

# Federated Learning for Short Text Clustering

Mengling Hu, Chaochao Chen, Weiming Liu, Xinting Liao, and Xiaolin Zheng,  
 Zhejiang University, China  
 {humengling, zjuccc, 21831010, xintingliao, xlzheng}@zju.edu.cn

## Abstract

Short text clustering has been popularly studied for its significance in mining valuable insights from many short texts. In this paper, we focus on the federated short text clustering (FSTC) problem, i.e., clustering short texts that are distributed in different clients, which is a realistic problem under privacy requirements. Compared with the centralized short text clustering problem that short texts are stored on a central server, the FSTC problem has not been explored yet. To fill this gap, we propose a Federated Robust Short Text Clustering (**FSTC**) framework. **FSTC** includes two main modules, i.e., *robust short text clustering module* and *federated cluster center aggregation module*. The robust short text clustering module aims to train an effective short text clustering model with local data in each client. We innovatively combine optimal transport to generate pseudo-labels with Gaussian-uniform mixture model to ensure the reliability of the pseudo-supervised data. The federated cluster center aggregation module aims to exchange knowledge across clients without sharing local raw data in an efficient way. The server aggregates the local cluster centers from different clients and then sends the global centers back to all clients in each communication round. Our empirical studies on three short text clustering datasets demonstrate that **FSTC** significantly outperforms the federated short text clustering baselines.

## 1 Introduction

Short text clustering has been proven to be beneficial in many applications, such as, news recommendation [Wu *et al.*, 2022], opinion mining [Stieglitz *et al.*, 2018], stance detection [Li *et al.*, 2022], etc. Existing short text clustering models [Xu *et al.*, 2017; Hadifar *et al.*, 2019; Rakib *et al.*, 2020; Zhang *et al.*, 2021] all assume that the short texts to be clustered are stored on a central server where their models are trained. However, the assumption may be invalid when the data is distributed among many clients and gathering the data on a central server is not feasible due to privacy regulations or communication concerns [Magdziarczyk, 2019;

McMahan *et al.*, 2017a; Stallmann and Wilbik, 2022]. For example, a multi-national company, with several local markets, sells similar commodities in all markets. Each local market has text data about their customers, e.g., personal information, purchased items, reviews, etc. As text clustering is one of the most fundamental tasks in text mining, the company wishes to cluster text data from all markets for text mining, which can mine more reliably valuable insights compared with only clustering local data. The mined valuable information can further guide the marketing strategy of the company. Because of strict privacy regulations, the company is not allowed to gather all data in a central server, e.g., European customer data is not allowed to be transferred to most countries outside of Europe [Otto, 2018; Stallmann and Wilbik, 2022].

In this paper, we focus on the federated short text clustering (FSTC) problem. Federated learning is widely used to enable collaborative learning across a variety of clients without sharing local raw data [Tan *et al.*, 2022]. Federated clustering is a kind of federated learning setting whose goal is to cluster data that is distributed among multiple clients. Unlike popular federated supervised classification task, federated unsupervised clustering task is less explored and existing federated clustering methods cannot work well with short texts. Specifically, existing federated clustering methods can be divided into two types, i.e., the k-means based federated clustering methods [Kumar *et al.*, 2020; Pedrycz, 2021; Dennis *et al.*, 2021; Stallmann and Wilbik, 2022] and the deep neural network based federated clustering method [Chung *et al.*, 2022]. The former methods are not applicable to short texts because short texts often have very sparse representations that lack expressive ability. It is beneficial to utilize deep neural network to enrich the short text representations for better clustering performance [Xu *et al.*, 2017; Hadifar *et al.*, 2019; Zhang *et al.*, 2021]. However, the latter deep learning based method [Chung *et al.*, 2022] cannot cope with real-world noisy data well. Therefore, existing federated clustering methods cannot be utilized to solve the FSTC problem.

The FSTC problem has not been explored yet, possibly because short texts are sparse and noisy, which makes it difficult to cluster short texts in the federated environment. [McMahan *et al.*, 2017b] proposes FedAvg to substitute synchronized stochastic gradient descent for the federated learning of deep neural networks. Combining the state-of-the-

<p>86 art short text clustering models with FedAvg [McMahan <i>et</i> al., 2017b] seems to be a reasonable way to solve the FSTC  87 problem. However, the combination cannot solve the FSTC  88 problem well. Firstly, existing short text clustering models  89 [Xu <i>et al.</i>, 2017; Hadifar <i>et al.</i>, 2019; Rakib <i>et al.</i>, 2020;  90 Zhang <i>et al.</i>, 2021] cannot learn sufficiently discriminative  91 representations due to lacking supervision information, caus-  92 ing limited clustering performance. Secondly, FedAvg needs  93 to aggregate models in every communication round, causing  94 limited communication efficiency. In summary, there are two  95 main challenges, i.e., <b>CH1</b>: How to provide supervision in-  96 formation for discriminative representation learning, and pro-  97 mote better clustering performance? <b>CH2</b>: How to exchange  98 knowledge across clients in a more efficient way?</p> <p>100 To address the aforementioned challenges, in this paper,  101 we propose <b>FSTC</b>, a novel framework for federated short  102 text clustering. In order to provide supervision information  103 (solving <b>CH1</b>) and exchange knowledge (solving <b>CH2</b>), we  104 utilize two modules in <b>FSTC</b>, i.e., <i>robust short text clustering module</i> and <i>federated cluster center aggregation module</i>. The robust short text clustering module aims to tackle  105 the first challenge by generating pseudo-labels as the super-  106 vision information. We leverage optimal transport to gener-  107 ate pseudo-labels, and introduce Gaussian-uniform mixture  108 model to estimate the probability of correct labeling for more  109 reliable pseudo-supervised data. The federated cluster cen-  110 ter aggregation module aims to tackle the second challenge  111 by aggregating cluster centers rather than models in every  112 communication round. We use the locally generated pseudo-  113 labels to divide the clusters of a client for obtaining the local  114 cluster centers, and align the local centers of all clients for  115 collaboration.</p> <p>116 We summarize our main contributions as follows: (1) We  117 propose a novel framework, i.e., <b>FSTC</b>, for federated short  118 text clustering. To our best knowledge, we are the first to ad-  119 dress short text clustering problem in the federated learning  120 setting. (2) We propose an end-to-end model for local short  121 text clustering, which can learn more discriminative repre-  122 sentations with reliable pseudo-supervised data and promote  123 better clustering performance. (4) We conduct extensive ex-  124 periments on three short text clustering datasets and the re-  125 sults demonstrate the superiority of <b>FSTC</b>.</p>	<p>[Hadifar <i>et al.</i>, 2019; Zhang <i>et al.</i>, 2021] integrate cluster-  143 ing with deep representation learning to learn the representa-  144 tions that are appropriate for clustering. Moreover, [Zhang  145 <i>et al.</i>, 2021] utilizes the pre-trained SBERT [Reimers and  146 Gurevych, 2019] and contrastive learning to learn discrimina-  147 tive representations. However, the learned representations are  148 still insufficiently discriminative due to the lack of supervi-  149 sion information [Hu <i>et al.</i>, 2021]. As a contrast, in this work,  150 we combine pseudo-label technology with Gaussian-uniform  151 mixture model to provide reliable supervision to learn more  152 discriminative representations.</p>
<b>2.2 Federated Clustering</b>	
Federated clustering aims to cluster data that is distributed 155 among multiple clients. Unlike the popularity of federat- 156 ed supervised classification task, federated unsupervised 157 clustering is underdeveloped. Existing federated clustering 158 methods can be divided into two types, i.e., the k-means 159 based federated clustering methods [Kumar <i>et al.</i> , 2020; 160 Pedrycz, 2021; Dennis <i>et al.</i> , 2021; Stallmann and Wilbik, 161 2022] and the deep neural network based federated cluster- 162 ing method [Chung <i>et al.</i> , 2022]. [Kumar <i>et al.</i> , 2020] ex- 163 tends k-means algorithm to the federated setting, they pro- 164 pose calculating a weighted average of local cluster centers 165 to update the global cluster centers, the weights are given by 166 the samples number in clusters. [Pedrycz, 2021] introduces 167 a fuzzy c-means federated clustering, which uses fuzzy as- 168 signments as weights instead of the samples number in clus- 169 ters. [Dennis <i>et al.</i> , 2021] introduces a one-shot k-means fed- 170 erated clustering method which utilize k-means to aggregate 171 and update the global cluster centers. [Stallmann and Wilbik, 172 2022] proposes a federated fuzzy c-means clustering method 173 which is similar to [Kumar <i>et al.</i> , 2020; Pedrycz, 2021; 174 Dennis <i>et al.</i> , 2021]. The deep neural network based federat- 175 ed clustering methods are not well studied. Only [Chung <i>et</i> 176 <i>al.</i> , 2022] develops a new generative model based clustering 177 method in the federated setting, based on IFCA [Ghosh <i>et al.</i> , 178 2020] algorithm. However, [Chung <i>et al.</i> , 2022] shows that 179 the method can obtain good clustering performance for syn- 180 thetic datasets, but always fails when training with real-world 181 noisy data. As a contrast, in this work, our method can obtain 182 good clustering performance for the real-world noisy data.	
<b>2 Related Work</b>	
<b>2.1 Short Text Clustering</b>	
<p>128 Short text clustering is not a trivial task due to the weak sig-  129 nal contained in each text instance. The existing short text  130 clustering methods can be divided into tree kinds: (1) tra-  131 ditional methods, (2) deep learning methods, and (3) deep  132 joint clustering methods. The traditional methods [Scott and  133 Matwin, 1998; Salton and McGill, 1983] often obtain very  134 sparse representations that lack discriminations. The deep  135 learning method [Xu <i>et al.</i>, 2017] leverages pre-trained word  136 embeddings [Mikolov <i>et al.</i>, 2013] and deep neural network  137 to enrich the representations. However, it does not combine  138 a clustering objective with the deep representation learning,  139 which prevents the learned representations from being ap-  140 propriate for clustering. The deep joint clustering methods</p>	
<b>3 Methodology</b>	
<b>3.1 An Overview of FSTC</b>	
<p>144 The goal of <b>FSTC</b> is to collaboratively train a global short  145 text clustering model with the raw data stored locally in mul-  146 tiple clients. The overall structure of our proposed <b>FSTC</b> is  147 illustrated in Fig.1. <b>FSTC</b> consists of two main modules, i.e.,  148 <i>robust short text clustering module</i> and <i>federated cluster cen-  149 ter aggregation module</i>. The robust short text clustering mod-  150 ule aims to train a short text clustering model with local data  151 in a client. The federated cluster center aggregation module  152 aims to efficiently exchange information across clients with-  153 out sharing their local raw data. In the end, we can obtain  154 the global model by averaging the final local models. We will  155 introduce these two modules in detail later.</p>	

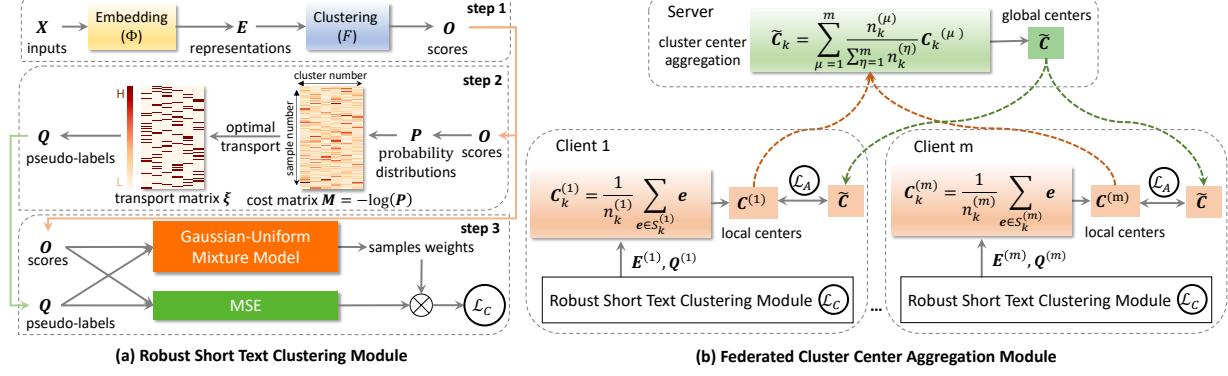


Figure 1: An overview of FSTC.

### 3.2 Robust Short Text Clustering Module

We first introduce the robust short text clustering module. Although the deep joint clustering methods [Hadifar *et al.*, 2019; Zhang *et al.*, 2021] on short text clustering are popular these days, their clustering performance is still limited. The reason is that lacking supervision information prevents those methods from learning more discriminative representations [Hu *et al.*, 2021]. Therefore, to provide supervision information for short text clustering, we propose to generate *pseudo-labels* for guiding the local model training, that is, unsupervised training samples turn into pseudo-supervised training samples. Moreover, because the pseudo-labels are inevitably noisy, we design a robust learning objective for fully exploiting the pseudo-supervised training samples. An overview of the robust short text clustering module is shown in Fig.1(a), which mainly has three steps: **Step 1**: predicting cluster assignment scores, **Step 2**: generating pseudo-labels, and **Step 3**: obtaining a robust objective. Note that, we divide the robust short text clustering module into three steps just for the sake of introduction convenience, the whole module is trained in an end-to-end way. We will introduce the details below.

**Step 1: predicting cluster assignment scores.** This step aims to predict and provide cluster assignment scores of the inputs for the other two steps. For inputs  $\mathbf{X}$ , we adopt SBERT [Reimers and Gurevych, 2019] as the embedding network  $\Phi$  to obtain the representations, i.e.,  $\Phi(\mathbf{X}) = \mathbf{E} \in \mathbb{R}^{N \times D}$ , where  $N$  denotes batch size and  $D$  is the dimension of the representations. We utilize fully connected layers as the clustering network  $F$  to predict cluster assignment scores, i.e.,  $F(\mathbf{E}) = \mathbf{O} \in \mathbb{R}^{N \times K}$ , where  $K$  is the number of clusters.

**Step 2: generating pseudo-labels.** This step aims to generate pseudo-labels of all samples by excavating information from the cluster assignment scores and provide the pseudo-labels for **Step 3**. To begin with, we use softmax [Bridle, 1990] to normalize the scores  $\mathbf{O}$  for obtaining the cluster assignment probability distributions  $\mathbf{P} \in \mathbb{R}^{N \times K}$ . For each sample, we expect its generated pseudo-label distribution to align its predicted probability distribution. Specifically, we denote the pseudo-labels as  $\mathbf{Q} \in \mathbb{R}^{N \times K}$ . We adopt KL-divergence minimization for aligning the pseudo-label distributions and the predicted probability distributions. Meanwhile, to avoid the trivial solution that all samples are assigned to one cluster, we add the constraint that the label assignments partition data equally [Asano *et al.*, 2020;

Caron *et al.*, 2020]. Formally, the objective is as follows:

$$\min_Q \text{KL}(\mathbf{Q} \parallel \mathbf{P}) \quad s.t. \quad \sum_{i=1}^N \mathbf{Q}_{ij} = \frac{N}{K},$$

where

$$\text{KL}(\mathbf{Q} \parallel \mathbf{P}) = - \sum_{i=1}^N \sum_{j=1}^K \mathbf{Q}_{ij} \log \mathbf{P}_{ij} + \sum_{i=1}^N \sum_{j=1}^K \mathbf{Q}_{ij} \log \mathbf{Q}_{ij}. \quad (1)$$

We turn this objective into a discrete optimal transport problem [Cuturi, 2013], i.e.,

$$\begin{aligned} \xi^* &= \underset{\xi}{\operatorname{argmin}} \langle \xi, \mathbf{M} \rangle + \epsilon H(\xi), \\ s.t., \quad \xi \mathbb{1}_K &= \mathbf{a}, \quad \xi^T \mathbb{1}_N = \mathbf{b}, \quad \xi \geq 0, \end{aligned} \quad (2)$$

where  $\xi = \mathbf{Q}$  denotes the transport matrix,  $\mathbf{M} = -\log \mathbf{P}$  denotes the cost matrix,  $H(\xi) = \sum_{i=1}^N \sum_{j=1}^K \xi_{ij} \log \xi_{ij}$  denotes the entropy constraint,  $\langle \cdot \rangle$  is the Frobenius dot-product between two matrices,  $\epsilon$  is the balance hyper parameter,  $\mathbf{a} = \mathbb{1}_N$ ,  $\mathbf{b} = \frac{N}{K} \mathbb{1}_K$ . We apply the Sinkhorn algorithm [Cuturi, 2013] to solve the optimal transport problem. Specifically, we introduce Lagrangian multipliers, then the objective turns into:

$$\min_{\xi} \langle \xi, \mathbf{M} \rangle + \epsilon H(\xi) - \mathbf{f}^T (\xi \mathbb{1}_K - \mathbf{a}) - \mathbf{g}^T (\xi^T \mathbb{1}_N - \mathbf{b}), \quad (3)$$

where  $\mathbf{f}$  and  $\mathbf{g}$  are both Lagrangian multipliers. Taking the differentiation of Equation (3) on the variable  $\xi$ , we can obtain:

$$\xi = \text{diag}(\mathbf{u}) \mathcal{K} \text{diag}(\mathbf{v}), \quad (4)$$

where  $\mathbf{u} = \exp(f/\epsilon - 1/2)$ ,  $\mathcal{K} = \exp(-\mathbf{M}/\epsilon)$ , and  $\mathbf{v} = \exp(g/\epsilon - 1/2)$ . Taking Equation (4) back to the original constraints  $\xi \mathbb{1}_K = \mathbf{a}$ ,  $\xi^T \mathbb{1}_N = \mathbf{b}$ , we can obtain:

$$\mathbf{u} = \mathbf{a} \oslash \mathcal{K} \mathbf{v}, \quad (5)$$

$$\mathbf{v} = \mathbf{b} \oslash \mathcal{K}^T \mathbf{u}, \quad (6)$$

where  $\oslash$  is the Hadamard division. Through iteratively solving Equation (5) and Equation (6), we can obtain the transport matrix  $\xi$  on Equation (4). Furthermore, although we let  $\xi = \mathbf{Q}$  before, we square the values in  $\xi$  for obtaining more reliable pseudo-labels  $\mathbf{Q}$  [Xie *et al.*, 2016]. Specifically,  $\mathbf{Q}$  is formulated as:

$$Q_{ik} = \frac{\xi_{ik}^2}{\sum_{k'=1}^K \xi_{ik'}^2}. \quad (7)$$

---

**Algorithm 1** Pseudo-label Generation.

---

**Input:** The cluster assignment scores  $\mathbf{O}$ ; the balance hyper-parameter  $\epsilon$ ; batch size  $N$ ; cluster number  $K$ .

**Procedure:**

- 1: Calculate the cluster assignment probability distributions  $\mathbf{P} = \text{softmax}(\mathbf{O})$
  - 2: Calculate the cost matrix  $\mathbf{M} = -\log \mathbf{P}$ .
  - 3: Calculate  $\mathcal{K} = \exp(-\mathbf{M}/\epsilon)$ .
  - 4: Randomly initialize vectors  $\mathbf{u} \in \mathbb{R}^N$  and  $\mathbf{v} \in \mathbb{R}^K$ .
  - 5: **for**  $i = 1$  to  $t$  **do**
  - 6:   Update  $\mathbf{u}$  by Equation (5).
  - 7:   Update  $\mathbf{v}$  by Equation (6).
  - 8: **end for**
  - 9: Calculate  $\xi$  by Equation (4).
  - 10: Calculate  $\mathbf{Q}$  by Equation (7).
  - 11: **return**  $\mathbf{Q}$
- 

267 We show the optimization scheme of pseudo-label generation  
268 in Algorithm 1.

269 **Step 3: obtaining a robust objective.** This step aims to de-  
270 sign a robust objective to fully exploit the pseudo-supervised  
271 samples. Although the generated pseudo-labels can be help-  
272 ful to learn more discriminative representations, not all of the  
273 pseudo-labels are correct and the wrong pseudo-labels may  
274 prevent our model from achieving better performance. Thus,  
275 to mitigate the influence of wrong pseudo-labels, we propose  
276 to estimate the probability of correct labeling, and use the  
277 probability to weight corresponding pseudo-supervised sam-  
278 ple. Specifically, inspired by [Lathuilière *et al.*, 2018], we use  
279 a Gaussian-uniform mixture model to model the distribution  
280 of a pseudo-label  $\mathbf{Q}_i$  conditioned by its cluster assignment  
281 score  $\mathbf{O}_i$ :

$$p(\mathbf{Q}_i|\mathbf{O}_i) = \pi\mathcal{N}(\mathbf{Q}_i; \mathbf{O}_i, \Sigma) + (1-\pi)\mathcal{U}(\mathbf{Q}_i; \gamma), \quad (8)$$

282 where  $\mathcal{N}(\cdot)$  denotes a multivariate Gaussian distribution and  
283  $\mathcal{U}(\cdot)$  denotes a uniform distribution. The Gaussian distribution  
284 models the correct pseudo-labels while the uniform distri-  
285 bution models the wrong pseudo-labels.  $\pi$  is the prior prob-  
286 ability of a correct pseudo-label,  $\Sigma \in \mathbb{R}^{K \times K}$  is the covariance  
287 matrix of the Gaussian distribution, and  $\gamma$  is the nor-  
288 malization parameter of the uniform distribution. The poste-  
289 rior probability of correct labeling for  $i$ -th sample is,

$$r_i \leftarrow \frac{\pi\mathcal{N}(\mathbf{Q}_i; \mathbf{O}_i, \Sigma)}{\pi\mathcal{N}(\mathbf{Q}_i; \mathbf{O}_i, \Sigma) + (1-\pi)\gamma}. \quad (9)$$

290 The parameters of Gaussian-uniform mixture models are  $\theta =$   
291  $\{\pi, \Sigma, \gamma\}$ . We update these parameters with:

$$\Sigma \leftarrow \sum_{i=1}^N r_i \boldsymbol{\delta}_i \boldsymbol{\delta}_i^T, \quad (10)$$

$$\pi \leftarrow \sum_{i=1}^N r_i / N, \quad (11)$$

$$\frac{1}{\gamma} \leftarrow \prod_{k=1}^K 2\sqrt{3(C_{2k} - C_{1k}^2)}, \quad (12)$$

294 where  $\boldsymbol{\delta}_i = \mathbf{Q}_i - \mathbf{O}_i$ ,  $C_{1k} \leftarrow \frac{1}{N} \sum_{i=1}^N \frac{1-r_i}{1-\pi} \delta_{ik}$ , and  $C_{2k} \leftarrow$   
295  $\frac{1}{N} \sum_{i=1}^N \frac{1-r_i}{1-\pi} \delta_{ik}^2$ . For more details about Gaussian-uniform

mixture model, please refer to [Coretto and Hennig, 2016;  
296 Lathuilière *et al.*, 2018]. For further mitigating the influence  
297 of wrong pseudo-labels, we discard samples with probability  
298 of correct labeling less than 0.5 following [Gu *et al.*, 2020],  
299 i.e., the weight of a pseudo-supervised sample is defined as,  
300

$$w_i = \begin{cases} r_i, & \text{if } r_i \geq 0.5, \\ 0, & \text{otherwise.} \end{cases} \quad (13)$$

301 With generated pseudo-labels and samples weights, our ro-  
302 bust clustering objective is defined as,  
303

$$\mathcal{L}_C = \frac{1}{\sum_{i=1}^N w_i} \sum_{i=1}^N w_i \|\mathbf{Q}_i - \mathbf{O}_i\|_2^2. \quad (14)$$

303 We adopt mean square error (MSE) as the objective to train  
304 our model with pseudo-supervised data, because MSE is  
305 more robust to label noise than the cross entropy loss in clas-  
306 sification task [Ghosh *et al.*, 2017]. Through the three steps,  
307 we can learn more discriminative representations and achieve  
308 better short text clustering performance with local data.  
309

### 3.3 Federated Cluster Center Aggregation Module

310 We then introduce the federated cluster center aggregation  
311 module. Existing short text clustering methods [Xu *et al.*,  
312 2017; Hadifar *et al.*, 2019; Rakib *et al.*, 2020; Zhang *et al.*,  
313 2021] all assume full access to data, i.e., the data is stored on  
314 a central server. However, the data may be distributed among  
315 many clients (e.g., companies), and gathering the data to a  
316 central server is not always feasible due to the privacy or com-  
317 munication concerns [McMahan *et al.*, 2017a]. Therefore, to  
318 enable collaborative learning across a variety of clients with-  
319 out sharing local raw data, we propose the federated clus-  
320 ter center aggregation module. To ensure the communica-  
321 tion efficiency, our federated learning module communicates  
322 the cluster centers rather than the model parameters during  
323 training process, inspired by [Tan *et al.*, 2022]. However,  
324 the partition of clusters is unknown in a clustering scenario,  
325 causing unavailable cluster centers. Thus, we use the lo-  
326 cally generated pseudo-labels to divide the clusters of a client.  
327 An overview of the federated learning module is shown in  
328 Fig.1(b). The server receives local centers from all clients and  
329 then averages these centers for obtaining global centers. The  
330 clients receive global centers and update their local centers  
331 by minimizing the clustering loss and the distance between  
332 global centers and local centers. We will provide the details  
333 below.

334 We assume that there are  $m$  clients, each client has  $K$  clus-  
335 ters. The sample  $i$  will be grouped into  $k$ -th cluster if the  
336  $k$ -th entry of its pseudo-label  $\mathbf{Q}_i$  is the largest. We denote  
337 the samples belonging to  $k$ -th cluster as  $S_k$ . We obtain the  
338 local cluster centers by averaging the representations of sam-  
339 ples in every cluster set. For client  $\mu$ , the center of cluster  $k$   
340 is computed as follows,

$$C_k^{(\mu)} = \frac{1}{n_k^{(\mu)}} \sum_{e \in S_k^{(\mu)}} e, \quad (15)$$

341 where  $n_k^{(\mu)}$  is the number of samples assigned to the  $k$ -th clus-  
342 ter of client  $\mu$ , i.e.,  $n_k^{(\mu)} = |S_k^{(\mu)}|$ . We obtain the global cluster  
343 centers by weighted averaging the local centers of all clients.  
344

---

**Algorithm 2** The optimization scheme of **FSTC**.

**Input:** The inputs of  $m$  clients:  $\mathbf{X}^{(\mu)}$  for  $\mu = 1, \dots, m$ .  
**ServerUpdate:**

- 1: Initialize client models with the same parameters.
- 2: Aggregate representations from  $m$  clients and perform k-means on them to initialize  $\mathbf{Q}^{(\mu)}$  for  $\mu = 1, \dots, m$ .
- 3: Initialize  $\mathbf{C}^{(\mu)}$  by Equation (15) for  $\mu = 1, \dots, m$ .
- 4: Initialize  $\tilde{\mathbf{C}}$  by Equation (16).
- 5: Initialize  $\mathbf{r}^{(\mu)} = \mathbb{1}_N$  for  $\mu = 1, \dots, m$ .
- 6: **for** each round **do**
- 7:   **for** each client  $\mu$  **do**
- 8:      $\mathbf{C}^{(\mu)} \leftarrow \text{ClientUpdate}(\mu, \tilde{\mathbf{C}})$  by Algorithm 3.
- 9:   **end for**
- 10:   Update  $\tilde{\mathbf{C}}$  by Equation (16).
- 11: **end for**
- 12: Aggregate the parameters of  $m$  client models to obtain the global model, i.e.,  $\Phi = \sum_{\mu=1}^m \alpha^{(\mu)} \Phi^{(\mu)}$  and  $F = \sum_{\mu=1}^m \alpha^{(\mu)} F^{(\mu)}$ , where  $\alpha^{(\mu)}$  is the samples proportion of client  $\mu$  in all clients.
- 13: **return**  $\Phi$  and  $F$ .

---

344 The weights are given by the number of samples assigned to  
 345 the local clusters. For cluster  $k$ , the global center is computed  
 346 as follows,

$$\tilde{\mathbf{C}}_k = \sum_{\mu=1}^m \frac{n_k^{(\mu)}}{\sum_{\eta=1}^m n_k^{(\eta)}} \mathbf{C}_k^{(\mu)}. \quad (16)$$

347 We expect the local centers  $\mathbf{C}^{(1)}, \mathbf{C}^{(2)}, \dots, \mathbf{C}^{(m)}$  to approach  
 348 global centers  $\tilde{\mathbf{C}}$  to align the local centers of all clients for ex-  
 349 changing information across clients. We achieve this aim by  
 350 the alignment loss, for client  $\mu$ , the alignment loss is defined  
 351 as follows,

$$\mathcal{L}_A = \|\mathbf{C}^{(\mu)} - \tilde{\mathbf{C}}\|^2. \quad (17)$$

## 352 4 Putting Together

353 The total loss of a client could be obtained by combining the  
 354 clustering loss and the alignment loss. That is, the loss of a  
 355 client is given as:

$$\mathcal{L} = \mathcal{L}_C + \lambda \mathcal{L}_A, \quad (18)$$

356 where  $\lambda$  is a hyper-parameter to balance the two losses. In  
 357 the end, we obtain the global model by averaging the final  
 358 client models. The cluster assignments are the column index  
 359 of the largest entry in each row of predicted scores  $\mathbf{O}$ . By  
 360 doing this, **FSTC** achieves effective and efficient short text  
 361 clustering with the raw data stored locally in multiple clients.  
 362 We show the optimization scheme of **FSTC** in Algorithm 2.

## 363 5 Experiment

364 In this section, we conduct experiments on three real-world  
 365 datasets to answer the following questions: (1) **RQ1**: How  
 366 does our approach perform compared with the federated short  
 367 text clustering baselines? (2) **RQ2**: How do the Gaussian-  
 368 uniform mixture model and the federated cluster center ag-  
 369 gregation module contribute to the performance improve-  
 370 ment? (3) **RQ3**: How does the performance of **FSTC** vary  
 371 with different values of the hyper-parameters? (4) **RQ4**:

---

**Algorithm 3** ClientUpdate( $\mu, \tilde{\mathbf{C}}$ )

- 1: **for** each local iteration **do**
- 2:   Update  $\mathbf{Q}^{(\mu)}$  by Algorithm 1 throughout the whole training iterations in a logarithmic distribution [Asano *et al.*, 2020].
- 3:   **for**  $j = 1$  to  $\tau$  **do**
- 4:     Update  $\mathbf{r}^{(\mu)}$  by Equation (9).
- 5:     Update  $\boldsymbol{\theta}^{(\mu)}$  by Equation (10),(11),(12).
- 6:   **end for**
- 7:   Update  $\mathbf{w}^{(\mu)}$  by Equation (13).
- 8:   Update  $\Phi^{(\mu)}$  and  $F^{(\mu)}$  by minimizing Equation (18).
- 9:   Update  $\mathbf{C}^{(\mu)}$  by Equation (15).
- 10: **end for**
- 11: **return**  $\mathbf{C}^{(\mu)}$

---

Dataset	#clusters	#samples	#words
AgNews	4	8,000	23
StackOverflow	20	20,000	8
Biomedical	20	20,000	13

Table 1: The statistics of the datasets. #words denotes the average word number per sample.

How dose the performance of **FSTC** vary with different numbers of clients? (5) **RQ5**: How dose the performance of **FSTC** vary with different non-IID levels?

## 372 5.1 Datasets

We conduct extensive experiments on three popularly used 376 real-world datasets. The details of each dataset are as follows. **AgNews** [Rakib *et al.*, 2020] is a subset of AG’s news 377 corpus collected by [Zhang *et al.*, 2015] which consists of 378 8,000 news titles in 4 topic categories. **StackOverflow** [Xu 379 *et al.*, 2017] consists of 20,000 question titles associated with 380 20 different tags, which is randomly selected from the challenge 381 data published in Kaggle.com<sup>1</sup>. **Biomedical** [Xu *et al.*, 382 2017] is composed of 20,000 paper titles from 20 different 383 topics and it is selected from the challenge data published in 384 BioASQ’s official website<sup>2</sup>. The detailed statistics of these 385 datasets are shown in Table 1. 386

We consider the experiments of both IID partition and non- 388 IID partition in multiple clients. **IID Partition**: The client 389 number is set to  $m = \{2, 4, 8, 10\}$ , the data is shuffled and 390 evenly partitioned into multiple clients. **Non-IID Partition**: 391 The client number is set to  $m = 2$ , the data is shuffled and 392 partitioned into 2 clients with different proportions  $\{6:4, 7:3,$  393  $8:2, 9:1\}$  which correspond to different non-IID levels  $\rho =$  394  $\{1, 2, 3, 4\}$ . 395

## 396 5.2 Evaluation Metrics

We report two widely used performance metrics of text clustering, i.e., accuracy (ACC) and normalized mutual information (NMI), following former short text clustering literatures [Xu *et al.*, 2017; Hadifar *et al.*, 2019; Zhang *et al.*, 2021].

<sup>1</sup><https://www.kaggle.com/c/predict-closed-questions-on-stackoverflow/>

<sup>2</sup><http://participants-area.bioasq.org/>

401 Accuracy is defined as:

$$ACC = \frac{\sum_{i=1}^N \mathbb{1}_{y_i=\text{map}(\hat{y}_i)}}{N}, \quad (19)$$

402 where  $y_i$  and  $\hat{y}_i$  are the ground truth label and the predicted  
403 label for a given text  $x_i$  respectively,  $\text{map}()$  maps each pre-  
404 dicted label to the corresponding target label by Hungarian  
405 algorithm[Papadimitriou and Steiglitz, 1998]. Normalized  
406 mutual information is defined as:

$$NMI(Y, \hat{Y}) = \frac{I(Y, \hat{Y})}{\sqrt{H(Y)H(\hat{Y})}}, \quad (20)$$

407 where  $Y$  and  $\hat{Y}$  are the ground truth labels and the predicted  
408 labels respectively,  $I()$  is the mutual information and  $H()$  is  
409 the entropy.

### 410 5.3 Experiment Settings

411 We build our framework with PyTorch [Paszke *et al.*,  
412 2019] and train the local models using the Adam optimizer  
413 [Kingma and Ba, 2015]. We choose distilbert-base-nli-stsb-  
414 mean-tokens in Sentence Transformer library [Reimers and  
415 Gurevych, 2019] to embed the short texts, and the maximum  
416 input length is set to 32. The learning rate is set  
417 to  $5 \times 10^{-6}$  for optimizing the embedding network, and  
418  $5 \times 10^{-4}$  for optimizing the clustering network. The dimensions  
419 of the text representations is set to  $D = 768$ . The batch size is set to  $N = 200$ . The hyper-parameter  $\epsilon$  is set  
420 to 0.1. We study the effect of hyper-parameter  $\lambda$  by varying  
421 it in  $\{0.001, 0.01, 0.1, 1, 10\}$ . The communication rounds  
422 is set to 40 and the local iterations is set to 100. Following  
423 previous short text clustering researches [Xu *et al.*, 2017;  
424 Hadifar *et al.*, 2019; Rakib *et al.*, 2020; Zhang *et al.*, 2021],  
425 we set the clustering numbers to the ground-truth category  
426 numbers. Moreover, we adopt the same augmentation strategy  
427 with [Zhang *et al.*, 2021] for achieving better representation  
428 learning.

### 430 5.4 Baselines

431 We compare our proposed approach with the following federated  
432 short text clustering baselines. **FBOW**: We apply k-  
433 FED [Dennis *et al.*, 2021] on the BOW [Scott and Matwin,  
434 1998] representations. **FTF-IDF**: We apply k-FED [Dennis  
435 *et al.*, 2021] on the TF-IDF [Salton and McGill, 1983] rep-  
436 resentations. **FSBERT**: We apply k-means on the represen-  
437 tations embedded by SBERT [Reimers and Gurevych, 2019].  
438 Note that, as the BOW representations and TF-IDF represen-  
439 tations reveal the raw texts, they cannot be transmitted to the  
440 central server and directly applying k-means. While SBERT  
441 representations do not reveal the raw texts, they can be trans-  
442 mitted to the central server and directly applying k-means.  
443 **FSCCL**: We combine FedAvg [McMahan *et al.*, 2017b] with  
444 SCCL [Zhang *et al.*, 2021]. SCCL is one of the state-of-the-  
445 art short text clustering models, it utilizes SBERT [Reimers  
446 and Gurevych, 2019] as the backbone, introduces instance-  
447 wise contrastive learning to support clustering, and uses the  
448 clustering objective proposed in [Xie *et al.*, 2016] for deep  
449 joint clustering. We use its released code<sup>3</sup> for achieving the  
450 local model.

<sup>3</sup><https://github.com/amazon-science/sccl>

### 451 5.5 Federated Clustering Performance (RQ1)

452 **Results and discussion.** The comparison results on three  
453 datasets are shown in Table 2. From them, we can find that:  
454 (1) Only adopting k-means based federated clustering with  
455 traditional text representations (**FBOW** and **FTF-IDF**) can-  
456 not obtain satisfying results due to the data sparsity prob-  
457 lem. (2) **FSBERT** outperforms the traditional text repre-  
458 sentation methods, indicating that adopting pre-trained word  
459 embeddings alleviates the sparsity problem, but the fixed  
460 SBERT without representation learning cannot obtain dis-  
461 criminative representations for clustering. (3) **FSCCL** ob-  
462 tains better clustering results by introducing instance-wise  
463 contrastive learning and utilizing the clustering objective pro-  
464 posed in [Xie *et al.*, 2016] for simultaneously representa-  
465 tion learning and clustering. Although **FSCCL** obtains dis-  
466 criminative representations by deep representation learning,  
467 it cannot learn sufficiently discriminative representations due  
468 to lacking supervision information, causing limited cluster-  
469 ing performance. (4) **FSTC** consistently achieves the best  
470 performance, which proves that the robust short text cluster-  
471 ing module with generated pseudo-labels as supervision and  
472 the federated cluster center aggregation module with efficient  
473 communications can significantly improve the federated clus-  
474 tering performance.

475 **Visualization.** To better show the discriminability of text  
476 representations, we visualize the representations using t-SNE  
477 [Van der Maaten and Hinton, 2008] for **FTF-IDF**, **FSBERT**,  
478 **FSCCL**, and **FSTC**. The results on **Stackoverflow** are shown  
479 in Fig.2(a)-(d). From them, we can see that: (1) **FTF-IDF** has  
480 no boundaries between clusters, and the points from different  
481 clusters have significant overlap. (2) Although there is less  
482 overlap in **FSBERT**, it still has no significant boundaries be-  
483 tween clusters. (3) **FSCCL** achieves clear boundaries to some  
484 extent, but there are a large proportion of points are grouped  
485 to the wrong clusters. (4) With reliable pseudo-supervised  
486 data, **FSTC** obtains best text representations with smaller  
487 intra-cluster distance, larger inter-cluster distance while more  
488 points are grouped to the correct clusters. The visualization  
489 results illustrate the validity of our **FSTC** framework.

### 490 5.6 In-depth Analysis (RQ2-RQ5)

491 **Ablation (RQ2).** To study how does each component of  
492 **FSTC** contribute on the final performance, we compare  
493 **FSTC** with its two variants, including **FSTC-STC** and  
494 **FSTC-RSTC**. **FSTC-STC** only adopts the pseudo-label gen-  
495 eration while **FSTC-RSTC** adopts the pseudo-label genera-  
496 tion and the Gaussian-uniform mixture model (i.e., the robust  
497 short text clustering module). The final global model is ob-  
498 tained by averaging the final local models. The comparison  
499 results are shown in Table 2. From it, we can observe that (1)  
500 **FSTC-STC** always get more accurate output predictions than  
501 baselines, which indicates that using generated pseudo-labels  
502 as supervision information to guide the training is essential.  
503 (2) **FSTC-RSTC** always get better results than **FSTC-STC**,  
504 indicating that the Gaussian-uniform mixture model is bene-  
505 ficial to obtain more reliable pseudo-supervised data for train-  
506 ing. (2) However, **FSTC-RSTC** still cannot achieve the best  
507 results against **FSTC**. Simply averaging the final local mod-

	AgNews		Stackoverflow		Biomedical	
	ACC	NMI	ACC	NMI	ACC	NMI
FBOW	28.14±0.89	3.26±0.65	12.32±1.23	6.37±1.53	13.92±1.68	8.53±1.81
FTF-IDF	30.58±1.35	7.48±1.81	42.66±2.72	43.79±3.57	25.87±0.77	23.86±1.42
FSBERT	65.95±0.00	31.55±0.00	60.55±0.00	51.79±0.00	39.50±0.00	32.63±0.00
FSCCL	81.17±0.11	54.78±0.18	70.45±1.84	61.87±0.71	42.10±0.72	37.40±0.18
<b>FSTC-STC</b>	84.38 ±0.92	60.75±1.60	77.90 ±1.12	67.81±0.51	45.31±0.75	38.25±0.34
<b>FSTC-RSTC</b>	84.75±0.60	61.33±0.95	78.58±0.73	68.11 ±0.25	45.64±0.58	38.36±0.20
<b>FSTC</b>	<b>85.10±0.25*</b>	<b>62.45±0.45*</b>	<b>79.70±1.13*</b>	<b>68.83±0.28*</b>	<b>46.67±0.72*</b>	<b>39.86±0.58*</b>

Table 2: Experimental results on three short text datasets when  $m = 4$ , where \* denotes a significant improvement with  $p$ -value  $< 0.01$  using the Mann-Whitney U test. For all the experiments, we repeat them five times. We bold the best result and underline the runner-up.

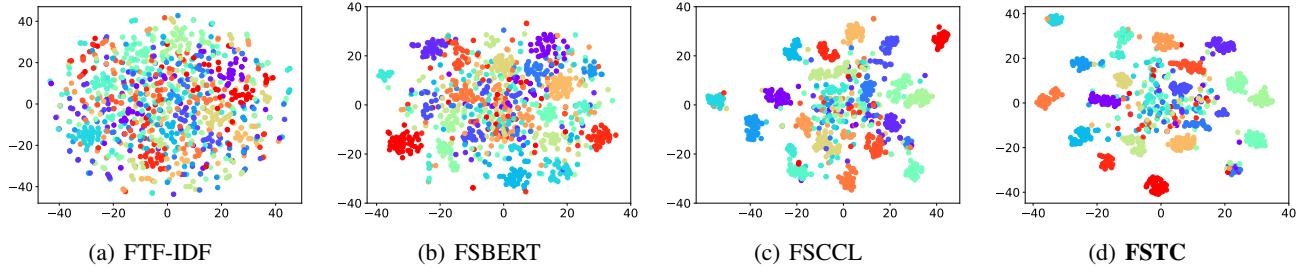


Figure 2: TSNE visualization of the representations on Stackoverflow, each color indicates a ground truth category.

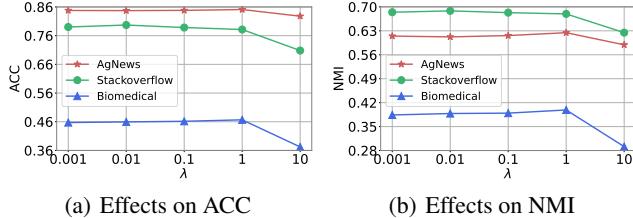


Figure 3: The effects of  $\lambda$  on model performance when  $m = 4$ .

els as the global model will sometimes cannot fully exploit all data for local training. Overall, the above ablation study demonstrates that our proposed robust short text clustering module and federated cluster center aggregation module are effective in solving the FSTC problem.

**Effect of hyper-parameters (RQ3).** We first study the effect of  $\lambda$  on model performance, where  $\lambda$  is a hyper-parameter to balance the clustering loss and the alignment loss in each client. We vary  $\lambda$  in  $\{0.001, 0.01, 0.1, 1, 10\}$  and report the results in Fig.3. Fig.3 shows that the performance first gradually increases and then decreases. It indicates that when  $\lambda$  approaches 0, the alignment loss cannot produce sufficiently positive effects. When  $\lambda$  becomes too large, the alignment loss will suppress the clustering loss, which also reduces the clustering performance. Empirically, we choose  $\lambda = 0.01$  on **Stackoverflow** while  $\lambda = 1$  on **AgNews** and **Biomedical**.

**Effect of client number (RQ4).** We also study the effect of client number  $m$  on **FSTC** and **FSCCL** by varying  $m$  in  $\{2, 4, 8, 10\}$  and report the results in Fig.4(a) on **Stackoverflow**. Fig.4(a) shows that **FSTC** keeps better performance than **FSCCL** all the time, which illustrates the validity of our framework. Besides, it is a normal phenomenon that the performance of **FSTC** declines as  $m$  increases, since the number of samples in each client decreases. The reason why the performance of **FSCCL** remains relatively stable as  $m$  increases may be that **SCCL** cannot obtain better performance

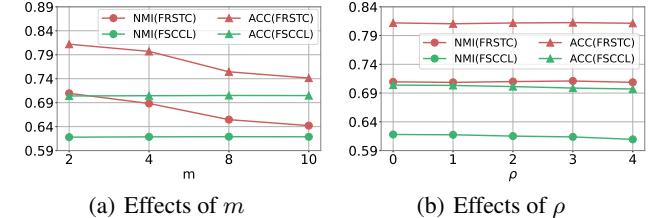


Figure 4: The effects of  $m$  and  $\rho$  on model performance.

with more samples in a client.

534

**Effect of non-IID level (RQ5).** We finally study the effect of non-IID level  $\rho$  on **FSTC** and **FSCCL** by varying  $\rho$  in  $\{0, 1, 2, 3, 4\}$  and report the results in Fig.4(b) on **Stackoverflow**, where  $\rho = 0$  denotes IID. Fig.4 shows that **FSTC** keeps better performance than **FSCCL** all the time, which illustrates the effectiveness of our framework. Besides, the performance of both models remains relatively stable as  $\rho$  increases, which indicates that our framework is robust to different non-IID levels with more efficient communications.

535

536

537

538

539

540

541

542

543

## 6 Conclusion

544

In this paper, we propose a federated robust short text clustering framework (**FSTC**), which includes the robust short text clustering module and the federated cluster center aggregation module. To our best knowledge, we are the first to address short text clustering problem in the federated setting. Moreover, we innovatively combine optimal transport to generate pseudo-labels with Gaussian-uniform mixture model to improve the reliability of the pseudo-supervised data. We also conduct extensive experiments to demonstrate the superior performance of our proposed **FSTC** on several real-world datasets.

545

546

547

548

549

550

551

552

553

554

555

## References

- [Asano *et al.*, 2020] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.
- [Bridle, 1990] John S Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *Neurocomputing*, pages 227–236. Springer, 1990.
- [Caron *et al.*, 2020] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.
- [Chung *et al.*, 2022] Jichan Chung, Kangwook Lee, and Kannan Ramchandran. Federated unsupervised clustering with generative models. In *AAAI 2022 International Workshop on Trustable, Verifiable and Auditable Federated Learning*, 2022.
- [Coretto and Hennig, 2016] Pietro Coretto and Christian Hennig. Robust improper maximum likelihood: tuning, computation, and a comparison with other methods for robust gaussian clustering. *Journal of the American Statistical Association*, 111(516):1648–1659, 2016.
- [Cuturi, 2013] Marco Cuturi. Sinkhorn distances: Light-speed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.
- [Dennis *et al.*, 2021] Don Kurian Dennis, Tian Li, and Virginia Smith. Heterogeneity for the win: One-shot federated clustering. In *International Conference on Machine Learning*, pages 2611–2620. PMLR, 2021.
- [Ghosh *et al.*, 2017] Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. Robust loss functions under label noise for deep neural networks. In Satinder Singh and Shaul Markovitch, editors, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 1919–1925. AAAI Press, 2017.
- [Ghosh *et al.*, 2020] Avishhek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. *Advances in Neural Information Processing Systems*, 33:19586–19597, 2020.
- [Gu *et al.*, 2020] Xiang Gu, Jian Sun, and Zongben Xu. Spherical space domain adaptation with robust pseudo-label loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9101–9110, 2020.
- [Hadifar *et al.*, 2019] Amir Hadifar, Lucas Sterckx, Thomas Demeester, and Chris Develder. A self-training approach for short text clustering. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 194–199, 2019.
- [Hu *et al.*, 2021] Weibo Hu, Chuan Chen, Fanghua Ye, Zibin Zheng, and Yunfei Du. Learning deep discriminative representations with pseudo supervision for image clustering. *Information Sciences*, 568:199–215, 2021.
- [Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.
- [Kumar *et al.*, 2020] Hemant H Kumar, VR Karthik, and Mydhili K Nair. Federated k-means clustering: A novel edge ai based approach for privacy preservation. In *2020 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)*, pages 52–56. IEEE, 2020.
- [Lathuilière *et al.*, 2018] Stéphane Lathuilière, Pablo Mesejo, Xavier Alameda-Pineda, and Radu Horaud. Deepgum: Learning deep robust regression with a gaussian-uniform mixture model. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part V*, volume 11209 of *Lecture Notes in Computer Science*, pages 205–221. Springer, 2018.
- [Li *et al.*, 2022] Jinning Li, Huajie Shao, Dachun Sun, Ruijie Wang, Yuchen Yan, Jinyang Li, Shengzhong Liu, Hanghang Tong, and Tarek Abdelzaher. Unsupervised belief representation learning with information-theoretic variational graph auto-encoders. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1728–1738, 2022.
- [Magdziarczyk, 2019] Małgorzata Magdziarczyk. Right to be forgotten in light of regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec. In *6th INTERNATIONAL MULTIDISCIPLINARY SCIENTIFIC CONFERENCE ON SOCIAL SCIENCES AND ART SGEM 2019*, pages 177–184, 2019.
- [McMahan *et al.*, 2017a] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In Aarti Singh and Xiaojin (Jerry) Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR, 2017.
- [McMahan *et al.*, 2017b] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

- 665 [Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai  
 666 Chen, Greg S Corrado, and Jeff Dean. Distributed rep-  
 667 resentations of words and phrases and their composition-  
 668 ality. *Advances in neural information processing systems*,  
 669 26, 2013.
- 670 [Otto, 2018] Marta Otto. Regulation (eu) 2016/679 on the  
 671 protection of natural persons with regard to the process-  
 672 ing of personal data and on the free movement of such  
 673 data (general data protection regulation–gdpr). In *Inter-  
 674 national and European Labour Law*, pages 958–981. Nomos  
 675 Verlagsgesellschaft mbH & Co. KG, 2018.
- 676 [Papadimitriou and Steiglitz, 1998] Christos H Papadim-  
 677 itriou and Kenneth Steiglitz. *Combinatorial optimization:  
 678 algorithms and complexity*. Courier Corporation, 1998.
- 679 [Paszke *et al.*, 2019] Adam Paszke, Sam Gross, Francisco  
 680 Massa, Adam Lerer, James Bradbury, Gregory Chanan,  
 681 Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca  
 682 Antiga, Alban Desmaison, Andreas Kopf, Edward Yang,  
 683 Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank  
 684 Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and  
 685 Soumith Chintala. Pytorch: An imperative style, high-  
 686 performance deep learning library. In *Advances in Neu-  
 687 ral Information Processing Systems 32*, pages 8024–8035.  
 688 Curran Associates, Inc., 2019.
- 689 [Pedrycz, 2021] Witold Pedrycz. Federated fcm: Clustering  
 690 under privacy requirements. *IEEE Transactions on Fuzzy  
 691 Systems*, 2021.
- 692 [Rakib *et al.*, 2020] Md Rashadul Hasan Rakib, Norbert  
 693 Zeh, Magdalena Jankowska, and Evangelos Milios. En-  
 694 hancement of short text clustering by iterative classifica-  
 695 tion. In *International Conference on Applications of Na-  
 696 tural Language to Information Systems*, pages 105–117.  
 697 Springer, 2020.
- 698 [Reimers and Gurevych, 2019] Nils Reimers and Iryna  
 699 Gurevych. Sentence-bert: Sentence embeddings using  
 700 siamese bert-networks. In *Proceedings of the 2019  
 701 Conference on Empirical Methods in Natural Language  
 702 Processing and the 9th International Joint Conference on  
 703 Natural Language Processing (EMNLP-IJCNLP)*, pages  
 704 3982–3992, 2019.
- 705 [Salton and McGill, 1983] Gerard Salton and Michael J  
 706 McGill. *Introduction to modern information retrieval.*  
 707 mcgraw-hill, 1983.
- 708 [Scott and Matwin, 1998] Sam Scott and Stan Matwin. Text  
 709 classification using wordnet hypernyms. In *Usage of  
 710 WordNet in natural language processing systems*, 1998.
- 711 [Stallmann and Wilbik, 2022] Morris Stallmann and Anna  
 712 Wilbik. Towards federated clustering: A federated fuzzy c-  
 713 means algorithm (FFCM). *CoRR*, abs/2201.07316, 2022.
- 714 [Stieglitz *et al.*, 2018] Stefan Stieglitz, Milad Mirbabaie,  
 715 Björn Ross, and Christoph Neuberger. Social media  
 716 analytics–challenges in topic discovery, data collection,  
 717 and data preparation. *International journal of informa-  
 718 tion management*, 39:156–168, 2018.
- 719 [Tan *et al.*, 2022] Yue Tan, Guodong Long, Lu Liu, Tianyi  
 720 Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. Fed-  
 721 proto: Federated prototype learning across heterogeneous  
 722 clients. In *AAAI Conference on Artificial Intelligence*, vol-  
 723 ume 1, page 3, 2022.
- 724 [Van der Maaten and Hinton, 2008] Laurens Van der Maaten  
 725 and Geoffrey Hinton. Visualizing data using t-sne. *Journal  
 726 of machine learning research*, 9(11), 2008.
- 727 [Wu *et al.*, 2022] Chuhan Wu, Fangzhao Wu, Yongfeng  
 728 Huang, and Xing Xie. Personalized news recomme-  
 729 dation: Methods and challenges. *ACM Transactions on In-  
 730 formation Systems (TOIS)*, 2022.
- 731 [Xie *et al.*, 2016] Junyuan Xie, Ross Girshick, and Ali  
 732 Farhadi. Unsupervised deep embedding for clustering  
 733 analysis. In *International conference on machine learn-  
 734 ing*, pages 478–487. PMLR, 2016.
- 735 [Xu *et al.*, 2017] Jiaming Xu, Bo Xu, Peng Wang, Suncong  
 736 Zheng, Guanhua Tian, Jun Zhao, and Bo Xu. Self-taught  
 737 convolutional neural networks for short text clustering.  
 738 *Neural Networks*, 88:22–31, 2017.
- 739 [Zhang *et al.*, 2015] Xiang Zhang, Junbo Zhao, and Yann  
 740 LeCun. Character-level convolutional networks for text  
 741 classification. *Advances in neural information processing  
 742 systems*, 28, 2015.
- 743 [Zhang *et al.*, 2021] Dejiao Zhang, Feng Nan, Xiaokai Wei,  
 744 Shang-Wen Li, Henghui Zhu, Kathleen R. McKeown,  
 745 Ramesh Nallapati, Andrew O. Arnold, and Bing Xi-  
 746 ang. Supporting clustering with contrastive learning. In  
 747 Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer,  
 748 Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cot-  
 749 terell, Tanmoy Chakraborty, and Yichao Zhou, editors,  
 750 *Proceedings of the 2021 Conference of the North Ameri-  
 751 can Chapter of the Association for Computational Linguis-  
 752 tics: Human Language Technologies, NAACL-HLT 2021,  
 753 Online, June 6-11, 2021*, pages 5419–5430. Association  
 754 for Computational Linguistics, 2021.