FoTo: Targeted Visual Topic Modeling for Focused Analysis of Short Texts

Sanuj Kumar, Tuan M. V. Le
New Mexico State University
Las Cruces, USA
{sanujkr, tuanle}@nmsu.edu

Abstract

Given a corpus of documents, focused analysis aims to find topics relevant to aspects that a user is interested in. The aspects are often expressed by a set of keywords provided by the user. Short texts such as microblogs and tweets pose several challenges to this task because the sparsity of word co-occurrences may hinder the extraction of meaningful and relevant topics. Moreover, most of the existing topic models perform a full corpus analysis that treats all topics equally, which may make the learned topics not be on target. In this paper, we propose a novel targeted topic model for semantic short-text embedding which aims to learn all topics and low-dimensional visual representations of documents, while preserving relevant topics for focused analysis of short texts. To preserve the relevant topics in the visualization space, we propose jointly modeling topics and the pairwise document ranking based on document-keyword distances in the visualization space. The extensive experiments on several real-world datasets demonstrate the effectiveness of our proposed model in terms of targeted topic modeling and visualization.

Keywords: focused analysis, targeted topic models, visualization, short texts

1. Introduction

Document visualization and topic modeling are widely used in text analysis. Document visualization based on dimensionality reduction aims to embed documents into 2- or 3-dimensional space for visualization (Van der Maaten and Hinton, 2008; Tang et al., 2016; McInnes et al., 2018). Meanwhile, topic modeling's objective is to discover latent topics discussed in the documents (Blei et al., 2003; Grootendorst, 2022). Recently, joint models are proposed to learn both topics and visualization such that the learned embeddings reflect the document similarities based on the underlying topics (Iwata et al., 2008; Pham and Le, 2020).

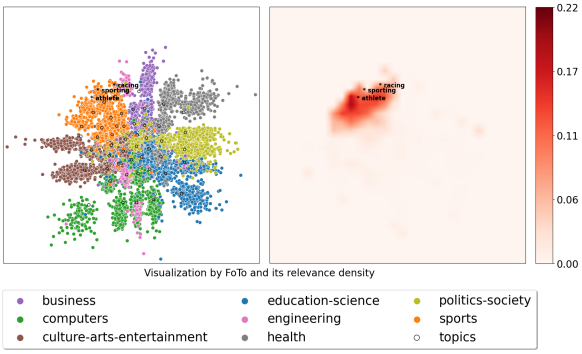

Figure 1: Visual focused analysis of short texts from SEARCHSNIPPET for query {sporting, athlete, racing}; (left) visualization of SEARCHSNIPPET by *FoTo*; (right) the corresponding document relevance density as per the aspects; black asterisks indicate keyword locations.

The above methods may suffer the following limitations when applied to short texts: 1) When documents are short in length (e.g., tweets or microblogs may contain less than dozens of words), these models may not perform well due to the sparsity of word co-occurrences; 2) Most of the existing topic models learn all topics by performing a full analysis on a large text corpus. Therefore, the extracted topics may be too coarse and not relevant to the aspects that users are interested in. Moreover, if the aspects of interest are relatively unpopular in the data, relevant topics may not be found because they may be overshadowed by other more prevalent ones in the corpus. Short texts make the relevant topics harder to extract because of the word sparsity issue; and 3) For joint visual topic models, if the focused topics are not prevalent, the models may not faithfully display the relevant topics or documents because the coordinates and quality of focused topics could be sacrificed when learning due to the information loss of the dimensionality reduction process.

Several short-text topic models can address the first limitation (Zuo et al., 2016; Wang et al., 2018; Dieng et al., 2020). However, these models suffer the second issue where users are not able to indicate the aspects of interest that should be focused on when training these topic models. Recently, targeted topic models are proposed to deal with this limitation (Wang et al., 2016; He et al., 2020; Wang et al., 2021). These models aim to find all topics relevant to interested aspects where each aspect is often expressed by a set of keywords provided by the user. However, for short texts, the

LREC-COLING 2024, pages 7406–7416
20-25 May, 2024.

issue still remains. If the user-provided list of keywords is not comprehensive enough, the extracted topics could be not relevant because the sparse word co-occurrences makes it challenging to infer and include the words similar to the provided keywords in the topics. Moreover, to visualize documents, the above targeted topic models need to perform an additional step to reduce dimension of the learned document-topic distributions. This two-step process does not ensure that the visualization will faithfully display the relevant topics and documents. In contrast, visual topic models for short texts can jointly produce visualization but they are not targeted topic models (Kumar and Le, 2021). In another direction, given a set of seed words, seed-guided topic models aim to find a topic for each seed word (Harandizadeh et al., 2022; Zhang et al., 2023). Therefore, the number of extracted topics is equal to the number of seed words. This is different from targeted topic models which aim to find all topics relevant to the given keywords.

To simultaneously address the above limitations, we propose a novel targeted visual topic model, named *FoTo*, that can extract and visualize topics, documents relevant to targeted aspects for focused analysis. Our proposed model leverages word embeddings to alleviate the data sparsity. Different from previous methods, our method tightly integrates topics, embeddings of documents and topics, word embeddings, targeted aspects, as well as the pairwise document ranking in a holistic model. The generated focused visualization shows not only documents, topics, but also user-provided keywords in the same visualization space. To ensure that the relevant topics and documents are well-preserved, we propose modeling the pairwise document ranking based on the distances between documents and keywords in the visualization.

As an example, Figure 1 shows the novel aspects of *FoTo*. We perform targeted visual topic modeling of short texts from SEARCHSNIPPET for query {sporting, athlete, racing}. *FoTo* is a joint targeted visual topic model that visualizes topics (hollow circles), keywords (black asterisks), and documents (color circles) while extracting and preserving relevant topics. In the visualization, users can see a comprehensive view of the corpus by considering the locations of topics and documents. Moreover, users can see how relevant topics and documents are distributed by viewing the relative locations of points near the keywords in the visualization. To show how relevant documents are distributed, we calculate the TF-IDF scores of documents w.r.t the query and estimate the relevance density by averaging the relevant scores of documents in a region[1]. The result is shown in the right figure. As we can see, most of the relevant documents are located near the keywords, indicating that our method preserves well the documents and topics of interest. We summarize our contributions as follows:

1. We propose a novel targeted visual topic model, named *FoTo*[2], for extracting and visualizing topics, documents that are relevant to targeted aspects.
2. To integrate targeted aspects, we propose modeling the pairwise document ranking based on distances between documents and keywords in the visualization space. We derive a stochastic variational inference algorithm for our proposed model.
3. We conduct extensive experiments on several real-world short-text datasets. The results show that *FoTo* consistently generates better focused topics and visualization, as compared to state-of-the-art models.

2. Targeted Visual Topic Modeling

2.1. Problem Definition

The input to the model are a corpus of N documents $\mathcal{W} = \{\boldsymbol{w}_1, ..., \boldsymbol{w}_N\}$ and a set of S targeted aspects $\mathcal{S} = \{a_1, ..., a_S\}$ that a user wants to find topics related to. Each aspect a_s is represented as a set of keywords. A document w_i is represented as a bag of words. The main objectives are:

- The model finds all topics in the corpus, with a focus on topics that are related to the given targeted aspects. The model will return word distributions of Z topics $\{\beta_z\}_{z=1}^{Z}$ and topic distributions $\{\theta_i\}_{i=1}^{N}$ of documents.
- For visualization, we learn for each document i a coordinate x_i, and for each topic z a coordinate ϕ_z. Here $x_i, \phi_z \in \mathbb{R}^D$, $D = 2$ or 3. Each aspect s will be also assigned a coordinate π_s in the visualization. The distances between documents, aspects and between topics, aspects in the visualization reflect how relevant documents and topics are to aspects. The closer documents and topics are to aspects, the more relevant they are.

2.2. Integrating Topics and Visualization

FoTo utilizes word embeddings as supplementary information for dealing with the sparsity in short texts. Let $\mathcal{V}$ be a finite vocabulary from documents and $V = |\mathcal{V}|$ is the size of the vocabulary. Each word $v \in \mathcal{V}$ is represented by an embedding vector $\omega_v \in \mathbb{R}^E$, where E is the dimensionality[3] of the word embedding vector. Let $\Omega \in \mathbb{R}^{E \times V}$ be the word embedding matrix where its column v is the

[1] Bilinear interpolation is used.

[2] https://github.com/sanujsriv/FoTo

[3] $E = 300$ in our experiments.

word embedding ω_v of the word v in the vocabulary. Let $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ be the document coordinate matrix where the row i is the visualization coordinate x_i of document i, and $\boldsymbol{\Phi} \in \mathbb{R}^{Z \times D}$ be the topic coordinate matrix where the row z is the visualization coordinate ϕ_z of topic z. Given word and topic embeddings, the topic word distribution β_z is modeled as a log-linear model as follows:

$$\beta_z = \text{softmax}(\Omega^\top \tau_z + b) \tag{1}$$

here τ_z is the embedding vector of topic z, and b is the bias term. For visualization, we reduce the dimension of τ_z by mapping it to the visualization space using a feed-forward neural network: $\phi_z = f(\tau_z)$. The architecture of f is shown in the supplementary material. Given document and topic coordinates, the topic distribution θ_i of a document i is defined using a softmax function over its negative Euclidean distances to all topics:

$$\theta_{iz} = p\left(z|x_i, \boldsymbol{\Phi}\right) = \frac{\exp\left(-\frac{1}{2}\left\|x_i - \phi_z\right\|^2\right)}{\sum_{z'=1}^{Z} \exp\left(-\frac{1}{2}\left\|x_i - \phi_{z'}\right\|^2\right)} \tag{2}$$

Intuitively, the closer the document is to a topic, the higher probability that it is about that topic. To visualize aspects, for each aspect s, we treat its set of keywords a_s as a pseudo document. Besides the keywords provided by users for each aspect s, to improve the relevance of extracted topics, we extend a_s to include more similar keywords that can be found in the vocabulary. More specifically, we add to a_s the top words in the vocabulary that have the highest maximum cosine similarities to keywords in a_s. In the experiments, we extend the a_s until its length is equal to the average document length of the corpus. Let a_s^* be the extended a_s. Since each aspect is now a pseudo document, we can learn its visualization coordinate π_s in the same way as for normal documents.

2.3. Integrating Targeted Aspects via Pairwise Document Ranking

To model the pairwise document ranking in the visualization w.r.t each aspect, we define the probability that an aspect s is more relevant to document i than to document j as:

$$p(i >_s j|x_i, x_j, \pi_s) = \sigma\left(\frac{\exp\left(-\frac{1}{2}\left\|x_i - \pi_s\right\|^2\right)}{\exp\left(-\frac{1}{2}\left\|x_j - \pi_s\right\|^2\right)}\right) \tag{3}$$

here $i >_s j$ indicates that i is ranked higher than j in terms of relevance to aspect s. When document x_i is closer to aspect π_s than x_j, the probability that an aspect s is more relevant to document i than to document j is higher.

We treat the ranking order as observed data which can help improve the relevance of extracted topics. In general, we can pass any available ranking order of documents, either partial or complete, to the model. If not passed by the user, the model will infer the ranking order based on how many keywords there are in the documents. A document that contains more keywords should be ranked higher than documents containing no or less keywords. This assumption is very reasonable for short texts. We note that our method does not explicitly model the order between documents that do not contain any keywords (i.e., the order is partial), but let the document similarities inferred by the topic model fill in the gaps. When working with the extended a_s, the added keywords should not have the same effects to the ranking order as original keywords in aspect s. Therefore, we weight the count of an extended keyword appearing in the documents by its cosine similarity to the original keyword. More specifically, we compute the weighted sum of counts of keywords appearing in each document i as in Eq. 4 and use that to approximate the ranking order of documents.

$$c_{si} = \sum_{u \in a_s^*} \max_{v \in a_s} \left(\text{cosine}(\omega_u, \omega_v)\right) * \text{count}(u, \boldsymbol{w}_i) \tag{4}$$

Based on the above assumption, if c_{si} is greater than c_{sj} then we have $i >_s j$. Documents do not contain any keywords will have $c_{si} = 0$ and hence there are no explicit order between them.

2.4. Generation

Putting everything together, we propose the following generative process to integrate targeted aspects, topics, and visualization:

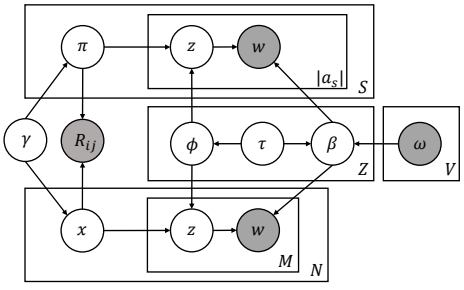

Figure 2: The graphical model of *FoTo*.

1. For each topic $z \in \{1, 2, \ldots, Z\}$
   (a) Obtain its coordinate $\phi_z = f(\tau_z)$
   (b) Obtain its word distribution: $\beta_z = \text{softmax}(\Omega^\top \tau_z + b)$
2. For each document $n \in \{1, \ldots, N\}$
   (a) Draw a document coordinate $\boldsymbol{x}_n \sim \text{Normal}\left(\boldsymbol{0}, \gamma \boldsymbol{I}\right)$
   (b) For each word $m \in \{1, 2, \cdots, M_n\}$
      i. Draw a topic $z_{nm} \sim \text{Multinomial}\left(\{P\left(z|\boldsymbol{x}_n, \boldsymbol{\Phi}\right)\}_{z=1}^{Z}\right)$

ii. Draw a $w_{nm} \sim \text{Multinomial}\left(\boldsymbol{\beta}_z\right)$

3. For each aspect s as a pseudo document:
   (a) Draw its coordinate $\pi_s \sim \text{Normal}\left(\mathbf{0}, \gamma \boldsymbol{I}\right)$
   (b) For each word $m \in \{1, 2, \ldots, |a_s^*|\}$
      i. Draw a topic $z_{sm} \sim \text{Multinomial}\left(\{P\left(z|\boldsymbol{\pi}_s, \boldsymbol{\Phi}\right)\}_{z=1}^{Z}\right)$
      ii. Draw a $w_{sm} \sim \text{Multinomial}\left(\boldsymbol{\beta}_z\right)$
   (c) For each document pair (i, j), where $i \neq j$, $1 \leq i, j \leq N$:
      i. Draw R_{sij} that indicates whether s is more relevant to document i than to document j: $R_{sij} \sim \text{Bernoulli}(p(i >_s j|x_i, x_j, \pi_s))$

The corresponding graphical model is shown in Figure 2. In this generative process, $P\left(z|\boldsymbol{x}_n, \boldsymbol{\Phi}\right)$ is the topic distribution and is computed as in Eq. 2. In step 3c(i), $p(i >_s j|x_i, x_j, \pi_s)$ is the probability that an aspect s is more relevant to document i than to document j and is computed as in Eq. 3.

2.5. Variational Inference

We estimate the parameters using the variational inference approach (Kingma and Welling, 2014). The parameters that need to be estimated are $\boldsymbol{\Psi} = \langle \boldsymbol{X}, \boldsymbol{\Phi}, \boldsymbol{\mathcal{S}}, \boldsymbol{\beta}, \boldsymbol{\tau} \rangle$. Here, $\boldsymbol{X}, \boldsymbol{\Phi}, \boldsymbol{\mathcal{S}}$ are document, topic, and aspect coordinates respectively. $\boldsymbol{\beta} \in \mathbb{R}^{Z \times V}$ is the topic-word probability matrix where the row z is the word distribution β_z, and $\boldsymbol{\tau} \in \mathbb{R}^{Z \times E}$ is the topic embedding matrix where its row z is the embedding vector τ_z. A document i is represented as a row vector of word counts: $\boldsymbol{w}_i \in \mathbb{Z}_{\geq}^{|\mathcal{V}|}$ and $\boldsymbol{w}_i^v$ is the number of occurrences of word $v \in \mathcal{V}$ in the document. The marginal likelihood of a document $\boldsymbol{w}_i$ and its aspect ranking order given coordinates of other documents $\boldsymbol{X}_{-x_i}$ and aspect coordinates $\boldsymbol{\mathcal{S}}$ is given by:

$$
\begin{aligned}
&p(\boldsymbol{w}_i, \mathbf{R}_i|\boldsymbol{X}_{-x_i}, \Phi, \boldsymbol{\beta}, \boldsymbol{\mathcal{S}}, \gamma) = \\
&\int_{x_i} p(\boldsymbol{w}_i|x_i, \Phi, \boldsymbol{\beta}, \gamma) p(\mathbf{R}_i|x_i, \boldsymbol{X}_{-x_i}, \boldsymbol{\mathcal{S}})) p(x_i|\gamma) dx_i \\
&= \int_{x_i} \left(\prod_{v=1}^{V} \left(\sum_{z=1}^{Z} p(v|z, \boldsymbol{\beta}) p\left(z|x_i, \boldsymbol{\Phi}\right) \right)^{\boldsymbol{w}_i^v} \right) \\
&\left(\prod_{s}^{S} \prod_{j \neq i, R_{sij}=1}^{N} p(R_{sij}|x_i, x_j, \boldsymbol{\mathcal{S}}) \right) p(x_i|\gamma) dx_i
\end{aligned}
\tag{5}
$$

here $\boldsymbol{\beta}$, $p\left(z|x, \boldsymbol{\Phi}\right)$ are computed using Eq. 1 and Eq. 2 respectively. Since each aspect s is considered as a pseudo document, the marginal likelihood of $\boldsymbol{w}_s$ is computed as follows:

$$
\begin{aligned}
&p(\boldsymbol{w}_s|\Phi, \boldsymbol{\beta}, \gamma) = \\
&\int_{s} \left(\prod_{v=1}^{V} \left(\sum_{z=1}^{Z} p(v|z, \boldsymbol{\beta}) p\left(z|\pi_s, \boldsymbol{\Phi}\right) \right)^{\boldsymbol{w}_s^v} \right) p(\pi_s|\gamma) d\pi_s
\end{aligned}
\tag{6}
$$

Since pseudo documents are aspects, they do not have ranking orders modeled in Eq. 6. To estimate the model parameters, we maximize the marginal log-likelihood above for each document and aspect. We have the following lower bound to the marginal log-likelihood (ELBO) of a document $\boldsymbol{w}_i$:

$$
\begin{aligned}
&\mathcal{L}\left(\eta|\boldsymbol{X}_{-x_i}, \Phi, \boldsymbol{\beta}, \boldsymbol{\mathcal{S}}, \gamma\right) = \\
&- \mathbb{D}_{\text{KL}}\left[q(x_i|\boldsymbol{w}_i, \eta) \| p(x_i|\gamma)\right] \\
&+ \mathbb{E}_{q(x_i|\boldsymbol{w}_i, \eta)}\left[\log p\left(\boldsymbol{w}_i|x_i, \boldsymbol{\Phi}, \boldsymbol{\beta}\right)\right] \\
&+ \mathbb{E}_{q(x_i|\boldsymbol{w}_i, \eta)}\left[\log p(\mathbf{R}_i|x_i, \boldsymbol{X}_{-x_i}, \boldsymbol{\mathcal{S}}))\right]
\end{aligned}
\tag{7}
$$

Similarly, the lower bound to the marginal log-likelihood (ELBO) of the pseudo document $\boldsymbol{w}_s$:

$$
\begin{aligned}
&\mathcal{L}_s\left(\eta|\gamma, \boldsymbol{\Phi}, \boldsymbol{\beta}\right) = \\
&- \mathbb{D}_{\text{KL}}\left[q(\pi_s|\boldsymbol{w}_s, \eta) \| p(\pi_s|\gamma)\right] \\
&+ \mathbb{E}_{q(\pi_s|\boldsymbol{w}_s, \eta)}\left[\log p\left(\boldsymbol{w}_s|\pi_s, \boldsymbol{\Phi}, \boldsymbol{\beta}\right)\right]
\end{aligned}
\tag{8}
$$

where $p(x|\gamma) = \text{Normal}\left(\mathbf{0}, \gamma \boldsymbol{I}\right)$ is the prior distribution of document coordinate x, $q(x|\boldsymbol{w}, \eta) = \text{Normal}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ is the variational distribution and $\boldsymbol{\mu}$, diagonal $\boldsymbol{\Sigma} \in \mathbb{R}^D$ are outputs of the encoding feed-forward neural network with variational parameters η. The KL divergence between two Gaussians in Eqs. 7 and 8 can be computed in a closed form as follows:

$$
\begin{aligned}
&\mathbb{D}_{\text{KL}}\left[q(x|\boldsymbol{w}, \eta) \| p(x|\gamma)\right] = \frac{1}{2}\Bigg(\text{tr}\left((\gamma \boldsymbol{I})^{-1} \boldsymbol{\Sigma}\right) \\
&+ (-\boldsymbol{\mu})^{\top} (\gamma \boldsymbol{I})^{-1} (-\boldsymbol{\mu}) - D + \log \frac{|\gamma \boldsymbol{I}|}{|\boldsymbol{\Sigma}|} \Bigg)
\end{aligned}
\tag{9}
$$

We approximate the expectations w.r.t $q(x|\boldsymbol{w}, \eta)$ in Eqs. 7 and 8 with Monte Carlo integration:

$$
\begin{aligned}
&\mathbb{E}_{q(x|\boldsymbol{w}, \eta)}\left[\log p\left(\boldsymbol{w}|x, \boldsymbol{\Phi}, \boldsymbol{\beta}\right)\right] \approx \\
&\frac{1}{L} \sum_{l=1}^{L} \log p\left(\boldsymbol{w}|x^{(l)}, \boldsymbol{\Phi}, \boldsymbol{\beta}\right) = \frac{1}{L} \sum_{l=1}^{L} \log\left(\theta^{(l)} \boldsymbol{\beta}\right) \boldsymbol{w}^T
\end{aligned}
\tag{10}
$$

$$
\begin{aligned}
&\mathbb{E}_{q(x_i|\boldsymbol{w}_i, \eta)}\left[\log p(\mathbf{R}_i|x_i, \boldsymbol{X}_{-x_i}, \boldsymbol{\mathcal{S}})\right] \approx \\
&\frac{1}{L} \sum_{l=1}^{L} \log p(\mathbf{R}_i|x_i^{(l)}, \boldsymbol{X}_{-x_i}^{(l)}, \boldsymbol{\mathcal{S}}^{(l)})
\end{aligned}
\tag{11}
$$

where $x^{(l)} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2} \boldsymbol{\epsilon}^{(l)}$, $\boldsymbol{\epsilon}^{(l)} \sim \text{Normal}\left(\mathbf{0}, \boldsymbol{I}\right)$, $\theta^{(l)} \in \mathbb{R}^Z$ is a row vector of topic distribution, and $\theta_z^{(l)} = p\left(z|x^{(l)}, \boldsymbol{\Phi}\right)$ is computed as in Eq. 2[4]. Computing Eq. 11 needs the coordinates of other documents $\boldsymbol{X}_{-x_i}$ and aspect coordinates $\boldsymbol{\mathcal{S}}$. We perform the sampling for all documents x_i, aspects π_s, and for every x_i we compute Eq. 11 given the sample l of coordinates of other documents, $\boldsymbol{X}_{-x_i}^{(l)}$, and aspect coordinates $\boldsymbol{\mathcal{S}}^{(l)}$. For the whole corpus,

[4] In our experiments, $L = 1$ works well for all settings.

we maximize the following final objective function:

$$\mathcal{L}(\mathbf{\Psi}) = \sum_{n=1}^{N,S} \Bigg[-\frac{1}{2} \Big(\mathrm{tr}\left((\gamma \boldsymbol{I})^{-1} \boldsymbol{\Sigma}_n\right) + (-\boldsymbol{\mu}_n)^T (\gamma \boldsymbol{I})^{-1} (-\boldsymbol{\mu}_n) - D + \log \frac{|\gamma \boldsymbol{I}|}{|\boldsymbol{\Sigma}_n|} \Big) + \log \left(\hat{\theta}_n \boldsymbol{\beta} \right) \boldsymbol{w}_n^T \Bigg] + \sum_{s}^{S} \sum_{i}^{N} \sum_{j \neq i, R_{sij}=1}^{N} \log p(R_{sij} | x_i, x_j, \hat{\boldsymbol{\mathcal{S}}}) \quad (12)$$

where $\boldsymbol{\mu}_n$, diagonal $\boldsymbol{\Sigma}_n \in \mathbb{R}^D$ are outputs of the encoding feed-forward neural network. In summary, the main steps of the inference algorithm are shown in Algorithm 1 and Figure 3 shows the inference network used for our proposed model. As network settings, we set $H1$, $H2$, $H3$, $H4$ = 100, dropout rate as 0.2, $relu$ as our activation function, and we apply batch normalization to fully connected layers of documents - $fc3$, $fc4$ and topics - $fc7$.

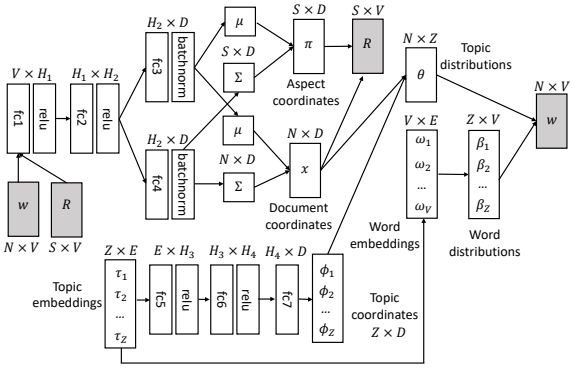

Figure 3: Inference Network of *FoTo*.

Algorithm 1 Inference Algorithm of *FoTo*

Require: A corpus of N documents $\mathcal{W} = \{\boldsymbol{w}_1, ..., \boldsymbol{w}_N\}$
A set of S targeted aspects $\mathcal{S} = \{a_1, ..., a_S\}$
Pairwise document ranking: $\mathbf{R}$
Inference model: $q(x|\boldsymbol{w}, \eta)$
Generative model: $p(\boldsymbol{w}, R|\boldsymbol{X}, \mathbf{\Phi}, \boldsymbol{\beta}, \boldsymbol{\mathcal{S}})$, $p(x|\gamma)$

1: **while** $notConverged()$ **do**
2: Sample a document $\boldsymbol{w} \sim \mathcal{W}$
3: Obtain its pairwise ranking R
4: Estimate its posterior parameters (μ, Σ)
5: Sample $x \sim q(x|\boldsymbol{w}, \eta)$
6: Evaluate $\mathcal{L}(\mathbf{\Psi})$ in Eq. 12
7: Update $\eta, \mathbf{\Phi}, \boldsymbol{\mathcal{S}}, \beta, \boldsymbol{\tau}$ using ADAM
8: **end while**

Time Complexity Analysis. In Algorithm 1, for one epoch, the main bottleneck is in steps 4 and 6. Step 4 is for computing $\boldsymbol{\mu}$, diagonal $\boldsymbol{\Sigma}_n \in \mathbb{R}^D$ which are outputs of the encoding feed-forward neural network with variational parameters η. Given the inference network as in Figure 3, $D = 2,3 \ll V$, and assuming that $H1$, $H2$ are fixed, the complexity of Step 4 is $\mathcal{O}((N+S)V)$ where N is the number of documents, S is the number of aspects, and V is the vocabulary size. Step 6 is to compute the loss function in Equation 12 whose complexity is $\mathcal{O}((N+S)ZV + SNC)$ where Z is the number of topics, $H3, H4, E$ are fixed, and C is the number of documents containing at least one of the keywords. If $S \ll N$, which is likely true because the query is often short, and $C \ll N$ (e.g., when the aspects of interest are rare), the asymptotic time complexity of one epoch in Algorithm 1 is $\mathcal{O}(NZV)$ which is linear in the number of documents, topics, and vocabulary size.

3. Experiments

3.1. Datasets and Experimental Settings

We conduct comprehensive experiments on four publicly available short text datasets of different sizes and domains. BBC [5] includes 2,095 news articles categorized into 5 classes. SEARCHSNIPPET [6] contains 12,076 web searched snippets from 8 different domains. YAHOOANSWERS (Zhang et al., 2015) has 10 classes and contains 40,802 titles and contents of questions from the Yahoo! Answers. NEWSCATEGORY[7] has 145,304 news headlines and short descriptions from HuffPost from 2012 to 2022. They are grouped into 15 categories. The average document length is from 10-14 words. The vocabulary size is 2k, 3k, 4k, 5k for BBC, SEARCHSNIPPET, YAHOOANSWERS, and NEWSCATEGORY respectively.

As in (Wang et al., 2016), for each dataset we select 4 queries of different sizes. The selected queries cover both popular and rare aspects. More specifically, for each dataset, we select 4 queries corresponding to the following templates: pp, rr, ppr, prr. Here p is a keyword for a popular aspect, and r is a keyword for a rare aspect. The popularity and rarity of a word are based on its frequency in the corpus. Here is the list of queries for all datasets: 1) BBC: Q1 = {sector, corporate}, Q2 = {law, immigration}, Q3 = {administration, policy, public}, Q4 = {private, bank, debt}; 2) SEARCHSNIPPET: Q1 = {biz, economics}, Q2 = {healthcare, fitness}, Q3 = {sporting, athlete, racing}, Q4 = {notebook, microprocessor, disk}; 3) YAHOOANSWERS: Q1 = {lifestyle, school}, Q2 = {disease, diabetes}, Q3 = {musical, song, guitar}, Q4 = {youtube, social, gaming}; 4) NEWSCATEGORY: Q1 = {republican,

[5] http://mlg.ucd.ie/datasets/bbc.html
[6] http://jwebpro.sourceforge.net/data-web-snippets.tar.gz
[7] https://www.kaggle.com/datasets/rmisra/news-category-dataset

election}, Q2 = {playoff, tournament}, Q3 = {family, child, medicare}, Q4 = {kid, halloween, cartoon}.

For each query, we extend it by adding more keywords as described in Section 2. We set dropout rate = 0.2, γ = 1, learning rate = 0.001. The batch size is set to 1000 for YAHOOANSWERS, NEWSCATEGORY and 250 for other datasets. For word embeddings, we train skip-gram on all corpora (Mikolov et al., 2013). We use Adam as the optimizing algorithm. All models are trained with 1000 epochs and the results are averaged across 5 independent runs.

3.2. Baselines

We compare the following state-of-the-art methods:

- Targeted topic modeling: TTM (Wang et al., 2016)[8], QDTM (Fang et al., 2021)[9]. These are targeted topic models designed for extracting topics relevant to aspects of interest. For completeness, we also compare our method with non-targeted topic models including BerTopic (Grootendorst, 2022)[10], ProdLDA (Srivastava and Sutton, 2017)[11], ETM (Dieng et al., 2020)[12], ASTM (Wang et al., 2018)[13], and TSCTM (Wu et al., 2022)[14]. Except for BerTopic and ProdLDA, others are short-text topic models. These models only extract topics and do not generate visualization. Therefore, to visualize documents, we use t-SNE to visualize the topic distributions.
- Visual topic modeling: PLSV (Iwata et al., 2008)[15], and WTM for short texts (Kumar and Le, 2021)[16]. These models can extract topics and generate visualization of documents. They are not targeted topic models.
- Targeted visual topic modeling: *FoTo* [this paper][17]. Our method learns both relevant topics and focused visualization of short texts that embeds queries, relevant documents and topics in the same visualization.

Since we consider queries as pseudo-documents, for a fair comparison, we pass them together along with other normal documents as inputs to the baseline models and extract similar documents. As shown in the experiments, this approach is not ideal because these models are still not enforced to focus on the aspects of interest.

[8] https://github.com/shuaiwanghk/TTM
[9] https://github.com/Fitz-like-coding/QDTM
[10] https://github.com/MaartenGr/BERTopic
[11] https://github.com/hyqneuron/pytorch-avitm
[12] https://github.com/adjidieng/ETM
[13] https://github.com/wjmzjx/ASTM
[14] https://github.com/BobXWu/TSCTM
[15] https://github.com/dangpnh2/plsv_vae, (Pham and Le, 2020)
[16] https://github.com/sanujsriv/WTM
[17] https://github.com/sanujsriv/FoTo

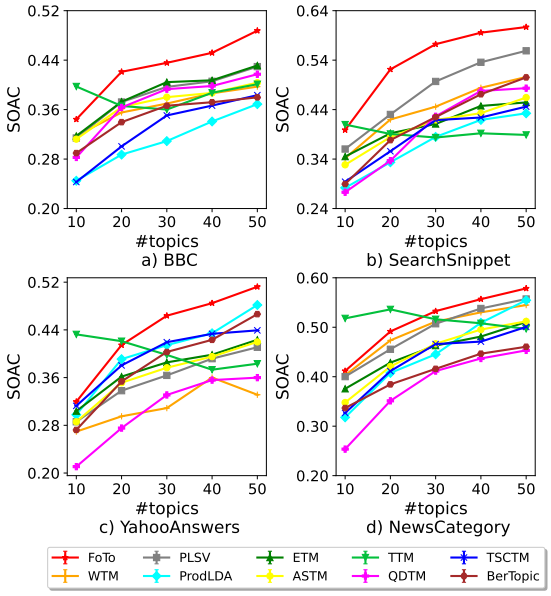

Figure 4: Comparison based on the sum of averaged cosine similarity (SOAC).

3.3. Extracting Focused Topics

In this section, we show that, as a targeted topic model, our method can extract quality focused topics that are relevant to the targeted aspects. We use the sum of averaged cosine to measure the quality of the focused topics. For each original aspect keyword, we compute its averaged cosine similarity to top 10 words of a topic using word embeddings. We take the sum of this averaged cosine similarity (SOAC) across all aspects. Since only a few topics would be relevant to the query, for every method, we take the average of SOAC values for the top 5 topics that have the highest SOAC values. We report the averaged results across queries and runs in Figure 4. Although TTM and QDTM are targeted topic models, their performance is quite comparable to the other methods. This could be because they are not designed for short texts. In contrast, since *FoTo* is a targeted topic model for short texts, it consistently achieves the highest SOAC value in most settings, which demonstrates that the topics extracted by *FoTo* are more relevant to the query than that of the other methods. We show some example relevant topics in Section 3.7.

3.4. Document Relevant Ranking in 2-D

A good focused visualization should preserve well relevant documents in the visualization space. Therefore, to measure the quality of the focused visualization, we rely on the task of document relevant ranking in visualization space. Based on the locations of the aspects in the visualization (i.e., π_s), we extract the nearest documents to the aspects in the visualization. Intuitively, a good

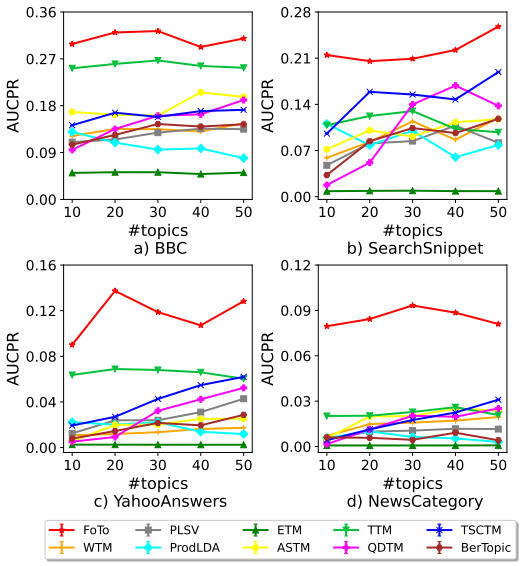

Figure 5: AUCPR with TF-IDF ground truth.

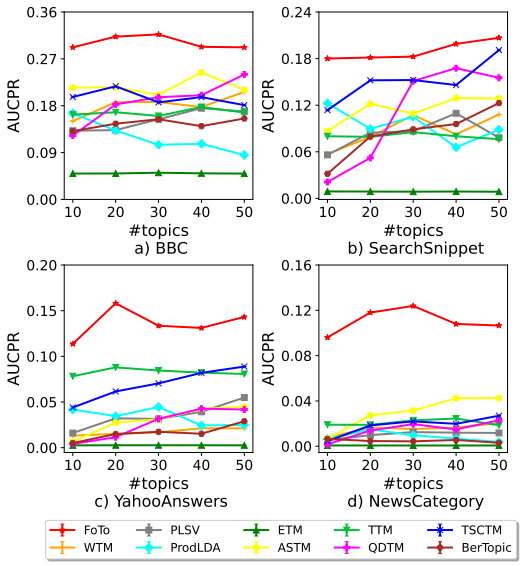

Figure 6: AUCPR with DESM ground truth.

focused visualization will put the most relevant documents close to the aspects. More specifically, we rank the relevance of a document to the query by its minimum distance to any one of the aspects in the query. The result ranking will be compared with ground-truth relevant documents and the area under the Precision-Recall curve (AUCPR) is reported. Higher AUCPR means better focused visualization. We use two different ways to obtain ground-truth documents. The first way is using the cosine similarity between TF-IDF vectors of documents and the query. The second way is using DESM score that utilizes word embeddings to rank documents (Mitra et al., 2016). We use top 100 documents as ground-truth documents.

Figure 5 shows averaged AUCPR across all queries and runs with TF-IDF ground truth. TTM outperforms the non-targeted topic models on BBC and YAHOOANSWERS. However, since it is not a joint model for visualization, its outperformance is not consistent across all datasets. Contrarily, *FoTo*, a joint model of targeted topic modeling and visualization, consistently outperforms all baselines in this task. The results show that *FoTo* generates good focused visualization placing relevant documents close to the queried aspects. The example visualizations shown in Section 3.7 will demonstrate this further. For AUCPR with DESM ground truth in Figure 6, we observe a similar trend.

3.5. k-NN Accuracy vs. AUCPR

In this task, we show that while extracting and visualizing relevant topics and documents, our method can still produce a good overview visualization of all documents. To measure the quality of the overview visualization, we rely on the document labels and compute the k-nearest neighbors (k-NN) classification accuracy in the visualization space. A good overview visualization should group documents of the same label together, which therefore will yield high k-NN accuracy. To see how methods balance the quality of overall visualization and focused visualization, we plot k-NN vs. AUCPR in Figure 7 ($k = 50, Z = 50$). *FoTo* is seen to be on the top right, which shows that the overview visualizations generated by our method are at par with others while being better in terms of AUCPR.

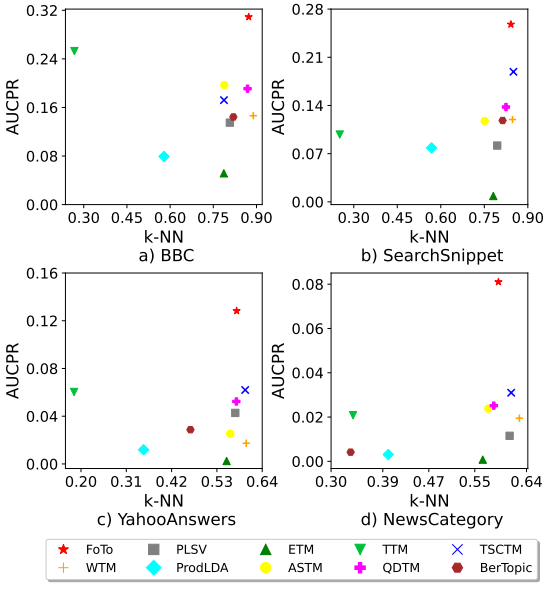

Figure 7: k-NN vs. AUCPR in visualization space.

3.6. Topic Coherence vs. SOAC

Topic coherence is a popular metric to evaluate the quality of topic models. In this task, we show that our method's extracted topics are focused and coherent. We use the Normalized Pointwise Mutual Information (NPMI) whose probabilities of words

are estimated based on a large external corpus (Lau et al., 2014). To see whether the quality of focused topics affects the topic coherence, we plot NPMI vs. SOAC in Figure 8. The results are averaged across top 5 topics that have the highest SOAC values and across topic settings. *FoTo* is seen to be on the top right, which indicates that the topics extracted by our method are coherent while being more focused.

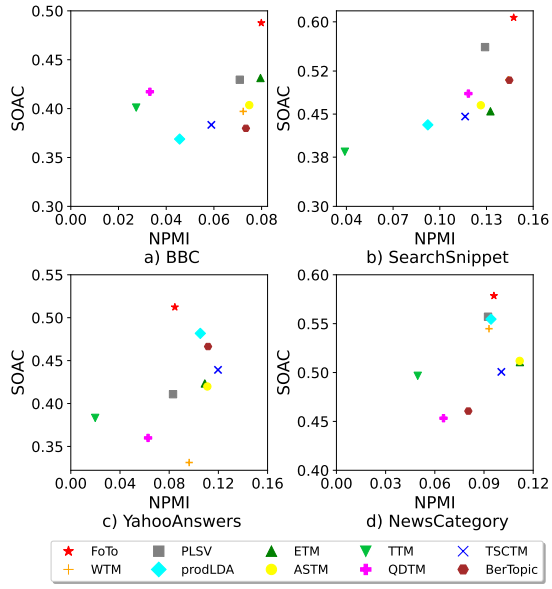

Figure 8: Topic coherence NPMI vs. SOAC.

3.7. Topic and Visualization Examples

We present the qualitative analysis of *FoTo* in terms of focused topics and visualization. Table 1 shows the top 2 most relevant topics based on the topics' SOAC values by *FoTo*, TTM, and TSCTM on SEARCHSNIPPET for the query {sporting, athlete, racing}. The words in red are extended words that are most similar to the queried keywords in terms of cosine similarity. It is evident that topics produced by *FoTo* are more focused than the topics extracted by other baselines because they contain more red words.

Table 1: Top 2 most relevant topics on SEARCHSNIPPET for the query {sporting, athlete, racing}.

FoTo		TTM		TSCTM	
Topic3	Topic9	Topic5	Topic2	Topic5	Topic9
sport	game	sport	horse	mlb	movie
football	news	game	yahoo	forehand	actress
game	car	online	race	volleyball	cruise
news	tennis	ebay	auto	nba	celebs
soccer	match	boxing	sporting	quarterfinal	celebrity
league	sport	horseracing	basketball	nhl	stranger
team	racing	olympics	electronics	league	actor
player	wheel	portland	save	softball	showtime
hockey	tournament	pool	extra	champion	filmography
espn	golf	facility	equipment	ncaa	julia

Figure 9 shows the visualizations of SEARCHSNIPPET by *FoTo*, WTM, TTM, and TSCTM for 50 topics on two different queries. On the left of Figures 9(a) and 9(b), *FoTo* gives a good overview of the corpus by grouping similar documents together with their topics. Moreover, it can preserve and visualize topics, documents that are relevant to targeted aspects. The right of Figures 9(a) and 9(b) show how relevant documents are distributed. We calculate the TF-IDF scores of documents w.r.t the query and estimate the relevance density by averaging the relevant scores of documents in a region. As we can see, most of the relevant documents are near the keywords, indicating that our method preserves well the documents and topics of interest. For TSCTM, WTM, and TTM, relevant documents are more scattered and mixed with non-relevant documents, as indicated by the lower document relevance density near the keywords in the visualization.

4. Related Work

Several short-text topic models have been developed (Yan et al., 2013; Zuo et al., 2016; Li et al., 2016; Wang et al., 2018; Dieng et al., 2020). In these models, one of the approaches is using word embeddings as supplementary information to enrich the learned topics. ASTM (Wang et al., 2018) further proposes an attention mechanism based on word embeddings to segment words of a document into different groups. A topic is then assigned to each group to obtain more coherent topics. Another approach is to aggregate short texts into pseudo-documents that are effective for topic learning (Zuo et al., 2016; Lin et al., 2020; Quan et al., 2015). Recently, there have been methods that opt contrastive learning for generating topic representations (Wu et al., 2022). We also have models (Sia et al., 2020; Grootendorst, 2022) that cluster document embeddings generated using pre-trained transformer-based language models for learning topic representations. These are not targeted topic models and do not generate visualization.

There have been topic models for visualizing long or short texts (Iwata et al., 2008; Le and Lauw, 2017; Kumar and Le, 2021). However, these models are not for focused analysis. Targeted topic models can be used to tackle this problem (Wang et al., 2016; Fang et al., 2021). They aim to extract all topics relevant to aspects of interest given a set of keywords. In another direction, given a set of seed words, seed-guided topic models aim to find a topic for each seed word (Meng et al., 2020; Harandizadeh et al., 2022; Zhang et al., 2023; Lin et al., 2023). Therefore, the number of extracted topics is equal to the number of seed words. This is different from targeted topics models which aim to find all topics relevant to the given keywords. None of these models learn embeddings of documents and topics for visualization.

(a) sporting , athlete, racing

(b) notebook, microprocessor, disk

Figure 9: Visualization of SEARCHSNIPPET by *FoTo*, WTM, TTM, TSCTM for different queries and their relevance densities.

5. Conclusion

We propose a novel targeted visual topic model for focused analysis of short texts. *FoTo* is a joint model to extract and visualize topics, documents that are relevant to targeted aspects. A unified generative process is proposed to model topics, document and topic embeddings, keywords, as well as pairwise document ranking for visual focused analysis. The extensive experiments show the effectiveness of our proposed model in terms of targeted topic modeling and visualization. For future work, we plan to extend the current model for generalizing to general texts.

6. Ethical Statement

There are no ethical issues.

7. Data and Source Code

The paper uses publicly available datasets. The source code of *FoTo* can be found at https://github.com/sanujsriv/FoTo.

8. Acknowledgements

This research is sponsored by NSF #1757207 and NSF #1914635.

9. Bibliographical References

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

Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Zheng Fang, Yulan He, and Rob Procter. 2021. A query-driven topic model. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1764–1777.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Bahareh Harandizadeh, J Hunter Priniski, and Fred Morstatter. 2022. Keyword assisted embedded topic model. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 372–380.

Jin He, Lei Li, Yan Wang, and Xindong Wu. 2020. Targeted aspects oriented topic modeling for short texts. *Applied Intelligence*, 50(8):2384–2399.

Tomoharu Iwata, Takeshi Yamada, and Naonori Ueda. 2008. Probabilistic latent semantic visualization: topic model for visualizing documents. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 363–371.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Sanuj Kumar and Tuan Le. 2021. A word embedding topic model for robust inference of topics and visualization. In *The First International Conference on AI-ML-Systems*, pages 1–7.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539.

Tuan Le and Hady Lauw. 2017. Semantic visualization for short texts with word embeddings. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence IJCAI-17*. IJCAI.

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 International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 165–174.

Hao Lin, Yuan Zuo, Guannan Liu, Hong Li, Junjie Wu, and Zhiang Wu. 2020. A pseudo-document-based topical n-grams model for short texts. *World Wide Web*, 23(6):3001–3023.

Yang Lin, Xin Gao, Xu Chu, Yasha Wang, Junfeng Zhao, and Chao Chen. 2023. Enhancing neural topic model with multi-level supervisions from seed words. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13361–13377.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29).

Yu Meng, Jiaxin Huang, Guangyuan Wang, Zihan Wang, Chao Zhang, Yu Zhang, and Jiawei Han. 2020. Discriminative topic mining via category-name guided text embedding. In *Proceedings of The Web Conference 2020*, pages 2121–2132.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Bhaskar Mitra, Eric Nalisnick, Nick Craswell, and Rich Caruana. 2016. A dual embedding space model for document ranking. *arXiv preprint arXiv:1602.01137*.

Dang Pham and Tuan Le. 2020. Auto-encoding variational bayes for inferring topics and visualization. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5223–5234.

Xiaojun Quan, Chunyu Kit, Yong Ge, and Sinno Jialin Pan. 2015. Short and sparse text topic modeling via self-aggregation. In *Twenty-fourth international joint conference on artificial intelligence*.

Suzanna Sia, Ayush Dalmia, and Sabrina J Mielke. 2020. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! *arXiv preprint arXiv:2004.14914*.

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

Jian Tang, Jingzhou Liu, Ming Zhang, and Qiaozhu Mei. 2016. Visualizing large-scale and high-dimensional data. In *Proceedings of the 25th international conference on world wide web*, pages 287–297.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Jiamiao Wang, Ling Chen, Lei Li, and Xindong Wu. 2021. Bittm: A core biterms-based topic model for targeted analysis. *Applied Sciences*, 11(21):10162.

Jiamiao Wang, Ling Chen, Lu Qin, and Xindong Wu. 2018. Astm: An attentional segmentation based topic model for short texts. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 577–586. IEEE.

Shuai Wang, Zhiyuan Chen, Geli Fei, Bing Liu, and Sherry Emery. 2016. Targeted topic modeling for focused analysis. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1235–1244.

Xiaobao Wu, Anh Tuan Luu, and Xinshuai Dong. 2022. Mitigating data sparsity for short text topic modeling by topic-semantic contrastive learning. *arXiv preprint arXiv:2211.12878*.

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.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28:649–657.

Yu Zhang, Yunyi Zhang, Martin Michalski, Yucheng Jiang, Yu Meng, and Jiawei Han. 2023. Effective seed-guided topic discovery by integrating multiple types of contexts. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 429–437.

Yuan Zuo, Junjie Wu, Hui Zhang, Hao Lin, Fei Wang, Ke Xu, and Hui Xiong. 2016. Topic modeling of short texts: A pseudo-document view. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2105–2114.