models, Table 3 reports some discovered topic examples. As shown by the comparison of topic diversity in Section 5.1.2, we can see baseline models produce some topics including repeated words. To be more specific, LDA generates topics with repetitive words like “subject’, “organization” from 20News-LT and “book”, “author” from Amazon-3M. Similar topics about “newcastle”, “orchestra” and “hockey” are yielded by NVDM, ProdLDA, and Scholar respectively. We also notice that NVDM, ProdLDA and Scholar all yield several same topics about “census” from Wiki-500K. These topics are coherent indeed, but they can trickily improve the C_V scores and are redundant in the downstream applications. On the contrary, we find only one coherent topic generated by DecTM corresponding to the aforementioned ones.

Models	Topic examples
LDA	hp nasa organization subject article new re subject organization access good pc support subject problem organization file world get scsi organization subject mark university ide thanks organization subject pt scott imagine life book god church author christian spiritual book students guide text chapter reading book story love read author stories characters books author lives writing new years book guide book design new using use techniques
NVDM	galaxy texas newcastle sky austin theta madrid edinburgh harbour newcastle fortress tunnel edwards leeds birmingham newcastle townships cdp islander nonfamilies couples females males husband nonfamilies
ProdLDA	orchestra hits songs symphony unreleased song concert orchestra opera biography symphony translation symphony subtitles orchestra mozart median capita nonfamilies residing household nonfamilies households householder residing residing township householder nonfamilies quot bmw yamaha honda macbook laptop
Scholar	guitarist pianist composer hockey player montreal nhl provincial provinces hockey championships finals medal olympics hockey householder nonfamilies households residing nonfamilies residing households householder township norway nonfamilies residing demographics median census hamlet town
DecTM	median residing nonfamilies household tires tire steering truck motorcycle honda wales welsh yorkshire scotland glasgow violin orchestra symphony concerto piano nonfiction copies manga novels bestseller championships mens olympic competed ink inkjet paper printer printers cartridges episodes episode season vol inspector series europe russian paris germany german spain

Table 3: Topic examples. Repeated words are underlined.

What is more, DecTM also discovers some latent topics like "printer", "series" and "europe" while these cannot be found by baseline models, which could verify the superior C_V-T performance of our model. These topic examples qualitatively show the overall topic quality performance of DecTM is adequately preferable.

6 Conclusion

In this paper, for discovering the topics in long-tailed corpora, we present a causal inference model to describe how the bias influences topic modeling, and to reduce the impact of the bias, we then propose a causal intervention method for deconfounding, relying on which we introduce the Deconfounded Topic Model (DecTM) with a deconfounded decoder. Comprehensive experiments demonstrate that our model can produce topics with better quality and mitigate the effect of long-tailed bias.

Acknowledgement

We want to thank all anonymous reviewers for their helpful comments. This work is supported by China NSFC under Grant 61672309 and MOST Fundamental Research Project under Grant 2017FY201407.

References

Peter C Austin. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424.

K. Bhatia, K. Dahiya, H. Jain, P. Kar, A. Mittal, Y. Prabhu, and M. Varma. 2016. The extreme classification repository: Multi-label datasets and code.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. 2017. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

David M Blei and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40.

Dallas Card, Chenhao Tan, and Noah A Smith. 2018. Neural Models for Documents with Metadata. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2031–2040.

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

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on

effective number of samples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9268–9277.

Fabian Dablander. 2020. An introduction to causal inference. *PsyArXiv*.

Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2019. The dynamic embedded topic model. *arXiv preprint arXiv:1907.05545*.

Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235.

Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. 2020. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*.

Salman H Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A Sohel, and Roberto Togneri. 2017. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems*, 29(8):3573–3587.

Diederik P Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *The International Conference on Learning Representations (ICLR)*.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 331–339.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Zongyang Ma, Aixin Sun, Quan Yuan, and Gao Cong. 2012. Topic-driven reader comments summarization. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 265–274. ACM.

David P MacKinnon, Amanda J Fairchild, and Matthew S Fritz. 2007. Mediation analysis. *Annu. Rev. Psychol.*, 58:593–614.

Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. 2018. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 181–196.

Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172.

Jon Mcauliffe and David Blei. 2007. Supervised topic models. *Advances in neural information processing systems*, 20:121–128.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *International conference on machine learning*, pages 1727–1736.

Feng Nan, Ran Ding, Ramesh Nallapati, and Bing Xiang. 2019. Topic modeling with Wasserstein autoencoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6345–6381, Florence, Italy. Association for Computational Linguistics.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108. Association for Computational Linguistics.

Judea Pearl. 1995. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic books.

William J Reed. 2001. The pareto, zipf and other power laws. *Economics letters*, 74(1):15–19.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. *In Proceedings ofthe 31th International Conference on Machine Learning*.

Lorenzo Richiardi, Rino Bellocco, and Daniela Zugna. 2013. Mediation analysis in epidemiology: methods, interpretation and bias. *International journal of epidemiology*, 42(5):1511–1519.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408. ACM.

Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *ICLR*.

Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. 2020. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *NeurIPS*.

Xuerui Wang, Andrew McCallum, and Xing Wei. 2007. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Seventh IEEE international conference on data mining (ICDM 2007)*, pages 697–702. IEEE.

Yi Wang, Xuemin Zhao, Zhenlong Sun, Hao Yan, Lifeng Wang, Zhihui Jin, Liubin Wang, Yang Gao, Ching Law, and Jia Zeng. 2015. Peacock: Learning long-tail topic features for industrial applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6(4):1–23.

Xiaobao Wu and Chunping Li. 2019. Short Text Topic Modeling with Flexible Word Patterns. In *International Joint Conference on Neural Networks*.

Xiaobao Wu, Chunping Li, Yan Zhu, and Yishu Miao. 2020a. Learning Multilingual Topics with Neural Variational Inference. In *International Conference on Natural Language Processing and Chinese Computing*.

Xiaobao Wu, Chunping Li, Yan Zhu, and Yishu Miao. 2020b. Short text topic modeling with topic distribution quantization and negative sampling decoder. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1772–1782, Online.

Yiquan Wu, Kun Kuang, Yating Zhang, Xiaozhong Liu, Changlong Sun, Jun Xiao, Yueting Zhuang, Luo Si, and Fei Wu. 2020c. De-biased court's view generation with causality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 763–780.

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1445–1456. ACM.

Ronghui You, Zihan Zhang, Ziye Wang, Suyang Dai, Hiroshi Mamitsuka, and Shanfeng Zhu. 2019. Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. In *Advances in Neural Information Processing Systems*, pages 5820–5830.

Xiangji Zeng, Yunliang Li, Yuchen Zhai, and Yin Zhang. 2020. Counterfactual generator: A weakly-supervised method for named entity recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7270–7280.

Ziqi Zhang and Lei Luo. 2019. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, 10(5):925–945.

Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. 2020. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9719–9728.