

# Fine-Grained Privacy Detection with Graph-Regularized Hierarchical Attentive Representation Learning

XIAOLIN CHEN and XUEMENG SONG, Shandong University, China

RUIYANG REN, Renmin University of China, China

LEI ZHU, Shandong Normal University, China

ZHIYONG CHENG, Qilu University of Technology (Shandong Academy of Sciences), China

LIQIANG NIE, Shandong University, China

Due to the complex and dynamic environment of social media, user generated contents (UGCs) may inadvertently leak users' personal aspects, such as the personal attributes, relationships and even the health condition, and thus place users at high privacy risks. Limited research efforts, thus far, have been dedicated to the privacy detection from users' unstructured data (i.e., UGCs). Moreover, existing efforts mainly focus on applying conventional machine learning techniques directly to traditional hand-crafted privacy-oriented features, ignoring the powerful representing capability of the advanced neural networks. In light of this, in this article, we present a fine-grained privacy detection network (GrHA) equipped with graph-regularized hierarchical attentive representation learning. In particular, the proposed GrHA explores the semantic correlations among personal aspects with graph convolutional networks to enhance the regularization for the UGC representation learning, and, hence, fulfil effective fine-grained privacy detection. Extensive experiments on a real-world dataset demonstrate the superiority of the proposed model over state-of-the-art competitors in terms of eight standard metrics. As a byproduct, we have released the codes and involved parameters to facilitate the research community.

CCS Concepts: • **Information systems** → **Retrieval tasks and goals**; • **Security and privacy** → *Privacy protections*;

Additional Key Words and Phrases: Fine-grained privacy detection, graph convolutional networks, hierarchical attention mechanism

This work is supported by the National Key Research and Development Project of New Generation Artificial Intelligence, No.:2018AAA0102502; the National Natural Science Foundation of China, No.:61702300, No.:61772310, and No.:U1936203; the Shandong Provincial Natural Science Foundation, No.:ZR2019JQ23; the Shandong Provincial Key Research and Development Program, No.:2019JZZY010118; the Innovation Teams in Colleges and Universities in Jinan, No.:2018GXRC014.

Authors' addresses: X. Chen, X. Song, and L. Nie, Shandong University, No. 72 Binhai Road, Jimo, Qingdao, Shandong Province, 266237, China; emails: {cxlicd, sxmustc, nieliqiang}@gmail.com; R. Ren, Renmin University of China, No. 59 Zhongguancun Street, Haidian District, Beijing, 100872, China; email: reyon.ren@ruc.edu.cn; L. Zhu, Shandong Normal University, No. 1 Daxue Road, Changqing District, Jinan, Shandong Province, 250358, China; email: leizhu0608@gmail.com; Z. Cheng, Qilu University of Technology (Shandong Academy of Sciences), No. 19 Keyuan Road, Lixia District, Jinan, Shandong Province, 250014, China; email: jason.zy.cheng@gmail.com.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

© 2020 Association for Computing Machinery.

1046-8188/2020/09-ART37 \$15.00

<https://doi.org/10.1145/3406109>

**ACM Reference format:**

Xiaolin Chen, Xuemeng Song, Ruiyang Ren, Lei Zhu, Zhiyong Cheng, and Liqiang Nie. 2020. Fine-Grained Privacy Detection with Graph-Regularized Hierarchical Attentive Representation Learning. *ACM Trans. Inf. Syst.* 38, 4, Article 37 (September 2020), 26 pages. <https://doi.org/10.1145/3406109>

---

**1 INTRODUCTION**

In modern society, online social media has become a popular platform for social interactions, where people can build up relationships, broadcast exciting news, and even share their personal lives. According to [30], information pertaining to users themselves accounts for up to 66% of the entire user generated contents (UGCs). However, the massive amount of personal data may put users at high privacy risks. Figure 1 illustrates two real-world UGCs that reveal the users' contact and occupation information, respectively.<sup>1</sup> Although users can manually check their historical UGCs to alleviate the problem of privacy leakage, this strategy suffers from the following two key limitations. (1) Manual check is rather time-consuming, especially for users who have contributed tremendous UGCs. And (2) even though the users manage to check their posts manually, they may not be acutely aware of the privacy leakage [71]. Therefore, automatic privacy detection on social media does merit our special attention. Notably, privacy leakage is highly subjective, as different users may have different privacy perceptions. Therefore, we focus on the potential privacy detection from the users' historical posts, and provide the detection results to users for further process.

As a matter of fact, users' growing concern regarding the privacy leakage on social media has attracted many researchers' attention [23, 73, 74, 76]. They mainly explored the well-structured data, such as users' profiles [74] and privacy settings [23]. Despite the great success achieved by these efforts, most of them overlook the unstructured-data (i.e., UGCs), whereby the information is more abundant and the privacy leakage issue is more prominent. Although some pioneer studies [13, 15, 55, 56, 73, 83] have tried to tackle the problem of privacy leakage detection from unstructured-data, they mainly resort to shallow learning techniques based on a set of hand-crafted privacy-oriented features. Moreover, they put their efforts on the coarse binary classification (i.e., private or not) of UGCs, which is suboptimal near enough to solve the practical task of privacy leakage detection. Therefore, in this work, we aim to explore the potential of incorporating deep learning techniques, which have shown compelling success in various machine learning tasks, in the context of fine-grained privacy detection.

In this work, we aim to comprehensively investigate the practical problem of fine-grained privacy detection over UGCs, where the personal aspects spanning from personal attributes (e.g., *occupation* and *age*) to life milestones (e.g., *get pregnant* and *graduation*) are detected for each user post. However, fine-grained privacy detection based on UGCs is non-trivial due to the following challenges: (1) In a sense, a UGC can be treated as a document with several sentences, each of which consists of a few words. Different sentences and even words may have different confidences pertaining to revealing the users' privacy. For example, given the tweet "*@user excuse me for disturbing you. I am a student of master degree majoring in physics, and want to go to university of Tokyo for PHD.*" it is apparent that the second sentence, especially the words "master degree," delivers more information regarding the user's education background. Therefore, how to accurately capture confidences of different words and sentences in privacy detection, and, hence, boost the performance of fine-grained privacy detection poses the main challenge for us. (2) Personal aspects are usually not independent but correlated due to their semantic correlations. For example,

---

<sup>1</sup>For the privacy concern, we replaced the sensitive information, like the names in the original tweets, with the general references such as "user1" and "XXXXX".

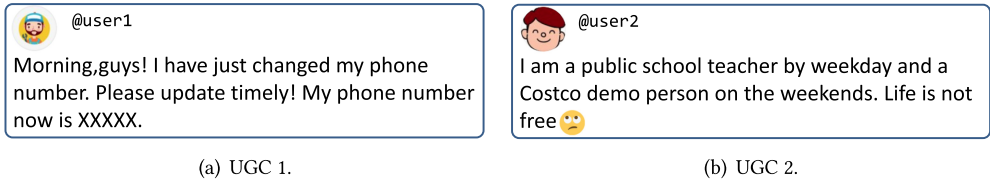


Fig. 1. Exemplars of UGCs that reveal the users’ contact and occupation information, respectively.

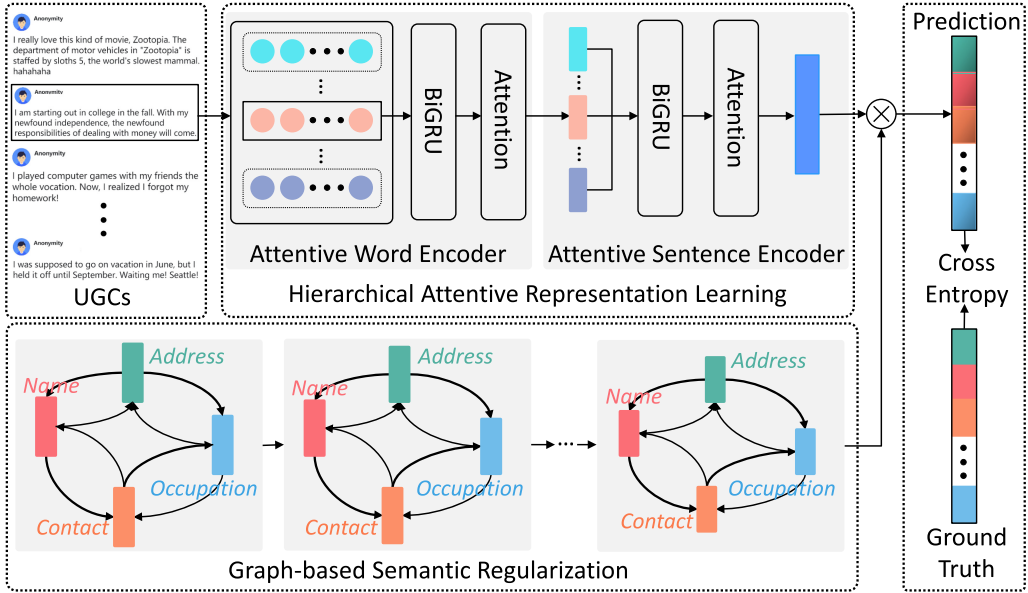


Fig. 2. Illustration of the proposed scheme for fine-grained privacy detection from UGCs. Aiming to learn that latent space that can characterize the correspondence between the UGC and its labels, namely, the personal aspects it reveals, GrHA is comprised of two key components: *hierarchical attentive representation learning* and *graph-based semantic regularization*.

given a tweet that reveals the user’s personal aspect “Health condition,” it is more likely that the tweet also indicates the user’s personal aspect “Treatment” rather than the aspect “Graduation.” Accordingly, modeling such semantic correlations among personal aspects is a tough challenge. And (3) how to utilize the semantic correlations among personal aspects to regularize the latent representation learning of UGCs and thus fulfil the fine-grained privacy detection in an end-to-end fashion is another crucial challenge.

To address the aforementioned challenges, we present a fine-grained privacy detection network with **Graph-regularized Hierarchical Attentive representation learning**, GrHA for short. As illustrated in Figure 2, GrHA aims to learn a latent space that is capable of characterizing the correspondence between the UGC and its labels, i.e., the personal aspects it reveals. In particular, the proposed GrHA consists of two key components: *hierarchical attentive representation learning* and *graph-based semantic regularization*. As for the hierarchical attentive representation learning, we devise a neural network with two layers of attention mechanisms, corresponding to distinguish the confidences of different words and sentences in privacy detection, respectively. Pertaining to the graph-based semantic regularization, we employ the Graph Convolutional Networks (GCNs) [36] to explore the semantic correlations that reside in personal aspects. Ultimately, the fine-grained

privacy detection is fulfilled by encoding the graph-based semantic regularization into the hierarchical attentive representation learning in an end-to-end manner.

Our main contributions can be summarized in threefold:

- We present a fine-grained privacy detection network, GrHA, which can seamlessly integrate the latent representation learning for UGCs and the graph-based semantic regularization in an end-to-end fashion. In addition, we introduce a hierarchical attentive network to distinguish the word-level and sentence-level confidences, and, hence, accurately capture the privacy indicators of each UGC.
- To the best of our knowledge, we are among the first to incorporate the graph-based semantic correlations among personal aspects as a regularization toward the latent representation learning for UGCs in the context of fine-grained privacy detection.
- Extensive experiments conducted on the real-world dataset demonstrate the superiority of our proposed model over the state-of-the-art methods. As a byproduct, we have released the codes and involved parameters to facilitate other researchers.<sup>2</sup>

In the remainder of this article, we briefly review the related work in Section 2. Section 3 formulates the research problem and details our proposed model. Experimental setup and result analyses are presented in Section 4. We finally conclude our work and discuss future research directions in Section 5.

## 2 RELATED WORK

Our work is related to the studies of privacy analyses, representation learning, and graph convolutional networks.

### 2.1 Privacy Analyses

In the last few years, increasing research efforts have been dedicated to the privacy analysis over social media, especially based on users' profiles [4, 14, 16, 29, 32, 48, 50, 64, 74] and users' privacy settings [23, 52, 66, 81]. For example, Song et al. [74] studied the risk of being re-identified from users' trajectory records with a human mobility dataset. Han et al. [23] studied the privacy issues in the context of people search by simulating different privacy settings in a public social network. Due to the concern of the privacy leakage over unstructured UGCs, Tran et al. [76] presented a binary classification framework with Convolutional Neural Networks (CNN) [37] to determine whether a given photo is private. Likewise, Mao et al. [55] built automatic binary classifiers to detect sensitive vacation tweets, drunk driving tweets, and disease tweets, respectively. Moreover, Li et al. [46] proposed an approach to preserve the user's privacy by explicitly obscuring the important author characteristics, while keeping the learned representations invariant to these attributes.

However, the above research efforts on privacy analyses mainly focus on the coarse-grained privacy detection, making the results less meaningful. Toward this end, Song et al. [73] proposed a taxonomy-guided multi-task learning model based on several hand-crafted privacy-oriented features to predict which personal aspects are revealed in the posts, where a comprehensive taxonomy characterizing the user's privacy is introduced. Although the pioneer studies have obtained remarkable achievements, they mainly utilize the shallow learning methods together with a set of hand-crafted privacy-oriented features. Beyond that, in this work, we focus on enhancing the performance of fine-grained privacy detection by utilizing the deep learning techniques, where the semantic correlations among personal aspects and the latent representation learning of UGCs can be explored thoroughly.

<sup>2</sup><https://github.com/Fine-grainedPrivacyDetection/GrHA/>.

## 2.2 Representation Learning

As an active research topic, representation learning has long been striving for learning more effective representations for data rather than hand-crafted features, which has achieved remarkable success in various tasks [22, 28, 34, 65, 72, 78, 84]. For example, Wang et al. [80] developed a semi-supervised algorithm for short text clustering, where the latent representations for short sentences are learned by a deep neural network model. In addition, Lai et al. [39] proposed a recurrent convolutional neural network to fulfil the task of text classification, where the contextual information is learned by a bidirectional recurrent structure. Likewise, Lee et al. [41] employed the CNN to learn the effective representation for each short text and, hence, tackled the problem of short-text classification.

Although these studies have achieved compelling success, they all suffer from the limitation of treating all the latent features or factors equally, but overlooking the different capabilities of features in representation. Toward this end, Bahdanau et al. [2] introduced the general attention mechanism working on identifying the important words from auxiliary textual information to provide more precise representations for data. Since then, many derivatives of the attention mechanism have been proposed to solve various tasks from the natural language processing domain [19, 21, 27, 68, 69, 85, 87] and the computer vision domain [61–63, 79, 90]. To be specific, in the natural language processing domain, Yin et al. [87] presented three attention schemes to incorporate the mutual influence between sentences into CNNs to learn the sentence representations. In addition, Seo et al. [69] combined the local and global attentions over review text to derive the better interpretable representations for users and items, respectively. Moreover, Yang et al. [85] introduced two levels of attention mechanisms to tackle the problem of document classification. Besides, in the computer vision domain, Peng et al. [62] proposed a modality-specific cross-modal similarity measurement approach, which constructs independent semantic spaces for different modalities, where the modality-specific characteristics can be well explored with attention mechanism. In addition, Zhao et al. [90] introduced a pyramid feature attention network for image saliency detection, which employs both the spatial attention and channel-wise attention to enhance the low-level spatial structural features and the high-level context features, respectively. What is more, Peng et al. [63] established a spatial-temporal attention model for video classification to jointly capture the video evolutions both in spatial and temporal domains.

Although the representation learning has shown remarkable performance in plenty of tasks, limited efforts have been dedicated to the task of fine-grained privacy detection. Toward this end, in this work, we employ a hierarchical attentive network to distinguish the confidences of different words and sentences in delivering users' personal aspects, and, hence, fulfil the task of fine-grained privacy detection on social media.

## 2.3 Graph Convolutional Network

As an extension of the convolutional network, graph convolutional network, introduced by Bruna et al. in [6], works on exploiting the adjacency matrix or the Laplacian matrix that characterizes the graph structure, and, hence, capturing the correlations among different nodes. Due to its powerful capability in exploring the correlation propagation between nodes, recent years have witnessed increasing research attention from both the natural language processing domain [3, 36, 42, 49] and the computer vision domain [9, 20, 25, 26] has been paid to the graph convolutional network. For example, Kipf and Welling [36] introduced a graph-based semi-supervised learning framework for node classifications, where the label information is smoothed over the graph via a Laplacian regularization term in the loss function. In addition, Peng et al. [60] proposed a graph-CNN based deep learning model for text classification, where the GCN is employed to exploit the

Table 1. Summary of the Main Notations

Notation	Explanation
$\mathcal{T}$	The set of training labeled tweets.
$\mathcal{C}$	The set of personal aspects.
$\mathbf{y}_i$	The label vector of the tweet $t_i$ .
$x_m^t$	The $t$ -th word in the $m$ -th sentence of the tweet.
$\mathbf{h}_m^t$	The hidden representation of the $t$ -th word in the $m$ -th sentence of the tweet.
$\mathbf{c}_w$	The word-level attention context vector.
$\alpha_m^t$	The confidence of the $t$ -th word in the $m$ -th sentence of the tweet.
$\mathbf{k}_m$	The attentive representation of the $m$ -th sentence of the tweet.
$\mathbf{h}_m$	The hidden representation of the $m$ -th sentence of the tweet.
$\mathbf{c}_s$	The sentence-level attention context vector.
$\beta_m$	The confidence of the $m$ -th sentence of the tweet.
$\mathbf{v}_i$	The attentive latent representation of the tweet $t_i$ .
$\mathbf{h}_j$	The latent embedding of the $j$ -th personal aspect.
$\mathbf{A}$	The predefined conditional semantic correlation matrix.
$\mathbf{z}_j$	The final representation of the $j$ -th personal aspect.
$\mathbf{Z}$	The formal latent representation of personal aspects.

word graph. Apart from the natural language processing tasks, Chen et al. [9] presented a multi-label image classification model, where the GCNs are employed to derive a set of inter-dependent object classifiers based on a directed graph with each node representing an object. In a sense, these studies have demonstrated the superiority of GCNs in learning the correlations among different nodes, which exactly inspires us to resort to GCNs for the semantic correlation modeling among different personal aspects.

### 3 METHODOLOGY

In this section, we first formulate the research problem of fine-grained privacy detection and then detail the proposed GrHA, which comprises two key components: *hierarchical attentive representation learning* and *graph-based semantic regularization*.

#### 3.1 Problem Formulation

Let us first declare some notations. In particular, we use bold uppercase letters (e.g.,  $\mathbf{X}$ ) and bold lowercase letters (e.g.,  $\mathbf{x}$ ) to represent matrices and vectors, respectively. We employ nonbold letters (e.g.,  $x$ ) to represent scalars and Greek letters (e.g.,  $\gamma$ ) as parameters. If not clarified, all vectors are in column forms.  $\|\mathbf{A}\|_F$  denotes the Frobenius norm of matrix  $\mathbf{A}$ . The main notations used in this article are summarized in Table 1.

Without loss of generality, we specifically investigate the privacy leakage of tweets in Twitter, one of the most popular social media platforms, while the cases of other social media can be explored in the same manner. In a sense, the privacy detection can be cast as a multi-label classification problem, as each tweet can reveal multiple personal aspects of the user simultaneously. Suppose we have  $N$  tweets  $\mathcal{T} = \{t_1, t_2, \dots, t_N\}$  labeled by a set of personal aspects  $\mathcal{C} = \{c_j | j = 1, 2, \dots, Q\}$ , where  $c_j$  represents the  $j$ -th personal aspect and  $Q$  is the total number of personal aspects. Let  $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N\}^T \in \mathbb{R}^{N \times Q}$  denote the corresponding label matrix, where  $\mathbf{y}_i = (y_1, y_2, \dots, y_Q) \in \{1, 0\}^Q$  represents the label vector for the  $i$ -th tweet, indicating the personal aspects revealed by the tweet.

In this work, we aim to learn an effective representation  $\mathbf{v}_i \in \mathbb{R}^D$  for the  $i$ -th tweet via a hierarchical attentive neural network, regularized by a graph-based semantic correlation modeling network, where a latent representation  $\mathbf{z}_j \in \mathbb{R}^{1 \times D}$  for each personal aspect  $c_j$  can be obtained.  $D$  is the dimension of the latent space. Based on  $\mathbf{v}_i$  and  $\mathbf{z}_j$ , we can thus measure the correspondence between the tweet  $t_i$  and the personal aspect  $c_j$ .

### 3.2 Hierarchical Attentive Representation Learning

In fact, a tweet can be treated as a document with several sentences, and each sentence comprises a sequence of words. Obviously, different sentences may play different roles in revealing the user's privacy. Moreover, even words in the same sentence can have different levels of confidences pertaining to delivering the user's personal information. For example, given the tweet "*The pictures on my wall are real. I graduated from the University of North Texas with Bachelors in Music and English,*" we can notice that the second sentence, especially the words "graduated," "university," and "bachelors," is most informative toward the privacy leakage of the user's education background. Therefore, to distinguish the privacy indicators and, hence, enhance the representation learning of tweets, we propose to utilize the hierarchical attentive neural network to encode the confidences of different words and sentences, adaptively.

In particular, suppose each tweet consists of  $M$  sentences,  $\mathcal{S} = \{s_1, s_2, \dots, s_M\}$ , and the  $m$ -th sentence  $s_m$  is comprised of  $P_m$  words,  $s_m = \{x_m^1, x_m^2, \dots, x_m^{P_m}\}$ . Following the bottom-up strategy, we first learn the representation for each sentence  $s_m$  with an attentive word encoder, and then derive that for the tweet based on an attentive sentence encoder.

**3.2.1 Attentive Word Encoder.** Due to the remarkable performance of the bidirectional gated recurrent units (BiGRU) in various natural language process tasks [2, 10, 33, 89], we employ it to encode words in the sentence  $s_m$ . One advantage of BiGRU lies in the fact that it can comprehensively summarize the sentence information from both directions by a forward GRU that reads the sentence from the word  $x_m^1$  to word  $x_m^{P_m}$ , as well as a backward GRU that scans from the word  $x_m^{P_m}$  to word  $x_m^1$ . Let  $\vec{\mathbf{h}}_m^t$  and  $\overleftarrow{\mathbf{h}}_m^t$  be the  $t$ -th hidden state of the forward GRU and the backward GRU, respectively. Here, we only briefly give the derivation of  $\vec{\mathbf{h}}_m^t$ , as  $\overleftarrow{\mathbf{h}}_m^t$  can be obtained similarly. According to BiGRU, we calculate  $\vec{\mathbf{h}}_m^t$  as follows:

$$\begin{cases} \mathbf{u}_t = \sigma(\mathbf{W}_u[\vec{\mathbf{h}}_m^{t-1}, \mathbf{e}_m^t] + \mathbf{b}_u), \\ \mathbf{r}_t = \sigma(\mathbf{W}_r[\vec{\mathbf{h}}_m^{t-1}, \mathbf{e}_m^t] + \mathbf{b}_r), \\ \mathbf{c}_t = \tanh(\mathbf{W}_c[\mathbf{r}_t \odot \vec{\mathbf{h}}_m^{t-1}, \mathbf{e}_m^t] + \mathbf{b}_c), \\ \vec{\mathbf{h}}_m^t = \mathbf{u}_t \odot \mathbf{c}_t + (1 - \mathbf{u}_t) \odot \vec{\mathbf{h}}_m^{t-1}, \end{cases} \quad (1)$$

where  $\mathbf{u}_t$  and  $\mathbf{r}_t$  are the update gate and the reset gate, respectively, and  $\mathbf{c}_t$  is the status of the memory cell.  $\mathbf{W}_u$ ,  $\mathbf{W}_r$ , and  $\mathbf{W}_c$  are the weight matrices, while  $\mathbf{b}_u$ ,  $\mathbf{b}_r$ , and  $\mathbf{b}_c$  are the bias vectors.  $\odot$  stands for the element-wise operation, and  $\sigma(\cdot)$  refers to the sigmoid activation function.  $\mathbf{e}_m^t \in \mathbb{R}^{D_e}$  is the embedding of word  $x_m^t$ , which can be pre-obtained with the help of the widely-used word2vec tool [24, 40].  $D_e$  denotes the dimension of the word embedding. Eventually, we can obtain the latent representation  $\mathbf{h}_m^t$  for the word  $x_m^t$  by concatenating  $\vec{\mathbf{h}}_m^t$  and  $\overleftarrow{\mathbf{h}}_m^t$  as follows:

$$\mathbf{h}_m^t = [\vec{\mathbf{h}}_m^t, \overleftarrow{\mathbf{h}}_m^t]. \quad (2)$$

Traditionally, BiGRU would represent the sentence  $s_m$  as the mean of all  $\mathbf{h}_m^t$ 's, which apparently overlooks the different confidence levels of words in disclosing the individual's privacy. Therefore, to distinguish informative indicators in the privacy leakage, we adopt the attention mechanism [2],

which has been proven to be effective in many machine learning tasks, such as the text summarization [70], multimedia recommendation [8], and entity representation [82]. In particular, we assign the confidences of different words as follows:

$$\begin{cases} \mathbf{u}_m^t = \tanh(\mathbf{W}_w \mathbf{h}_m^t + \mathbf{b}_w), \\ \alpha_m^t = \frac{\exp(\mathbf{c}_w^T \mathbf{u}_m^t)}{\sum_t \exp(\mathbf{c}_w^T \mathbf{u}_m^t)}, \end{cases} \quad (3)$$

where the weight matrix  $\mathbf{W}_w$  and bias vector  $\mathbf{b}_w$  are the to-be-learned layer parameters.  $\alpha_m^t$  refers to the normalized confidence of the word  $x_m^t$ , which measures the similarity between the latent representation  $\mathbf{u}_m^t$  of the word  $x_m^t$  and a context vector  $\mathbf{c}_w$ . To be specific, the word-level attention context vector  $\mathbf{c}_w$  can be treated as the latent representation of the reference query “*is it an informative word toward the privacy leakage,*” which can be automatically learned during the training process. Thereafter, we compute the final representation  $\mathbf{k}_m$  for the sentence  $s_m$  as the weighted sum of word representations  $\mathbf{h}_m^t$ 's:

$$\mathbf{k}_m = \sum_t \alpha_m^t \mathbf{h}_m^t. \quad (4)$$

**3.2.2 Attentive Sentence Encoder.** Similar to the attentive word encoder, we also employ BiGRU to encode sentences and learn the representation for the tweet as follows:

$$\vec{\mathbf{h}}_m = \vec{g}_1(\mathbf{k}_m), m \in \{1, 2, \dots, M\}, \quad (5)$$

where  $\vec{\mathbf{h}}_m$  is the  $m$ -th hidden state of the forward GRU  $\vec{g}_1$ , while  $\overleftarrow{\mathbf{h}}_m$  is that of the backward GRU  $\overleftarrow{g}_2$ . Similarly, we obtain the latent representation of the sentence  $s_m$  as  $\mathbf{h}_m$ , which can be calculated as the concatenation of the  $\vec{\mathbf{h}}_m$  and  $\overleftarrow{\mathbf{h}}_m$ , i.e.,  $\mathbf{h}_m = [\vec{\mathbf{h}}_m, \overleftarrow{\mathbf{h}}_m]$ .

To distinguish the informative sentences toward the user's privacy leakage, we adopt the attention mechanism again as follows:

$$\begin{cases} \mathbf{u}_m = \tanh(\mathbf{W}_s \mathbf{h}_m + \mathbf{b}_s), \\ \beta_m = \frac{\exp(\mathbf{c}_s^T \mathbf{u}_m)}{\sum_m \exp(\mathbf{c}_s^T \mathbf{u}_m)}, \end{cases} \quad (6)$$

where  $\beta_m$  is the normalized confidence of the sentence  $s_m$ .  $\mathbf{c}_s$  refers to the sentence-level attention context vector, representing the query “*is it an informative sentence toward the privacy leakage.*” Ultimately, we represent the  $i$ -th tweet with  $\mathbf{v}_i \in \mathbb{R}^D$  as follows:

$$\mathbf{v}_i = \sum_m \beta_m \mathbf{h}_m. \quad (7)$$

### 3.3 Graph-based Semantic Regularization

In a sense, we aim to learn a latent space, capable of characterizing the correspondence between the UGC and its labels, i.e., the personal aspects it reveals. In light of this, the UGC's labels can be employed to regularize the representation learning of each UGC. As a matter of fact, a tweet tends to reveal the user's multiple personal aspects simultaneously, due to their semantic correlations. For example, given a tweet that indicates the user's personal aspect “status change” (e.g., get married or become pregnant), it is more likely to leak the user's “gender” rather than his/her “home address.” Toward this end, we argue that the semantic correlations among personal aspects should be taken into consideration to enhance the regularization for the UGC representation learning.

To model the semantic correlations among personal aspects, a natural way is to utilize the conditional co-occurrence between personal aspects. In particular, we adopt the directed graph to characterize the conditional co-occurrence among personal aspects. As shown in Figure 3, each



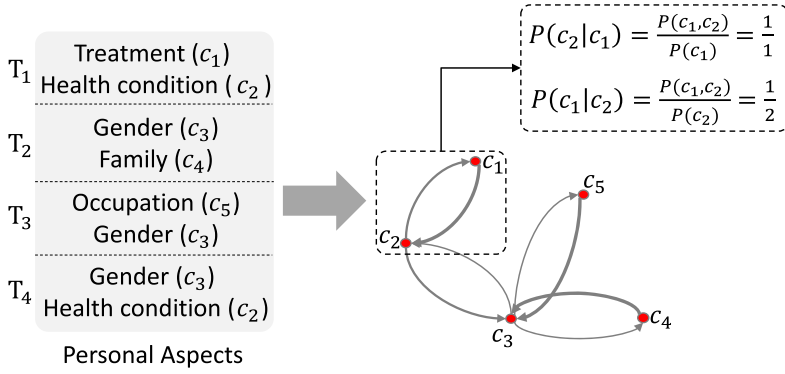


Fig. 3. Illustration of the directed graph reflecting the conditional co-occurrence between personal aspects. Each node represents a personal aspect, and the edge stands for the conditional co-occurrence probability between the two corresponding personal aspects. The thickness of the edge indicates the magnitude of the conditional probability. Taking “Treatment” ( $c_1$ ) and “Health condition” ( $c_2$ ) as an example, the edge from  $c_1$  to  $c_2$  is thicker than that from  $c_2$  to  $c_1$ , which means  $P(c_2|c_1) > P(c_1|c_2)$ . “T”: Tweet.

node of the graph refers to a personal aspect, and the edge from the node  $c_j$  to  $c_k$  reflects the conditional co-occurrence probability that a UGC will reveal the user’s personal aspect  $c_k$  if it reveals the personal aspect  $c_j$ .

In particular, we define the conditional co-occurrence probability as  $P(c_k|c_j) = \frac{P(c_k, c_j)}{P(c_j)}$ , where  $P(c_k, c_j) = \frac{n(j, k)}{N}$ , and  $n(j, k)$  denotes the number of UGCs that simultaneously reveal personal aspects  $c_j$  and  $c_k$ .  $P(c_j) = \frac{n(c_j)}{N}$  represents the probability that a UGC will reveal the personal aspect  $c_j$ , where  $n(c_j)$  is the number of UGCs labeled with the personal aspect  $c_j$ . Accordingly, we define the conditional co-occurrence adjacent matrix  $\mathbf{P} = \{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_Q\}^T \in \mathbb{R}^{Q \times Q}$ , where  $\mathbf{p}_j = (p_j^1, p_j^2, \dots, p_j^Q)$  and  $p_j^k = P(c_k|c_j)$ . Notably, the conditional co-occurrence adjacent matrix  $\mathbf{P}$  is asymmetric, as usually  $p_j^k \neq p_k^j$ . For example, as shown in Figure 3, given the four tweets that reveal five personal aspects in total, we can calculate that  $P(c_1, c_2) = 1/4$ ,  $P(c_1) = 1/4$ , and  $P(c_2) = 2/4 = 1/2$ . Accordingly, we have  $p_1^2 = 1$  and  $p_2^1 = 1/2$ , where  $p_1^2 \neq p_2^1$ .

To alleviate the noisy co-occurrence caused by the sparse real-world dataset, inspired by [9], we binarize the conditional co-occurrence adjacent matrix  $\mathbf{P}$  with a threshold  $\tau$  as follows:

$$\hat{p}_j^k = \begin{cases} 0, & \text{if } p_j^k < \tau, \\ 1, & \text{if } p_j^k \geq \tau. \end{cases} \quad (8)$$

Regarding the diagonal elements of  $\mathbf{P}$ , one naive strategy is to set  $p_j^j = 0$  to avoid the self-loops. However, in this manner, the representation of each personal aspect would be completely determined by its co-occurrence distribution with all the other personal aspects, where the own feature of each personal aspect would be neglected. This makes certain personal aspects (e.g., “Education” and “Graduation”) that share much similar co-occurrence distribution indistinguishable. Therefore, we further revise  $\hat{p}_j^k$  to  $a_j^k$  as follows:

$$a_j^k = \begin{cases} \frac{\varphi}{\sum_{x=1}^Q \hat{p}_j^x}, & \text{if } j \neq k, \\ 1 - \varphi, & \text{if } j = k, \end{cases} \quad (9)$$

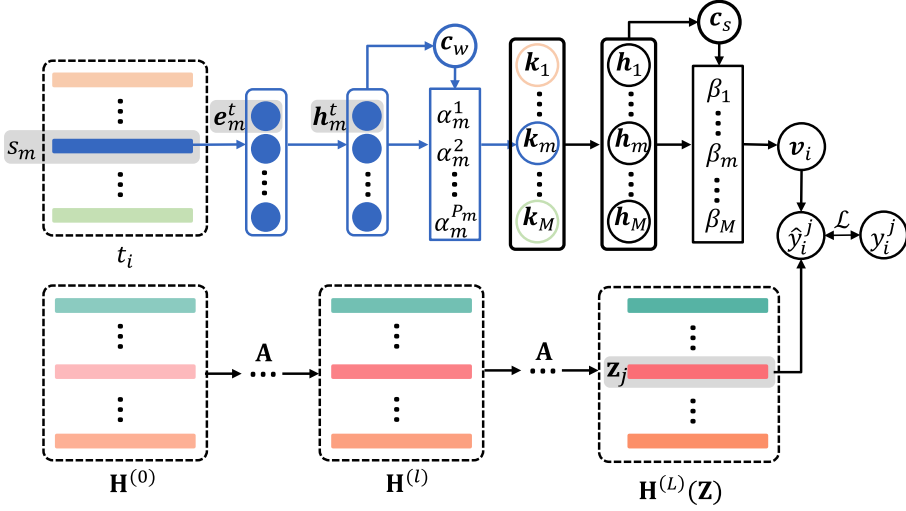


Fig. 4. Workflow of the proposed GrHA network, which consists of two key components: the hierarchical attentive representation learning and the graph-based semantic regularization.

where  $a_j^k$  denotes the conditional semantic correlation between the personal aspects  $c_k$  and  $c_j$ , and  $\varphi$  is a tradeoff parameter that determines the tradeoff between the personal aspect itself and its correlated personal aspects. In a sense, in this way, we allocate a fixed weight to each personal aspect itself but flexible weights to its correlated personal aspects based on their conditional co-occurrence distribution. In particular, when  $\varphi \rightarrow 1$ , the own feature of each personal aspect tends to be ignored, while those of its correlated personal aspects would be disabled when  $\varphi \rightarrow 0$ . Ultimately, we obtain the final conditional semantic correlation matrix  $\mathbf{A} = \{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_Q\}^T \in \mathbb{R}^{Q \times Q}$ , where  $\mathbf{a}_j = (a_j^1, a_j^2, \dots, a_j^Q)$  denotes the conditional semantic correlation vector for the  $j$ -th personal aspect.

Based on the aforementioned semantic correlation matrix, we then employ GCNs to explore the latent semantic-oriented representation for each personal aspect due to its conspicuous performance in various tasks, such as relation classification [47], text classification [36], and machine translation [3]. One advantage of GCNs is that they can update the personal aspect's representation according to the properties of its correlated personal aspects [9]. Specifically, given the conditional semantic correlation matrix  $\mathbf{A}$ , each GCN layer works as a nonlinear transformation as follows:

$$\mathbf{H}^{(l+1)} = g(\mathbf{A}\mathbf{H}^{(l)}\mathbf{W}^{(l)}), l \in \{0, 1, \dots, L-1\}, \quad (10)$$

where  $L$  refers to the total number of GCN layers, and  $\mathbf{W}^{(l)} \in \mathbb{R}^{d_l \times d_{(l+1)}}$  is the to-be-learned transformation matrix for the  $l$ -th layer.  $g(\cdot)$  stands for a non-linear operation, where we adopt the LeakyReLU [53].  $d_l$  and  $d_{(l+1)}$  are the embedding dimensions of the  $l$ -th and  $(l+1)$ -th layers, respectively.  $\mathbf{H}^{(l)} = \{\mathbf{h}_1^{(l)}, \mathbf{h}_2^{(l)}, \dots, \mathbf{h}_Q^{(l)}\}^T \in \mathbb{R}^{Q \times d_l}$ , where  $\mathbf{h}_j^{(l)}$  is the latent embedding of the  $j$ -th personal aspect at the  $l$ -th layer. In particular,  $\mathbf{h}_j^{(0)}$  is the initial embedding vector of the  $j$ -th personal aspect, which is initialized randomly and keeps updating at each following layer. Ultimately, we treat the output of the  $L$ -th layer as the final latent representation of each personal aspect, namely,  $\mathbf{z}_j = \mathbf{h}_j^{(L)} \in \mathbb{R}^{1 \times d_L}$ , where we set  $d_L = D$ . Simultaneously, we are able to obtain the formal latent representation of personal aspects, i.e.,  $\mathbf{Z} = \mathbf{H}^{(L)} = \{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_Q\}^T$ .

**ALGORITHM 1:** Fine-grained Privacy Detection Network.**Input:**  $\mathcal{T}$ ,  $Y$ ,  $A$ ,  $\lambda$ ,  $\tau$ ,  $\varphi$ .**Output:** Predicted scores of every personal aspect  $\hat{y}_i^j$ .

- 1: Initialize neural network parameters  $\Theta$ .
- 2: Initialize the latent embedding matrix of personal aspects  $\mathbf{H}^{(0)}$ .
- 3: **repeat**:
- 4:   Draw  $t_i$  from  $\mathcal{T}$ .
- 5:   Compute the latent word representations  $[\mathbf{h}_m^1, \dots, \mathbf{h}_m^{P_m}]$  according to Equations (1) and (2).
- 6:   Compute the word confidences  $[\alpha_m^1, \dots, \alpha_m^{P_m}]$  and the attentive sentence representation  $\mathbf{k}_m$  according to Equation (3).
- 7:   Compute the latent sentence representations  $[\mathbf{h}_1, \dots, \mathbf{h}_M]$  according to Equation (5).
- 8:   Compute the sentence confidences  $[\beta_1, \dots, \beta_M]$  and the attentive tweet representation  $\mathbf{v}_i$  according to Equation (6).
- 9:   Learn the latent label representation  $\mathbf{z}^j$  according to Equation (10).
- 10:   Compute the correspondence score  $\hat{y}_i^j$  between the tweet  $t_i$  and the personal aspect  $c_j$  according to Equation (11).
- 11:   Update  $\Theta$  according to Equation (12).
- 12: **until** : Objective value converges.

**3.4 Optimization**

To facilitate the end-to-end semantic regularization on the latent representation learning for UGCs, we define the corresponding score  $\hat{y}_i^j$  between the  $i$ -th UGC and the  $j$ -th personal aspect  $c_j$  below:

$$\hat{y}_i^j = \mathbf{z}_j \mathbf{v}_i. \quad (11)$$

Ultimately, adopting the cross-entropy loss [44], we thus reach the final objective function for fine-grained privacy detection as follows:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^Q \left[ -y_i^j \log(\sigma(\hat{y}_i^j)) - (1 - y_i^j) \log(1 - \sigma(\hat{y}_i^j)) \right] + \lambda \|\Theta\|_F^2, \quad (12)$$

where  $\lambda$  is the non-negative hyperparameter. The last term is designed to avoid overfitting and  $\Theta$  refers to the set of parameters (e.g.,  $\mathbf{W}_u$ ,  $\mathbf{b}_u$ ,  $\mathbf{W}_r$ ,  $\mathbf{b}_r$ ,  $\mathbf{W}_c$ ,  $\mathbf{b}_c$ ,  $\mathbf{W}_w$ ,  $\mathbf{b}_w$ ,  $\mathbf{W}_s$ ,  $\mathbf{b}_s$ , and  $\mathbf{W}^{(l)}$ ) of the proposed end-to-end network.

As to the optimization of the parameters  $\Theta$  in the proposed network, we adopt the back-propagation strategy, where the core step is to calculate the partial derivative with respect to these parameters. Here, we only introduce the calculation for  $\frac{\partial \mathcal{L}}{\partial \mathbf{W}_s}$  as an example, while the other partial derivatives can be solved in a similar fashion. We can calculate  $\frac{\partial \mathcal{L}}{\partial \mathbf{W}_s}$  as follows.

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial \mathbf{v}_i} = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^Q \left( \frac{e^{z_j \mathbf{v}_i}}{1 + e^{z_j \mathbf{v}_i}} - y_i^j \right) \mathbf{z}_j, \\ \frac{\partial \mathcal{L}}{\partial \beta_m} = \mathbf{h}_m, \\ \frac{\partial \mathcal{L}}{\partial \mathbf{u}_m} = \frac{\exp(\mathbf{c}_s^T \mathbf{u}_m) \left[ \sum_{n \neq m} \exp(\mathbf{c}_s^T \mathbf{u}_n) \mathbf{c}_s \right]}{\left[ \sum_m \exp(\mathbf{c}_s^T \mathbf{u}_m) \right]^2}. \end{cases} \quad (13)$$

As  $\frac{\partial \mathbf{u}_m}{\partial \mathbf{W}_s}$  can be derived from  $\mathbf{u}_m = \tanh(\mathbf{W}_s \mathbf{h}_m + \mathbf{b}_s)$ , we can easily access  $\frac{\partial \mathcal{L}}{\partial \mathbf{W}_s}$ . The network is optimized in mini batches by the Adam optimizer [35], which is a variant of Stochastic Gradi-

Table 2. The Distribution of Tweets with Respect to the Number of Personal Aspects with Which They Are Associated

#Personal aspects	1	2	3	4	5
#Tweets	8,546	2,215	533	65	9

ent Descent (SGD) [5] with adaptive moment estimation. The workflow of GrHA is illustrated in Figure 4, while the procedure is summarized in Algorithm 1.

## 4 EXPERIMENTS

To evaluate the proposed model, we conducted extensive experiments on the real-world dataset by answering the following research questions:

- **RQ1.** Does our GrHA outperform state-of-the-art methods?
- **RQ2.** How does the hierarchical attention mechanism affect the performance of GrHA?
- **RQ3.** What is the contribution of the graph-based semantic regularization?
- **RQ4.** How about the sensitivity of GrHA with respect to certain important hyperparameters?

### 4.1 Data Preprocessing

According to [73], it is intractable to build the large-scale dataset for fine-grained privacy detection from unstructured UGCs, which leads to the lack of public available datasets. We thus conducted our experiments only on the public real-world dataset introduced in [73], which consists of 11,368 tweets annotated with 32 personal aspects. It is worth noting that each tweet may be labeled with multiple aspects, as it may reveal more than one personal aspect of the user. Table 2 shows the distribution of tweets with the number of personal aspects they are associated with. On average, each tweet has 1.31 personal aspects.

To get a better understanding of the dataset we adopted, we listed several tweet examples in Table 3. As we can see, users’ occupations are mainly revealed by tweeting their new jobs, their thoughts about their occupations, or merely self-promotion. Users’ gender information can be embedded in their roles in relationships (e.g., daughter and girlfriend) or the distinct gender characteristic (e.g., period for women). In addition, users’ current locations are usually discussed with sharing their current feelings or the events they are joining, while users’ places-to-go can be tweeted when they are preparing for the trips, or expressing their eagerness to trips. Users may mention their age more when their birthdays are coming. Last but not least, although certain neutral statements may also talk about “career promotion,” “my home address” and other personal aspects, they are usually revealing others’ privacy or providing no detailed personal information.

To boost the performance of our model on the real-world dataset, we first sanitized the noisy tweets with the following steps: (1) We replaced the Internet slang with their corresponding formal expressions by the Internet Slang Dictionary & Translator.<sup>3</sup> (2) We performed the lemmatization using the Stanford NLP tool [54] to link word derivatives. And (3) we corrected words that contain repeated sequential letters (e.g., “coooooool” is changed to “cool”).

<sup>3</sup><http://www.noslang.com/>.

Table 3. Examples of Some Personal Aspects

Personal Aspect	Examples
Occupation	“Just got a job other at an eye laser clinic debating if I should take it.”
	“Working at plaza is gonna get me so much more money than what I get now. I’m so excited!!”
	“I used to be a swimmer...now I’m a coach. And I love torturing my kids. #evilmutantswimcoach”
Gender	“I seriously going to buy tacos, but the laziness took over. I am my father’s daughter.”
	“My girlfriend broke up with me...”
	“The worst thing you do is piss me off while I’m on my period.”
Current location	“Get to stay in Washington DC tonight...too bad I have to sleep in the airport.”
	“At the Bell Performing Arts Centre for the LTS Jazz Band Concert. #sweet”
	“She told the doctor tomorrow is my birthday. I can’t be in the hospital.”
Place to go	“In exactly one month I will be headed to the airport to depart for Cambodia... #WhatIsLife”
	“Good morning friends..preparing for my trip to Sweden..im driving to Kiruna through Riksgrnsen and Abisko to Kiruna airport..”
	“Going to SF this weekend for the Beenzino concert! I can’t wait to get my picture with.”
Age	“It’s still sinking in how next month I’ll be 30... Never married but feel damn near divorced and no kids. Wow.”
	“..when I told him I’m only 24.”
	“Can it be June? so I can be drunk off my ass in Vegas for my 21st birthday.”
Neutral statement	“Chelsea look like they got promoted last season.”
	“Do you want my home address and social security too?”

## 4.2 Experiment Settings

For a given tweet  $t_i$ , based on  $\hat{y}_i^j$ 's, we can generate a ranking list of all the personal aspects. As it is essential to position all the true personal aspects in the top places, we selected the following evaluation metrics.

**Average Precision.** Average precision assesses the overall effectiveness of the ranking list of predicted aspects, which is widely used in information retrieval systems [57].

**One-Error.** One-Error stands for the average probability that the first predicted personal aspect is not the ground truth [88].

**S@K.** S@K represents the mean probability that a correct personal aspect is captured within the top K recommended aspects. In our experiment, we set K as 1, 3, and 5, respectively.

**P@K.** P@K stands for the proportion of correct aspects among the top K recommended results. In our experiment, we set K as 1, 3, and 5, respectively.

Experimental results reported are the average values of the 10-fold cross-validation. We adopted the grid search strategy to obtain the optimal values for the regularization parameter  $\lambda$  and tradeoff

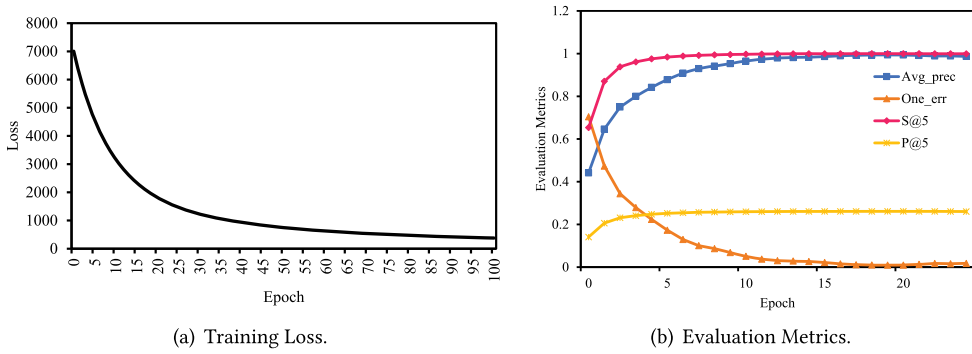


Fig. 5. Training loss and training evaluation metrics with respect to each epoch. For clarity, we only showed the  $S@K$  and  $P@K$ , where  $K = 5$ .

parameters (i.e.,  $\tau$  and  $\varphi$ ) of the semantic correlation matrix  $\mathbf{A}$  among values  $\{10^r | r \in \{-4, \dots, -1\}\}$ ,  $[0.001, 0.01, 0.1]$ , and  $[0.1, 0.9]$ , respectively. Meanwhile, the mini-batch size, the number of hidden units and the learning rate are searched in ranges of  $[30, 60, 90, 120]$ ,  $[50, 75, 100, 125]$ , and  $[0.0001, 0.001, 0.01, 0.1]$ , respectively. Moreover, we fine-tuned the proposed model with 50 epochs, and reported the performance on the testing set. In addition, we utilized TensorFlow to implement our model and all the experiments are conducted over a server equipped with an NVIDIA Titan X GPU. For optimization, we employed the adaptive moment estimation method [35] with the learning rate 0.0006.

We first experimentally demonstrated the convergence of our algorithm. Figure 5 shows the changes of the objective loss in Equation (12) and the training evaluation metrics with one run of the algorithm. As we can see, the values first change rapidly and then tend to go steady at last, which well validates the convergence of our model.

### 4.3 On Model Comparison (RQ1)

As a matter of fact, existing methods only focus on developing the hand-crafted privacy-oriented features and employ the shallow learning methods to tackle the problem of privacy detection. To validate the effectiveness of our GrHA for fine-grained privacy detection, we chose five state-of-the-art (shallow learning) methods and developed three deep learning methods as baselines.

**SVM.** The first shallow learning baseline is SVM [11], which simply concatenates the privacy-oriented features into a single vector and learns each personal aspect individually. We chose the formulation with the kernel of radial-basis function [59] and implemented this method with the help of a library for support vector machines (LIBSVM) [7].

**MTL\_Lasso.** The second baseline is the multi-task learning (MTL) with Lasso [75], which implements the  $l_1$ -penalization to the regression objective function. This method leaves out the semantic correlations among personal aspects.

**GO\_MTL.** The third baseline is the grouping and overlap in multi-task learning proposed in [38], which is able to learn the semantic correlations among personal aspects from the data.

**CMTL.** The fourth baseline is the clustered multi-task learning (CMTL) [31, 58], which assumes personal aspects can be clustered into several groups and those in one group can be learned together.

**TOKEN.** The fifth baseline is the latent group MTL [73], which utilizes the predefined personal aspect taxonomy to learn the group-sharing and aspect-specific latent features of personal aspects simultaneously.

**TextRNN.** Owing to the remarkable performance of TextRNN in text classification tasks [1, 51], we chose TextRNN [18] as one deep learning baseline. In particular, we employed Recurrent Neural Networks (RNN) to obtain the latent representations for UGCs, and based on that utilized the logistic regression [12] objective function to fulfil the task of fine-grained privacy detection.

**TextCNN.** Similarly, we selected the TextCNN model in [34] as our seventh baseline, where we employed CNN to derive the latent representations of tweets. In the same manner, we utilized the logistic regression as the loss function. Notably, both TextRNN and TextCNN overlook the semantic correlations among different personal aspects.

**D-TOKEN.** Due to the excellent performance of TOKEN reported in [73], we introduced the last end-to-end baseline D-TOKEN, which is an extension of TOKEN [73], where hand-crafted privacy-oriented features are replaced by the representation automatically learned by our hierarchical attentive network.

As five of the above baselines are shallow-learning methods, we chose the following common hand-crafted privacy-oriented features [73] for them.

- **LIWC.** Linguistic Inquiry and Word Count (LIWC) features have been widely used to characterize a given document from the content perspective [67]. The key component of LIWC is a dictionary, comprising the mappings from words to a set of predefined categories.<sup>4</sup> Given a document, LIWC generates a vector to represent the statistics of words falling into each personal aspect. Considering that the categories in the LIWC dictionary, such as “pronouns,” “job,” and “home,” provide different privacy hints, LIWC features can be adopted to capture the sensitive information of a given tweet.
- **Privacy Dictionary.** This dictionary is devised by [77] and derived from a wide range of privacy-sensitive empirical materials, offering a new linguistic resource for automated content analysis on privacy related texts. This dictionary consists of eight high-level categories: *Law*, *OpenVisible*, *OutcomeState*, *NormsRequisites*, *Restriction*, *NegativePrivacy*, *Intimacy*, and *PrivateSecret*. With the help of this dictionary, we can obtain the similar output as LIWC.
- **Sentence2Vector.** Due to the compelling success of neural networks in representation learning, considering the short-length nature of tweets, we treated each tweet as a sentence and employed the textual feature extraction tool Sentence2Vector<sup>5</sup> (i.e., a derivative of word2vec), which has been found to be sufficient to alleviate the semantic problem of word sparseness [17, 43] to generate the vector representation for each tweet.
- **Sentiment Analysis.** Since different personal aspects are frequently conveyed with different sentiments, we utilized the Stanford NLP sentiment classifier<sup>6</sup> to judge tweets’ polarity. In particular, we can assign each tweet with a value ranging from 0 to 4, corresponding to *very negative*, *negative*, *neutral*, *positive*, and *very positive* sentiment, respectively.
- **Meta-features.** Apart from the aforementioned linguistic features, we can also extract several meta-features due to the observation that tweets revealing different personal aspects may have certain special characteristics. For example, tweets describing what is happening are more likely to contain images, while tweets that reveal users’ “status change” or “friendship” may contain user mentions. Therefore, we extracted the meta-features, including the presence of hashtags, slang words, images, emojis,<sup>7</sup> user mentions, and the timestamp at hour level.

<sup>4</sup><http://www.liwc.net/>.

<sup>5</sup><https://github.com/klb3713/sentence2vec>.

<sup>6</sup><http://stanfordnlp.github.io/CoreNLP/>.

<sup>7</sup>An emoji refers to a “picture character,” which can express facial expressions, concepts, activities, and so on.

Table 4. Performance Comparison of Different Models

Models		Avg_prec	One_err	S@1	S@3	S@5	P@1	P@3	P@5
Shallow	SVM	52.91%	69.35%	30.65%	72.98%	80.47%	30.65%	26.33%	18.47%
	MTL_Lasso	58.00%	56.09%	43.91%	73.18%	82.11%	43.91%	27.38%	19.31%
	GO_MTL	58.68%	56.02%	43.98%	74.24%	83.92%	43.98%	27.65%	19.78%
	CMTL	58.99%	55.84%	44.16%	74.41%	83.30%	44.16%	27.81%	19.63%
	TOKEN	59.05%	55.96%	44.04%	74.72%	84.34%	44.04%	27.96%	19.92%
Deep	TextRNN	61.84%	49.61%	50.39%	74.50%	82.64%	50.39%	28.67%	19.90%
	TextCNN	69.31%	39.99%	60.01%	81.97%	88.40%	60.01%	31.77%	21.44%
	D-TOKEN	69.43%	39.96%	60.04%	83.39%	89.35%	60.04%	32.15%	21.56%
	<b>GrHA</b>	<b>71.44%*</b>	<b>39.17%*</b>	<b>60.83%*</b>	<b>85.83%*</b>	<b>92.20%*</b>	<b>60.83%*</b>	<b>33.51%*</b>	<b>22.60%*</b>


The symbol \* denotes that the performance improvement of our model is statistically significant with  $p < 0.01$  compared against all the baselines.

Table 4 shows the evaluation results of different methods with respect to various metrics. To validate the performance improvement is significant, we also conducted a pairwise significance test (i.e., t-test) between our proposed GrHA and the best baseline for each evaluation metric. Based on this table, we have the following observations.

- GrHA consistently outperforms both shallow learning and deep learning baselines across different evaluation metrics with all p-values  $p < 0.01$ , which indicates that GrHA can significantly improve the performance of fine-grained privacy detection over state-of-the-art methods. In a sense, this suggests that it is reasonable to incorporate the hierarchical attentive representation learning for UGCs as well as the graph-based semantic regularization toward fine-grained privacy detection.
- Deep learning methods (i.e., GrHA, D-TOKEN, TextCNN, and TextRNN) achieve better performance than shallowing learning methods do (i.e., SVM, MTL\_Lasso, GO\_MTL, CMTL, and TOKEN). In particular, we noticed that D-TOKEN exceeds TOKEN. This confirms the benefit of adopting the deep neural networks in the context of privacy detection from UGCs rather than the hand-crafted privacy-oriented features.
- GrHA surpasses all the deep learning baselines, namely, TextCNN, TextRNN and D-TOKEN, indicating the advantage of incorporating the hierarchical attentive representation learning for UGCs as well as the graph-based semantic regularization. One plausible explanation is that personal aspect representations learned by GCNs are more capable of capturing the semantic correlations among personal aspects, providing more accurate regularization over the latent representation learning of UGCs, and thus achieving more effective fine-grained privacy detection.
- Among all the shallow learning methods that adopt the hand-crafted privacy-oriented features, the single task learning SVM achieves the worst performance, which implies the correlations among personal aspects do exist. In addition, GO\_MTL, CMTL, and TOKEN show superiority over MTL\_Lasso, suggesting the semantic correlations among personal aspects do need to be taken into account in the context of fine-grained privacy detection.

Moreover, to demonstrate the practical value of our proposed GrHA, we applied it to the historical posts of a user, which consists of 118 tweets, and automatically generated a temporal privacy-aware profile, as shown in Figure 6. For clarity, we only presented the entries that are correctly captured by our method.





**Relation Status**

- My coworker is mad at me for not telling him I have a boyfriend because he wanted to set me up with his son.
- Having my brother and my boyfriend be really good friends make some awkward conversations.

**Education**

- Having a mid December birthday sucks in college because it's always during finals.
- I've cried about my tuition being \$17,000 and a lot of people pay over twice that much.
- Now that I have bands my hair is just a throwback to sophomore year.

**Family association**

- My sister and I both stayed close to home for college and my brothers like "Seeya I'm going to North Carolina."
- I'm texting my mom quotes from my textbook about miscommunication between husbands and wives.

**Outside**

- Sorry professor I can't come to class anymore because I'm auditioning for Minnesota Brass.
- My audition went terrible but hey ! I DIDN'T GET CUT!!!(Yet)

Fig. 6. Illustration of the privacy-aware profiling.

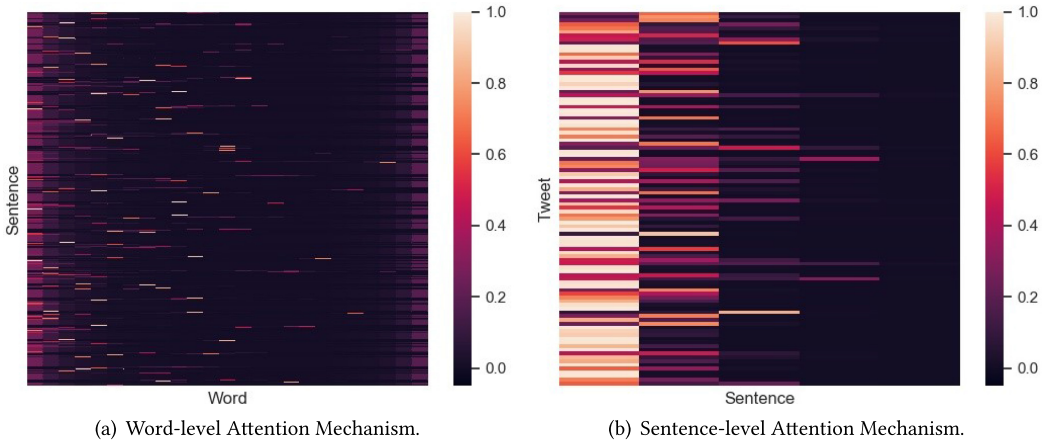


Fig. 7. Illustration of the hierarchical attention mechanism results.

#### 4.4 On Hierarchical Attention Mechanism (RQ2)

To get a deep understanding of the proposed GrHA toward fine-grained privacy detection, we particularly analyzed its one essential component: the hierarchical attention mechanism, which consists of two levels (i.e., word-level and sentence-level) of attention mechanisms.

We first intuitively illustrated the macro word/sentence confidences assigned by the hierarchical attention mechanism for testing samples with a thermodynamic diagram in Figure 7. The lighter the color, the higher the weight assigned to the word/sentence. As can be seen from Figure 7(a), the word-level attention mechanism does assign different levels of confidences to different words of a sentence, while the similar observation regarding the sentence-level attention mechanism can be found in Figure 7(b). This generally suggests that both levels of attention mechanisms contribute to the fine-grained privacy detection, especially the privacy indicator learning.

Then, we studied the micro working principle of the hierarchical attention mechanism, we showed the experimental results on the confidence assignment with several tweet samples in Figure 8. To be specific, we can extract the confidences of word-level and sentence-level according to Equations (3) and (6) on the basis of TensorFlow framework, respectively. The depth of the orange/blue bar stands for the confidence of the word/sentence learned by the attention mechanism, where the darker color refers to the larger attention weight. As we can see, given the tweet1

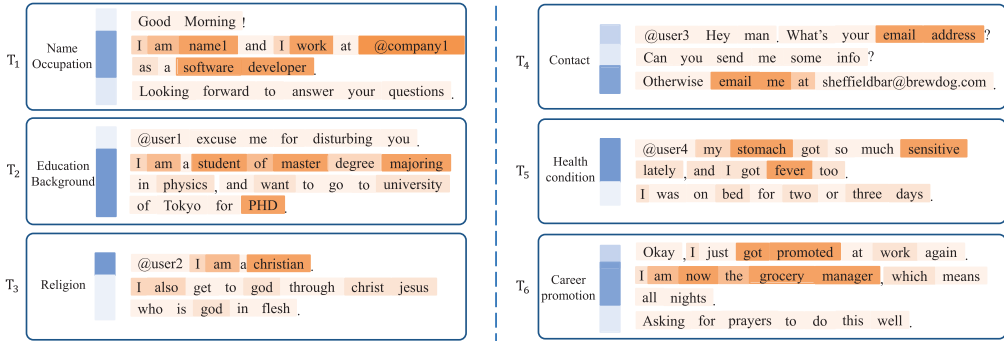


Fig. 8. Visualization of both word-level and sentence-level attention mechanisms. The color depth of the orange/blue bar stands for the confidence of the word/sentence learned by the attention mechanism. The darker color refers to the larger attention weight. “T”: Tweet.

Table 5. Effects of the Word-Level and Sentence-Level Attention Mechanisms

Model	Attention		Avg_prec	One_err	S@1	S@3	S@5	P@1	P@3	P@5
	Word	Sentence								
GrHA	No	No	68.33%	43.72%	56.28%	83.43%	90.04%	56.28%	32.40%	22.03%
	No	Yes	68.79%	43.30%	56.70%	83.85%	90.83%	56.70%	32.65%	22.18%
	Yes	No	70.04%	40.78%	59.22%	84.29%	90.98%	59.22%	32.72%	22.15%
	Yes	Yes	<b>71.44%</b>	<b>39.17%</b>	<b>60.83%</b>	<b>85.83%</b>	<b>92.20%</b>	<b>60.83%</b>	<b>33.51%</b>	<b>22.60%</b>

“Yes”/“No” refers to the presence/absence of the corresponding attention mechanism.

that leaks the user’s name and occupation, our model does identify the informative keywords, such as the user name “name1” and occupation-oriented indicators: “work,” “@company1,” “software,” and “developer.” Meanwhile, we noticed that our model assigns the highest confidence to the second sentence that contains the most of the indicator words of the given tweet, which is obviously reasonable. In fact, similar observations can be derived from other examples in Figure 8, which intuitively verifies the necessity of incorporating the hierarchical attention mechanism in the context of fine-grained privacy detection and the complementary relation between the word-level and sentence-level attention mechanisms.

Apart from the macro and micro qualitative illustration, we also quantitatively compared GrHA with its three derivatives, where different combinations of attention mechanisms at different levels are disabled by assigning the same weights to all hidden states of the corresponding BiGRU.

Table 5 shows the effects of the word-level and sentence-level attention mechanisms in our model. As can be seen, disabling the attention mechanism at any level hurts the performance of GrHA to a certain extent, and removing both levels of attention mechanisms results in the worst performance. This implies that the attention mechanisms at two levels complement each other and both contribute to the tweet representation learning in the context of fine-grained privacy detection. This may be attributed to the fact that the sentence-level attention mechanism can capture the global privacy leakage, while the word-level attention mechanism can distinguish the local privacy indicator. In addition, interestingly, we found that removing the word-level attention mechanism from GrHA devastates the performance more compared with removing the sentence-level attention mechanism. This suggests that the word-level attention mechanism contributes more to the fine-grained privacy detection, as compared to the sentence-level attention mechanism. One













ID	Personal Aspect	Content	Detection
1	Birthday	I told zari my birthday tomorrow, she gon say "aww the day I gave birth to you".	GrHA  GrHA-NoHA 
2	Occupation	Aye @user1's gonna be doing a show in Orlando while I am working at Disney! @user2 do not worry I got ur back again.	GrHA  GrHA-NoHA 
3	Contact	@user3. That I will send you the details of the fund so i want you to reply me to this email address. (XXXXX@gmail.com )	GrHA  GrHA-NoHA 
4	Salary	I earn \$2,500,000 a year with wordpress. Go quit your jobs and ask for fair compensation now . Talk pay!	GrHA  GrHA-NoHA 
5	Neutral	@user4. Please feel free Im usually around on tumblr or I can give you my email address .	GrHA  GrHA-NoHA 
6	Education	If I do not get above a C on my geometry final, I am dropping out and working at mcdonald's until I am of then becoming a stripper .	GrHA  GrHA-NoHA 

Fig. 9. Comparison between GrHA and GrHA-NoHA on several testing tweets. For the privacy concern, we replaced the sensitive information with the general symbols, such as “user1” and “XXXXX”. In addition, we represented the correct judgment of the model with the green circle and the wrong one with the red cross.

possible explanation is that words are usually more concise and discriminative than sentences in characterizing the privacy information revealed by the UGC.

To acquire a deeper understanding of the hierarchical attention mechanism, we also provided the concrete comparison between our GrHA and its derivative, GrHA-NoHA, where the hierarchical attention mechanism is completely removed, with several testing samples. As we can see from Figure 9, GrHA usually outperforms GrHA-NoHA in cases where there are the privacy indicator keywords, like the “birthday,” “email address,” and “earn,” and GrHA does distinguish them by the hierarchical attentive representation learning network. Nevertheless, this property cannot avoid certain failing cases for GrHA, especially when the tweet does contain privacy indicators but it reveals other people’s personal aspects rather than the user’s, which, hence, belongs to the neutral statement in our context. Overall, the observations from Figure 9 intuitively show the advantage of GrHA over GrHA-NoHA in the fine-grained privacy detection.

#### 4.5 On Graph-Based Semantic Regularization (RQ3)

To mine the deep insight into the graph-based semantic correlations among personal aspects, we first visualized the probability adjacency matrix based on the training data in Figure 10, where, for neatness, we only selected a few representative personal aspects. As can be seen from Figure 10, each node in the graph represents a specific personal aspect and the edge between two personal aspects stands for the conditional probability. In addition, the thickness of the edge reflects the magnitude of the corresponding conditional probability. For example, the edge from the personal aspect “Salary” to “Occupation” is thicker than that from “Salary” to “Graduation”, indicating that when a tweet reveals the user’s “Salary,” the personal aspect “Occupation” is more likely to be disclosed at the same time rather than the “Graduation.” Meanwhile, we noticed that personal aspects “Age” and “Birthday” are highly correlated, and so are “Education” and “Graduation,” which is reasonable according to common sense. These observations show that personal aspects are indeed not independent but correlated due to their semantic connections.

Then, to thoroughly verify the effectiveness of the graph-based semantic regularization, we adopted two settings: the whole GrHA and its derivative GrHA-NoHA, whose hierarchical attention mechanism is disabled. Accordingly, we introduced two derivatives of them: GrHA-NoGr and GrHA-NoHA-NoGr, respectively, where the graph-based semantic regularization of the

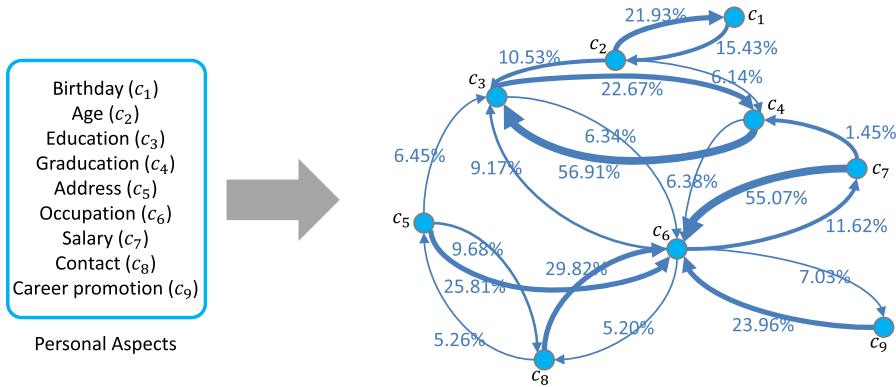


Fig. 10. Visualization of the probability adjacency matrix of selected personal aspects based on the training data. Each node stands for a specific personal aspect, and the edge between two nodes reflects the conditional probability between the two corresponding personal aspects.

Table 6. Performance Comparison of Our Proposed Network with Its Derivatives Excluding the Semantic Correlations among Personal Aspects

Models	Avg_prec	One_err	S@1	S@3	S@5	P@1	P@3	P@5
GrHA-NoHA-NoGr	65.39%	45.27%	54.73%	78.39%	85.77%	54.73%	30.37%	20.74%
GrHA-NoHA	68.33%	43.72%	56.28%	83.43%	90.04%	56.28%	32.40%	22.03%
GrHA-NoGr	67.84%	42.73%	57.27%	81.28%	88.27%	57.27%	31.50%	21.48%
GrHA	<b>71.44%</b>	<b>39.17%</b>	<b>60.83%</b>	<b>85.83%</b>	<b>92.20%</b>	<b>60.83%</b>	<b>33.51%</b>	<b>22.60%</b>

corresponding method is removed. Table 6 shows the effects of the semantic correlation regularization in GrHA-NoHA and GrHA, respectively. As can be seen, both GrHA and GrHA-NoHA show superiority over their derivatives, i.e., GrHA-NoGr and GrHA-NoHA-NoGr, respectively, indicating that the semantic regularization over the representation learning for UGCs does enhance the model performance and should be taken into account in the context of fine-grained privacy detection. Meanwhile, as a by-product, we found that GrHA significantly outperforms both GrHA-NoGr and GrHA-NoHA, while GrHA-NoHA-NoGr presents the worst performance. This enables us to draw the conclusion that the two essential components of GrHA (i.e., the hierarchical attention mechanism and graph-based semantic regularization) complement each other and both are crucial in the fine-grained privacy detection.

Apart from the quantitative evaluation, we also studied the success and failure cases for GrHA and GrHA-NoGr. As can be seen from Figure 11, GrHA usually exceeds GrHA-NoGr in cases where the given tweet reveals more than one personal aspect. For example, we found that the user who posts the first tweet tends to leak his/her “Contact” and “Current location” simultaneously, and this privacy leakage is correctly identified by GrHA rather than GrHA-NoGr. One possible explanation is that there is an obvious semantic correlation between personal aspects “Contact” and “Current location,” and this can be captured by GrHA rather than GrHA-NoGr. In a sense, the capability of capturing the semantic correlations among personal aspects contributes to the better performance of GrHA in fine-grained privacy detection. Unfortunately, GrHA can also yield several failure tweets, especially when the personal aspects revealed simultaneously in one tweet have no obvious semantic correlation. As can be seen from Figure 11, GrHA fails to identify the privacy leakage of the last tweet example that reveals the users’ personal aspects “negative emotion” and “career promotion,” the aspects of which seldom occur simultaneously in one tweet.

ID	Personal Aspect	Content	Detection
1	Contact Current location	@user1. Can you send someone near to Stanley shop? I can hand over the food items. Contact me – XXXXX.	GrHA <span style="color: green;">○</span> GrHA-NoGr <span style="color: red;">✘</span>
2	Activities at work Positive emotion	I love working at New Seasons! I get so much free stuff I never have to grocery shop!	GrHA <span style="color: green;">○</span> GrHA-NoGr <span style="color: red;">✘</span>
3	Positive emotion	Here' s another reason I love twitter. I am about to tweet the author who is a professor of economics at u Michigan.	GrHA <span style="color: green;">○</span> GrHA-NoGr <span style="color: red;">✘</span>
4	Friendship Positive emotion	Getting woken up by one of your besties texting you from Mexico that she just got engaged on the beach. So Fricken excited for them !	GrHA <span style="color: green;">○</span> GrHA-NoGr <span style="color: red;">✘</span>
5	Gender Place to go	Driving to Chicago while being on my period is not my cup of tea.	GrHA <span style="color: red;">✘</span> GrHA-NoGr <span style="color: green;">○</span>
6	Negative emotion Career promotion	I just got promoted and love doing grand openings, but my home store is just so stressful and I am miserable.	GrHA <span style="color: red;">✘</span> GrHA-NoGr <span style="color: green;">○</span>

Fig. 11. Comparison between GrHA and GrHA-NoGr on several testing tweets. For the privacy concern, we replaced the sensitive information with general symbols such as “user1” and “XXXXX”. In addition, we represented the correct judgment of the model with the green circle and the wrong one with the red cross.

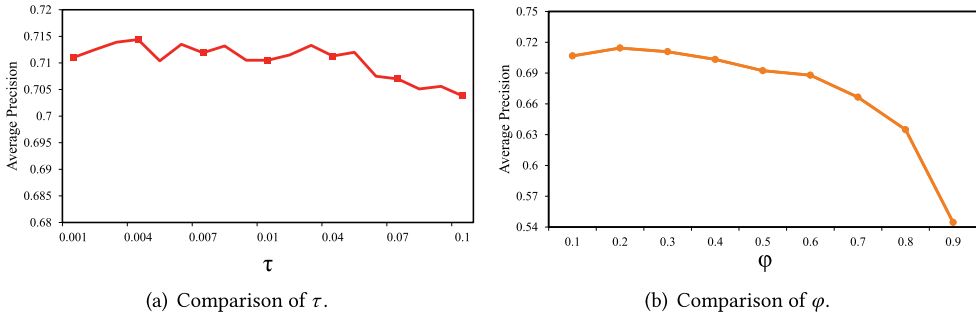


Fig. 12. Average precision comparison with different values of  $\tau$  and  $\phi$ .

#### 4.6 On Sensitivity Analysis (RQ4)

In this part, we performed the sensitivity analysis of the proposed GrHA, where we particularly studied the effects of the threshold parameter  $\tau$ , the tradeoff parameter  $\phi$ , and the depth of GCNs.

**Effect of the threshold parameter  $\tau$ .** Figure 12(a) shows the average precision of our GrHA with respect to the threshold parameter  $\tau$  for the conditional co-occurrence matrix binarization in Equation (8). In particular, we changed the value of  $\tau$  within the ranges of  $[0.001, 0.009]$  and  $[0.01, 0.1]$  with the steps of 0.001 and 0.01, respectively. As we can see, when  $\tau$  ranges from 0.001 to 0.004, the performance of our proposed GrHA keeps increasing and achieves the optimal performance at  $\tau = 0.004$ . In addition, we noticed that when  $\tau$  approaches 0.1 from 0.05, the performance of GrHA drops rapidly. This suggests that it is reasonable to filter out the edges with too small co-occurring probabilities (i.e.,  $0.001 \leq \tau \leq 0.003$ ), but inadvisable to leave out too many edges (i.e.,  $\tau > 0.05$ ).

**Effect of the tradeoff parameter  $\phi$ .** To explore the effect of  $\phi$  in controlling the balance between the personal aspects themselves and their neighborhood in learning the personal aspect representations (see Equation (9)), we conducted the sensitivity analysis of GrHA with respect to  $\phi$  by varying the value of  $\phi$  within the range of  $[0.1, 0.9]$  with a step of 0.1. As can be seen from Figure 12(b), GrHA obtains the best performance at  $\phi = 0.2$ , while suffering from a dramatic drop

Table 7. Performance Comparison with Different Depths of GCNs in Our Model

Model	Layer	Avg_prec	One_err	S@1	S@3	S@5	P@1	P@3	P@5
GrHA	1-Layer	68.65%	41.24%	58.76%	81.90%	89.28%	58.76%	31.78%	21.60%
	2-Layer	<b>71.44%</b>	<b>39.17%</b>	<b>60.83%</b>	<b>85.83%</b>	<b>92.20%</b>	<b>60.83%</b>	<b>33.51%</b>	<b>22.60%</b>
	3-Layer	68.30%	43.27%	56.73%	82.63%	89.66%	56.73%	31.84%	21.79%
	4-Layer	61.96%	49.77%	50.23%	75.92%	84.16%	50.23%	28.70%	20.01%

in performance when  $\varphi$  increases from 0.6 to 0.9. According to Equation (9), the own features of personal aspects would not be considered when  $\varphi \rightarrow 1$ , while those of the correlated personal aspects would be ignored when  $\varphi \rightarrow 0$ . Therefore, the observation allows us to conclude that it is important to balance the weights between the own features of personal aspects and those of their neighbors in the personal aspect representation learning. Especially, it is inadvisable to set  $\varphi \rightarrow 1$ , i.e., the diagonal element  $p_j^j \rightarrow 0$ , where the own features of personal aspects tend to be totally discarded resulting in the worst performance of GrHA.

**Effect of the depth of GCNs for semantic regularization.** As the depth of GCNs may affect the semantic correlation modeling and, hence, affect the graph-based semantic regularization, we also performed the corresponding sensitivity analysis of our model. Table 7 shows the performance comparison of our model with different numbers of GCN layers. As we can see, our model obtains the optimal performance when we chose GCNs with two layers, which is similar to what is reported in [9, 36, 45, 86]. In addition, we noticed that when the number of layers is larger than two, with the increasing number of layers, the performance of our proposed GrHA keeps decreasing. One possible explanation is that when using more layers for GCNs, the semantic propagation among personal aspects would be accumulated, which may result in overfitting and thus hurt the performance [9, 45].

## 5 CONCLUSION AND FUTURE WORK

In this work, to tackle the practical problem of fine-grained privacy detection, we present a graph-regularized hierarchical attentive representation learning network, termed GrHA. In particular, the proposed GrHA consists of two essential components: *hierarchical attentive representation learning* and *graph-based semantic regularization*. As for the hierarchical attentive representation learning, we introduce a hierarchical attentive network to distinguish the privacy indicators, and, hence, obtain the accurate representations for UGCs. Pertaining to the graph-based semantic regularization, we employ the GCNs to explore the semantic correlations that reside in personal aspects. Extensive experiments on a real-world dataset well validate our proposed GrHA and demonstrate the necessity of integrating both the *hierarchical attentive representation learning* and *graph-based semantic regularization* in the context of fine-grained privacy detection. Interestingly, we find that different words/sentences do have different confidences in revealing the users' privacy, and the word-level attention mechanism contributes more to the privacy detection compared to the sentence-level one.

In this work, we mainly focus on the potential privacy detection from the users' historical posts, but ignore the factor of users' social connections. In the future, we plan to investigate the second-order privacy leakage on UGCs.

## REFERENCES

- [1] Garen Arevian. 2007. Recurrent neural networks for robust real-world text classification. In *IEEE / WIC / ACM International Conference on Web Intelligence*. IEEE Computer Society, 326–329.

- [2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*.
- [3] Joost Bastings, Ivan Titov, Wilker Aziz, Diego Marcheggiani, and Khalil Sima'an. 2017. Graph convolutional encoders for syntax-aware neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1957–1967.
- [4] Joanna Asia Biega, Rishiraj Saha Roy, and Gerhard Weikum. 2017. Privacy through Solidarity: A user-utility-preserving framework to counter profiling. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 675–684.
- [5] Léon Bottou. 2010. Large-scale machine learning with stochastic gradient descent. In *Proceedings of the International Conference on Computational Statistics*. Springer, 177–186.
- [6] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. 2014. Spectral networks and locally connected networks on graphs. In *Proceedings of the International Conference on Learning Representations*.
- [7] Chih Chung Chang and Chih Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Intelligent Systems and Technology, TIST 2*, 3 (2011), 27:1–27:27.
- [8] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive collaborative filtering: Multimedia recommendation with item- and component-level attention. In *Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 335–344.
- [9] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. 2019. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Computer Vision Foundation / IEEE, 5177–5186.
- [10] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR abs/1412.3555* (2014).
- [11] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning* 20, 3 (1995), 273–297.
- [12] David R. Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)* 20, 2 (1958), 215–232.
- [13] Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11–21.
- [14] Sedigheh Eslami, Asia J. Biega, Rishiraj Saha Roy, and Gerhard Weikum. 2017. Privacy of hidden profiles: Utility-preserving profile removal in online forums. In *Proceedings of the ACM on Conference on Information and Knowledge Management*. ACM, 2063–2066.
- [15] Natasha Fernandes, Mark Dras, and Annabelle McIver. 2019. Generalised differential privacy for text document processing. In *Proceedings of the International Conference Principles of Security and Trust*. Springer, 123–148.
- [16] Casey Fiesler, Michaelanne Dye, Jessica L. Feuston, Chaya Hiruncharoenvate, Clayton J. Hutto, Shannon Morrison, Parisa Khanipour Roshan, Umashanthi Pavalanathan, Amy S. Bruckman, Munmun De Choudhury, and Eric Gilbert. 2017. What (or who) is public?: Privacy settings and social media content sharing. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 567–580.
- [17] Debasis Ganguly, Dwaipayan Roy, Mandar Mitra, and Gareth J. F. Jones. 2015. Word embedding based generalized language model for information retrieval. In *Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9–13, 2015*. ACM, 795–798.
- [18] C. Lee Giles, Gary M. Kuhn, and Ronald J. Williams. 1994. Dynamic recurrent neural networks: Theory and applications. *IEEE Transactions on Neural Network Learning Systems* 5, 2 (1994), 153–156.
- [19] Xinyu Guan, Zhiyong Cheng, Xiangnan He, Yongfeng Zhang, Zhibo Zhu, Qinke Peng, and Tat-Seng Chua. 2019. Attentive aspect modeling for review-aware recommendation. *ACM Transactions on Information Systems* 37, 3 (2019), 28:1–28:27.
- [20] Longteng Guo, Jing Liu, Jinhui Tang, Jiangwei Li, Wei Luo, and Hanqing Lu. 2019. Aligning linguistic words and visual semantic units for image captioning. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 765–773.
- [21] Yangyang Guo, Zhiyong Cheng, Liqiang Nie, Yinglong Wang, Jun Ma, and Mohan S. Kankanhalli. 2019. Attentive long short-term preference modeling for personalized product search. *ACM Transactions on Information Systems* 37, 2 (2019), 19:1–19:27.
- [22] Christophe Van Gysel, Maarten de Rijke, and Evangelos Kanoulas. 2018. Neural vector spaces for unsupervised information retrieval. *ACM Transactions on Information Systems* 36, 4 (2018), 38:1–38:25.
- [23] Shuguang Han, Daqing He, and Zhen Yue. 2014. Benchmarking the privacy-preserving people search. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 13–18.
- [24] Tianyi Hao and Longbo Huang. 2018. A social interaction activity based time-varying user vectorization method for online social networks. In *Proceedings of the International Joint Conference on Artificial Intelligence*. AAAI Press, 3790–3796.

- [25] Tao He and Xiaoming Jin. 2019. Image emotion distribution learning with graph convolutional networks. In *Proceedings of the International Conference on Multimedia Retrieval*. ACM, 382–390.
- [26] Mikael Henaff, Joan Bruna, and Yann LeCun. 2015. Deep convolutional networks on graph-structured data. *CoRR* abs/1506.05163 (2015).
- [27] Heyan Huang, Xiaochi Wei, Liqiang Nie, Xianling Mao, and Xin-Shun Xu. 2019. From question to text: Question-oriented feature attention for answer selection. *ACM Transactions on Information Systems* 37, 1 (2019), 6:1–6:33.
- [28] Minlie Huang, Qiao Qian, and Xiaoyan Zhu. 2017. Encoding syntactic knowledge in neural networks for sentiment classification. *ACM Transactions on Information Systems* 35, 3 (2017), 26:1–26:27.
- [29] Xiaolei Huang and Michael J. Paul. 2019. Neural user factor adaptation for text classification: Learning to generalize across author demographics. In *Proceedings of the Joint Conference on Lexical and Computational Semantics*. Association for Computational Linguistics, 136–146.
- [30] Lee Humphreys, Phillipa Gill, and Er Krishnamurthy. 2010. How much is too much? Privacy issues on Twitter. *Conference of International Communication Association* (2010).
- [31] Laurent Jacob, Francis R. Bach, and Jean-Philippe Vert. 2008. Clustered multi-task learning: A convex formulation. In *International Conference on Neural Information Processing Systems*. Curran Associates, Inc., 745–752.
- [32] Rosie Jones, Ravi Kumar, Bo Pang, and Andrew Tomkins. 2007. “I know what you did last summer”: Query logs and user privacy. In *Proceedings of the ACM Conference on Information and Knowledge Management*. ACM, 909–914.
- [33] Rafal Jozefowicz, Wojciech Zaremba, and Ilya Sutskever. 2015. An empirical exploration of recurrent network architectures. In *Proceedings of the International Conference on Machine Learning*. JMLR.org, 2342–2350.
- [34] Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1746–1751.
- [35] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.
- [36] Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*. OpenReview.net.
- [37] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Proceedings of the Annual Conference on Neural Information Processing Systems*. Curran Associates, Inc., 1097–1105.
- [38] Abhishek Kumar and Hal Daumé III. 2012. Learning task grouping and overlap in multi-task learning. In *Proceedings of the International Conference on Machine Learning*. JMLR.org.
- [39] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Proceedings of the Conference on Artificial Intelligence*. AAAI Press, 2267–2273.
- [40] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the International Conference on Machine Learning*. JMLR.org, II–1188–II–1196.
- [41] Ji Young Lee and Franck Dernoncourt. 2016. Sequential short-text classification with recurrent and convolutional neural networks. In *The Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. The Association for Computational Linguistics, 515–520.
- [42] Ao Li, Zhou Qin, Runshi Liu, Yiqun Yang, and Dong Li. 2019. Spam review detection with graph convolutional networks. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. ACM, 2703–2711.
- [43] Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. 2016. Topic modeling for short texts with auxiliary word embeddings. In *Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17–21, 2016*. ACM, 165–174.
- [44] Chun Hung Li and C. K. Lee. 1993. Minimum cross entropy thresholding. *Pattern Recognition* 26, 4 (1993), 617–625.
- [45] Qimai Li, Zhichao Han, and Xiao-Ming Wu. 2018. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, 3538–3545.
- [46] Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 25–30.
- [47] Yifu Li, Ran Jin, and Yuan Luo. 2019. Classifying relations in clinical narratives using segment graph convolutional and recurrent neural networks (Seg-GCRNs). *Journal of the American Medical Informatics Association* 26, 3 (2019), 262–268.
- [48] Bin Liu, Deguang Kong, Lei Cen, Neil Zhenqiang Gong, Hongxia Jin, and Hui Xiong. 2015. Personalized mobile app recommendation: Reconciling app functionality and user privacy preference. In *Proceedings of the ACM International Conference on Web Search and Data Mining*. ACM, 315–324.
- [49] Bang Liu, Mingjun Zhao, Di Niu, Kunfeng Lai, Yancheng He, Haojie Wei, and Yu Xu. 2019. Learning to generate questions by learning what not to generate. In *The World Wide Web Conference*. ACM, 1106–1118.



- [50] Kun Liu and Evimaria Terzi. 2010. A framework for computing the privacy scores of users in online social networks. *ACM Transactions on Knowledge Discovery from Data* 5, 1 (2010), 6:1–6:30.
- [51] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*. AAAI Press, 2873–2879.
- [52] Yabing Liu, Krishna P. Gummadi, Balachander Krishnamurthy, and Alan Mislove. 2011. Analyzing Facebook privacy settings: User expectations vs. reality. In *Proceedings of the ACM SIGCOMM Conference on Internet Measurement Conference*. ACM, 61–70.
- [53] Andrew L. Maas, Awni Y Hannun, and Andrew Y. Ng. 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of the International Conference on Machine Learning*. JMLR.org, 3.
- [54] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. The Association for Computer Linguistics, 55–60.
- [55] Huina Mao, Xin Shuai, and Apu Kapadia. 2011. Loose tweets: An analysis of privacy leaks on Twitter. In *Proceedings of the Annual ACM Workshop on Privacy in the Electronic Society*. ACM, 1–12.
- [56] Ahmadreza Mosallanezhad, Ghazaleh Beigi, and Huan Liu. 2019. Deep reinforcement learning-based text anonymization against private-attribute inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, 2360–2369.
- [57] Cam Tu Nguyen, De Chuan Zhan, and Zhi Hua Zhou. 2013. Multi-modal image annotation with multi-instance multi-label LDA. In *Proceedings of the International Joint Conference on Artificial Intelligence*. AAAI Press, 1558–1564.
- [58] Feiping Nie, Heng Huang, Xiao Cai, and Chris Ding. 2010. Efficient and robust feature selection via joint  $l_{2,1}$ -norms minimization. In *Proceedings of the International Conference on Neural Information Processing Systems*. Curran Associates, Inc., 1813–1821.
- [59] Sinno Jialin Pan, James T Kwok, and Qiang Yang. 2008. Transfer learning via dimensionality reduction. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, 677–682.
- [60] Hao Peng, Jianxin Li, Yu He, Yaopeng Liu, Mengjiao Bao, Lihong Wang, Yangqiu Song, and Qiang Yang. 2018. Large-scale hierarchical text classification with recursively regularized deep graph-CNN. In *Proceedings of the World Wide Web Conference on World Wide Web*. ACM, 1063–1072.
- [61] Yuxin Peng, Jinwei Qi, and Yuxin Yuan. 2018. Modality-specific cross-modal similarity measurement with recurrent attention network. *IEEE Transactions on Image Processing* 27, 11 (2018), 5585–5599.
- [62] Yuxin Peng, Jinwei Qi, and Yunkan Zhuo. 2020. MAVA: Multi-level adaptive visual-textual alignment by cross-media bi-attention mechanism. *IEEE Transactions on Image Processing* 29, 1 (2020), 2728–2741.
- [63] Yuxin Peng, Yunzhen Zhao, and Junchao Zhang. 2019. Two-stream collaborative learning with spatial-temporal attention for video classification. *IEEE Transactions on Circuits and Systems for Video Technology* 29, 3 (2019), 773–786.
- [64] NhatHai Phan, Yue Wang, Xintao Wu, and Dejing Dou. 2016. Differential privacy preservation for deep auto-encoders: An application of human behavior prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, 1309–1316.
- [65] Tiejun Qian, Bei Liu, Quoc Viet Hung Nguyen, and Hongzhi Yin. 2019. Spatiotemporal representation learning for translation-based POI recommendation. *ACM Transactions on Information Systems* 37, 2 (2019), 18:1–18:24.
- [66] Frederic Raber and Antonio Krüger. 2018. Deriving privacy settings for location sharing: Are context factors always the best choice? In *IEEE Symposium on Privacy-Aware Computing*. IEEE, 86–94.
- [67] Virgile Landeiro Dos Reis and Aron Culotta. 2015. Using matched samples to estimate the effects of exercise on mental health via Twitter. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence, January 25–30, 2015, Austin, Texas*. AAAI Press, 182–188.
- [68] Pengjie Ren, Zhumin Chen, Zhaochun Ren, Furu Wei, Liqiang Nie, Jun Ma, and Maarten de Rijke. 2018. Sentence relations for extractive summarization with deep neural networks. *ACM Transactions on Information Systems* 36, 4 (2018), 39:1–39:32.
- [69] Sungyong Seo, Jing Huang, Hao Yang, and Yan Liu. 2017. Interpretable convolutional neural networks with dual local and global attention for review rating prediction. In *Proceedings of the ACM Conference on Recommender Systems*. ACM, 297–305.
- [70] Abhishek Kumar Singh, Manish Gupta, and Vasudeva Varma. 2018. Unity in diversity: Learning distributed heterogeneous sentence representation for extractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press.
- [71] Manya Sleeper, Justin Cranshaw, Patrick Gage Kelley, Blase Ur, Alessandro Acquisti, Lorrie Faith Cranor, and Norman Sadeh. 2013. “I read my Twitter the next morning and was astonished”: A conversational perspective on Twitter regrets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3277–3286.

- [72] Xuemeng Song, Fuli Feng, Jinhuan Liu, Zekun Li, Liqiang Nie, and Jun Ma. 2017. NeuroStylist: Neural compatibility modeling for clothing matching. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 753–761.
- [73] Xuemeng Song, Xiang Wang, Liqiang Nie, Xiangnan He, Zhumin Chen, and Wei Liu. 2018. A personal privacy preserving framework: I let you know who can see what. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 295–304.
- [74] Yi Song, Daniel Dahlmeier, and Stephane Bressan. 2014. Not so unique in the crowd: A simple and effective algorithm for anonymizing location data. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 19–24.
- [75] Robert Tibshirani. 2011. Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society* 73, 3 (2011), 267–288.
- [76] Lam Tran, Deguang Kong, Hongxia Jin, and Ji Liu. 2016. Privacy-CNH: A framework to detect photo privacy with convolutional neural network using hierarchical features. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, 1317–1323.
- [77] Asimina Vasalou, Alastair J. Gill, Fadhila Mazanderani, Chrysanthi Papoutsis, and Adam N. Joinson. 2011. Privacy dictionary: A new resource for the automated content analysis of privacy. *Journal of the Association for Information Science and Technology* 62, 11 (2011), 2095–2105.
- [78] Quan Wang, Jun Xu, Hang Li, and Nick Craswell. 2013. Regularized latent semantic indexing: A new approach to large-scale topic modeling. *ACM Transactions on Information Systems* 31, 1 (2013), 5:1–5:44.
- [79] Xudong Wang, Zhaowei Cai, Dashan Gao, and Nuno Vasconcelos. 2019. Towards universal object detection by domain attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Computer Vision Foundation / IEEE, 7289–7298.
- [80] Zhiguo Wang, Haitao Mi, and Abraham Ittycheriah. 2016. Semi-supervised clustering for short text via deep representation learning. In *Proceedings of the Conference on Computational Natural Language Learning*. ACL, 31–39.
- [81] Jennifer Williams, Carinda Feild, and Kristina James. 2011. The effects of a social media policy on pharmacy students' Facebook security settings. *American Journal of Pharmaceutical Education* 75, 9 (2011), 177.
- [82] Ruobing Xie, Zhiyuan Liu, Jia Jia, Huanbo Luan, and Maosong Sun. 2016. Representation learning of knowledge graphs with entity descriptions. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, 2659–2665.
- [83] Qiongfai Xu, Lizhen Qu, Chenchen Xu, and Ran Cui. 2019. Privacy-aware text rewriting. In *Proceedings of the International Conference on Natural Language Generation*. Association for Computational Linguistics, 247–257.
- [84] Su Yan and Xiaojun Wan. 2015. Deep dependency substructure-based learning for multidocument summarization. *ACM Transactions on Information Systems* 34, 1 (2015), 3:1–3:24.
- [85] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. The Association for Computational Linguistics, 1480–1489.
- [86] Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, 7370–7377.
- [87] Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. 2016. ABCNN: Attention-based convolutional neural network for modeling sentence pairs. *Transactions of the Association for Computational Linguistics* 4 (2016), 259–272.
- [88] Min Ling Zhang and Zhi Hua Zhou. 2007. Multi-label learning by instance differentiation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, 669–674.
- [89] Richong Zhang, Yue Wang, Yongyi Mao, and Jinpeng Huai. 2019. Question answering in knowledge bases: A verification assisted model with iterative training. *ACM Transactions on Information Systems* 37, 4 (2019), 40:1–40:26.
- [90] Ting Zhao and Xiangqian Wu. 2019. Pyramid feature attention network for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Computer Vision Foundation / IEEE, 3085–3094.

Received November 2019; revised May 2020; accepted June 2020