SECL: Separated Embedding and Correlation Learning for Demographic Prediction in Ubiquitous Sensor Scenario

Yiwen Jiang[1,2,3],Wei Tang[1,2,3], Neng Gao[1,3], Chenyang Tu[1,3], Jia Peng[1,3], Min Li[1,3]
[1]*State Key Laboratory of Information Security, Chinese Academy of Sciences, Beijing, China*
[2]*School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China*
[3]*Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China*
{jiangyiwen,tangwei,gaoneng,tuchenyang,pengjia,minli} @iie.ac.cn

***Abstract*—Knowing exact demographic attributes of users is crucial for human-computer interaction, intelligent marketing and automatic advertising. Ubiquitous sensor devices yield massive volumes of temporal data which hide a lot of valuable demographic information. In this paper, we bridge the gap between sensor data and demographic prediction to obtain real attributes of users from popular sensor devices: pedometer, which is widely used in mobile devices. We propose a novel model named Separated Embedding and Correlation Learning (SECL) for demographic prediction. Specifically, SECL first process the input data with a separated embedding layer to disentangle task-specific features for interference eliminating, and then capture the hidden correlations between different tasks via a correlation learning layer, finally the refined task-specific features are fed into a multi-task prediction layer to predict demographic attributes. Experimental results show impressive performance of our model on a real-world pedometer dataset, which is made publicly available on https://github.com/deepdeed/SECL.**

***Index Terms*—demographic prediction, sensor data, sequence learning**

I. Introduction

Recently, sensing devices are ubiquitous in people's daily life. For example, many mobile devices like mobile phone embed pedometer, gyroscope, accelerometer, vibrometer and magnetometer. Some popular wearable devices such as Fitbit, Apple Watch, and Android Wear use pedometer, accelerometer and heart rate monitor [3]. All these sensing devices generate trillions of sensor data points per year, including rich signals such as step count variability, which closely correlate with users' daily activities as diverse as walking, exercise, or trip and indirectly hide the user's demographic attributes characteristics. As a result, extracting knowledge and emerging patterns from sensor data for user attribute prediction is a nontrivial task.

Obtaining individual demographic attributes is crucial for the applications of human-computer interaction, intelligent marketing and automatic advertising. Beyond conventional applications of user attribute inference, knowing demographic attributes of users via sensor data has it's own unique applications in internet of things. For example, in smart home system, explicit attribute could be used to enable human-computer interaction more humanized and friendly. More specifically, when responding to a human with known gender, the computer could select a gender-aware response from many possible candidates to make the user more comfortable, which significantly enhance the competitiveness of the products [19]. However, it is usually not easy for smart device to obtain exact users' attributes.

In this paper, we make effort on the reasonable utilize of pedometer data (step count sequence) for demographic prediction. Most of earlier studies on attribute prediction are primarily involve analysis of the user-generated data derived from social media, including Facebook [25], Twitters [5], [26], microblogs [33], telephone conversations [12], YouTube [11], web search queries [15], social networking chats [24], and forum posts [9]. In this paper, we extend our sight to the ubiquitous mobile and sensing device to bridge the gap between sensor data and users' demographic attributes. We attempt to extract knowledge and emerge users' daily walking patterns from pedometer data, thereby inferring users' demographic. To the best of our knowledge, there is only one existing work that has used sensor data for prediction task in 2018, Ballinger et al. [2] combined step count with heart rate and proposed a semi-supervised learning method to predict cardiovascular risk in medical field. Nevertheless, the heart rate data is hard to obtain and full of privacy sensitivity on personal health. In this case, we use only step count data but make the finer granularity of analysis for a more general problem of demographic prediction.

Previous work on demographic prediction, for example, Structured Neural Embedding (SNE) [31], usually employ shared embedding to capture the shared feature of user attribute. The advantages of this model are relatively simple structure and less parameters, but it ignores the interferences of multiple tasks. Another method of Embedding Transformation Network (ETN) [17] address this problem using separated transform embedding upon the shared embedding to extract task-specific features. But ETN also have a significant limitation: insufficient ability to learn hidden correlation features between multi-tasks. Commonly in multi-task learning, optimal correlation features are helpful for model to achieve better performance, especially in the case of demographic prediction

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

tasks such as gender prediction and work type prediction, which is complementary to each other.

To tackle the above problem, we present a novel model named Separated Embedding and Correlation Learning (SECL) for demographic prediction. In SECL, we first leverage an separated embedding layer and disentangle the task-specific features. Attention mechanism is employed in separated embedding layer to highlight the dominant days in every week for each task, as we find that weekends are more important for gender prediction. Then, we design a correlation learning layer to capture the informative correlations between different tasks for feature enhancement. Another attention model is adopted here to distinguish the degree of relevance between multiple tasks, for example, work is more relevant to gender than age. Finally, the refined task-specific features are fed into a multi-task prediction layer for demographic prediction. In addition, to better learn the patterns and trends of users for demographic prediction from fallible sensor data, we carry out a good deal of effective data pre-processing work for noise tolerance.

In experiments, several state-of-the-art baselines on demographic prediction are taken into comparison. The experimental results prove that our model outperforms all these baselines on the typical tasks such as Partial Label Prediction and New User Prediction with the popular evaluation metrics of F1 score. Furthermore, we release our pedometer dataset for promoting research in related fields. To our best knowledge, this is the first public sensor data with exact demographic annotations.

Overall, our contributions are as follows:

- We first extend the sight to the ubiquitous mobile and sensing device to bridge the gap between sensor data and users' demographic attributes. And new dataset of pedometer record with exact demographic annotation is released.
- We propose a novel model for demographic prediction, named Separated Embedding and Correlation Learning (SECL). It first disentangle task-specific features using separated embedding and then extract optimal correlations between multi-tasks via correlation learning. This model is more reasonable and explainable than previous models.
- Extensive experiments are conducted on a real world pedometer dataset. Results prove the effectiveness of the proposed SECL model. Furthermore, effective data resampling are carried out for noise tolerance on pedometer data.

The rest of this paper is organized as follows. Section II summarizes the related works. Section III gives the problem formalization of multi-task demographic prediction. Section IV introduces some data pre-processing methods used in paper for data enhancement. Section V discusses our approach in detail, and section VI presents the experimental results and analysis. And finally, in section **??**, we conclude our work.

II. Related Work

A. Demographic Prediction

Many studies have been devoted to the problem of demographic prediction using various types of data. Schler et al. [28] learned the differences in writing style and content between male and female bloggers to determine an unknown author's gender on the basis of a blog vocabulary. With the advent of the big data era, social network and search queries data are used to infer demographic attributes [4], [8]. Also, location and mobile application usage data have been used [21], [34]. And some recent works used purchasing history for demographic prediction [17], [27], [31].

Demographic Prediction is commonly considered as a multi-task learning problem. Early work on demographic prediction often infer each attribute independently but there are features helpful for each task learned from other tasks. Dong et al. [10] considered the interrelation between gender and age, and employed a Double Dependent-Variable Factor Graph model to predict them simultaneously based on various human-defined features. Wang et al. [31] concatenated multiple attribute labels and generated a single structured label to leverage the potential correlations for multi-task learning based on a shared embedding. Raehyun et al. [17] proposed an Embedding Transformation Network (ETN) model that leveraged a embedding transformation layer to capture task-specific features but ignore the informative correlations between multi-tasks. Here, we go a step further to adding a correlation learning layer after embedding transformation layer to retain optimal correlation features for multi-tasks.

B. Knowledge Discovery from Sensor Data

Wide-area sensor infrastructures, remote sensors, and wireless sensor networks yield massive volumes of disparate, dynamic, and chronologically distributed data. Given the unique characteristics of sensor data, particularly its spatiotemporal nature and presence of constraints associated with the data collection, there have been many research efforts to analyze the sensor data which build upon the general research in the data mining community [6], [23].

Sashank et al. [22] inferred vehicular users' location and traveled routes using gyroscope, accelerometer, and magnetometer information without any users' knowledge by modeling the problem as a maximum likelihood route identification on a graph generated from the OpenStreetMap publicly available database of roads. Ballinger et al. [2] presented two semi-supervised training methods, semi-supervised sequence learning and heuristic pre-training for the task of cardiovascular risk prediction using users' heart rate and step count data derived from wearable devices. In this paper, we pay attention to a more general problem of demographic prediction using sensor data.

C. Sequence Learning

A large amount of work show the effectiveness of Recurrent Neural Networks (RNNs) to learn hidden properties of sequential data. In 1997, M. Hochreiter [13] introduced

an efficient, gradient based method called long short-term memory (LSTM), which has been widely applied for reducing the vanishing and exploding gradient problems and learning longer term dependencies. In attribute inference, Ballinger et al. [2] presented a semi-supervised sequence learning method using LSTM to predict the risk of cardiovascular.

Attention mechanism has been proven to be significant in many tasks. Attention mechanism can selectively focus on more informative data, and it was first presented to capture the most relevant information at each step in machine translation task [1]. Then, other natural language processing tasks such as text classification [32] also used it. In other domains such as recommendation and computer vision also adopted attention mechanism [7], [20], which can not only improve performance, but also make models more interpretable. Given the representation ability of RNN, we use it to learn separated attribute representations, and adopt attention mechanism to selectively learn correlation features for demographic prediction.

III. Problem Formalization

Following previous work, we formalize demographic prediction as a multi-task prediction problem. Let $\mathbf{D} = [\{\mathbf{x}_1, \mathbf{y}_1\}, ..., \{\mathbf{x}_N, \mathbf{y}_N\}]$ be a list of all data samples, where N is the number of samples. As each sample corresponds to an individual user, $\mathbf{x}_n$ and $\mathbf{y}_n$ are all pedometer record and demographic attributes of the n-th user, respectively. $\mathbf{y}_n$ can be viewed as a list of labels for each task $[y_n^1, ..., y_n^M]$. The number of possible classes for m-th attribute is C^m. The pedometer record $\mathbf{x}_n = [x_n^1, ..., x_n^L)]$ can be an ordered list of daily step counts depending on data sets, where L is the number of days. In a real world scenario, we might have full or partial demographic attributes of users. Our goal is to predict all the missing attributes in the dataset or predict the attributes for new users. We follow two types of problem used in [17], [31]:

- **Partial Label Prediction** is for the situation that users gave some part of their demographic attributes, so that we want to know the remain unknown attributes. Let $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N]$ be a set of users' pedometer records and $\mathbf{Y}^{ob} = [\bar{\mathbf{y}}_1, ..., \bar{\mathbf{y}}_N]$ be the users' demographic attribute that are partially observed. Given $\mathbf{X}$ with $\mathbf{Y}^{ob}$, the objective is to learn a function to predict the unknown attributes $\mathbf{Y}^{un} = [\ddot{\mathbf{y}}_1, ..., \ddot{\mathbf{y}}_N]$. Note that $\bar{\mathbf{y}}_n \cup \ddot{\mathbf{y}}_n = [y_n^1, ..., y_n^M] = \mathbf{y}_n$.
- **New User Prediction** is to predict demographic attributes for new users. Given $\mathbf{X}^{ob}$ with partially/fully observed attributes $\mathbf{Y}^{ob}$ the objective is to learn a function to predict demographic attributes for new users. New users' pedometer record are $\mathbf{X}^{un}$ and corresponding labels are $\mathbf{Y}^{un}$. Note that unlike partial label prediction where $\mathbf{X}$ is used as the input for both training and test sets, $\mathbf{X}$ is split into $\mathbf{X}^{ob}$ for the training set and $\mathbf{X}^{un}$ for the test set, which implies $\mathbf{X}^{ob} \cap \mathbf{X}^{un} = \varnothing$.

IV. Data Pre-processing

In this section, we talk about data pre-processing to enhance the fallible sensor data for noise tolerance. Due to the slightly difference in sensitivity and the sensors themselves may have arbitrary errors, data derived from sensors may have different levels of noise, and low-quality sensors can even produce abnormal data. To address this problem, we design a resampling method to transform the raw data into partition intervals for noise tolerance.

Specifically in our application, the pedometer profile of n-th user is represented as a sequence of daily walking step counts, the raw step counts sequence could be presented as follow:

$$\mathbf{x}_n : [x_n^1, ..., x_n^L]$$

where L is the number of days.

Commonly, the process of the vanilla interval resampling is: first get the global $[min, max]$ of daily step counts by considering all users profiles; next calculate the $(k+1)$-dimensional partition vector based on a naive method such as uniform partition; then use this partition vector for sequence quantization. The $(k+1)$-dimensional partition vector $\mathbf{p}$ could be presented as:

$$\mathbf{p} = [0, \frac{1}{k}, \frac{2}{k}, ..., \frac{(k-1)}{k}, 1] * (max - min)$$

obviously, the proposed resampling can also be easily extended to other sequence-based applications.

Note that directly using the global $[min, max]$ will lead to a serious sparsity problem. For example, an occasional extremely large global max value will cause a large number of data points cluster near the global min value (see Figure 1). To address this issue, we redefine the naive global $[min, max]$ by global mean value (denoted as $mean$) and standard deviation (denoted as std):

$$[min^{re}, max^{re}] = [mean - std, mean + std]$$

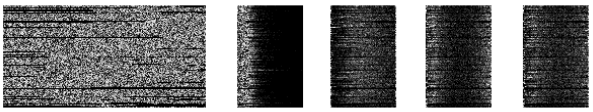

Fig. 1. The sparsity problem is well solved by using the $[min^{re}, max^{re}]$ with proposed non-linear partition method. In this figure, the white dots represent the relatively larger values. The first image is the original data of 100 users where each row represents 200 days step count sequences. Here we set the intervals number $k = 64$. The second image is the data distributions produced by naive $[min, max]$, the third image is produced by $[min^{re}, max^{re}]$ with uniform partition, the fourth image is produced by $[min^{re}, max^{re}]$ with non-linear partition ($\mu = 0$, $\sigma^2 = 0.2$), and the last image is produced by $[min^{re}, max^{re}]$ with non-linear partition ($\mu = 0$, $\sigma^2 = 0.3$).

Although the problem of sparsity has been partially solved, a large number of data points are near the mean value when we use the uniform partition (see Figure 1). In this situation, we design a non-linear function based on Gaussian distribution to

transform the linear partition into non-linear. This non-linear function is presented as:

$$g(z) = (\frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(z-\mu)^2}{2\sigma^2}})^{-1}$$

$$f(z) = \frac{\int_0^z g(z)dz}{\int_0^1 g(z)dz}$$

where x is the integration variable limited in range $[0, 1]$, μ and σ^2 are the mean value and variance of Gaussian distribution. A simple example of the proposed non-liner function is illustrated in Figure 2, and the positive effect of using non-linear partition with is presented in Figure 1.

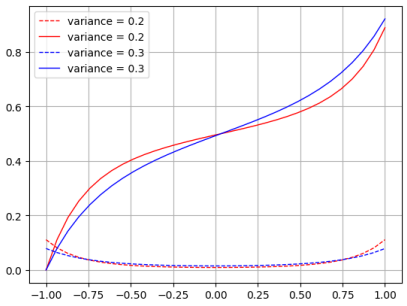

Fig. 2. The visualization of non-linear function. The μ is set to be 0, and σ^2 are set to be 0.2 and 0.3. The dotted lines are the reciprocal of Gaussian distributions $g(z)$, and the solid lines are the proposed non-linear functions $f(z)$ based on the discrete integration of dotted lines. Note that a large σ^2 will make the non-linear partition tend to be linear.

Based on aforementioned resampling method, all pedometer records of users could be quantified as a regular value with limited noise. Specifically, each element of input sequence will be reset to the minimum of its located interval, which could be presented as:

$$\tilde{x}_n^i = p_j \text{ if } x_n^i \in [p_j, p_{j+1}]$$

where p_j denotes the j-th element of aforementioned partition vector, $\tilde{x}_n^i$ is the processed step count value on i-th day. Then the processed step count sequence could be presented as:

$$\tilde{\mathbf{x}}_n : [\tilde{x}_n^1, ..., \tilde{x}_n^L]$$

V. Our Approach

In this section, we present the details of the proposed SECL model. An overview of SECL is illustrated in Figure 3. Different from previous work typically using shared embedding, we build a separated embedding layer to disentangle task-specific features. In addition, a correlation learning layer is used to learn the optimal correlations between different tasks. Finally, the multi-task prediction layer make the predictions of all tasks simultaneously.

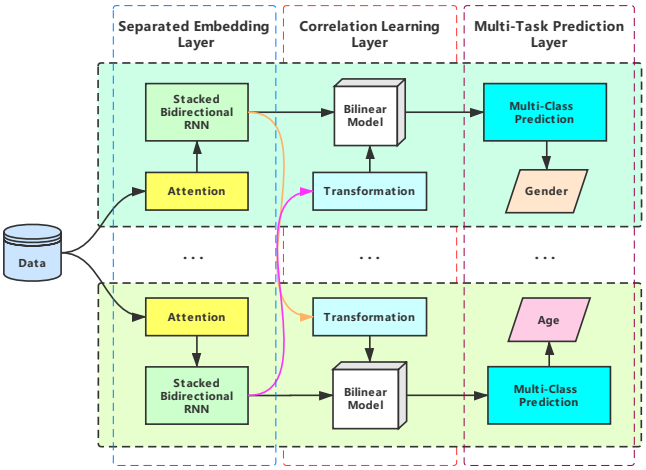

Fig. 3. The architecture of SECL. First, the separated embedding layer outputs the task-specific representations. Then, the correlation learning layer learns the hidden correlations between different tasks. Finally, the multi-task prediction layer simultaneously infers multiple attributes.

A. Separated Embedding Layer

We use separated embedding branches for learning task-specific features. Different from shared embedding mapping user profiles to a shared features that ignores the interferences among multiple tasks, the separated embedding branches eliminate these interferences and produce a relatively pure task-specific representations directly. We have to admit that shared embedding may retain informative correlations among multiple attributes, and separated embedding method seems to ignore these correlations. Nevertheless, this shared embedding lacks explanations for correlation extraction. In the next section, we will introduce correlation learning layer to learn these correlations in an more interpretable way.

Considering that the sensor data has a strong temporality, we adopt the most popular sequence learning model of LSTM [13] as the backbone of each embedding branches. In this paper, we assume that the contexts from both past and future are useful and complementary to each other. Therefore we combine forward (left to right) and backward (right to left) recurrent to build a bidirectional LSTM (Bi-LSTM) [29]. Moreover, the stacked recurrent layers are used to build a deep RNN model to enhance the representation ability of the separated embedding layer. Here, we use same stacked bidirectional LSTM branches to respectively learn task-specific features.

Note that the different days in a week play a diverse role in each tasks. For example, the step count of weekday is more important than the weekend for occupation inference, intuitively. From this point of view, we adopt attention mechanism [1] to give the relative significant time periods higher weight. Finally the separated embedding branches could be presented as:

$$\mathbf{w}_i = softmax(tanh(\mathbf{v}_i\tilde{\mathbf{x}} + \mathbf{b}_i))$$

$$\mathbf{T}_i = \mathcal{B}_{\theta_i}(\mathbf{w}_i\tilde{\mathbf{x}})$$

where $\mathcal{B}$ is the trainable model of Bi-LSTM, θ_i is the trainable parameters of embedding branches for i-th task, $\tilde{\mathbf{x}}$ is the input sequence, and $\mathbf{T}_i$ is the output task-specific feature for i-th task. Moreover, $\mathbf{v}_i$ and $\mathbf{b}_i$ are trainable parameters of the attention model in i-th branch, $\mathbf{w}_i$ is the attention weights describe the importance assigned to each element of the input. When we neglect the difference of importance in every elements and abandon the attention mechanism, all the attention weights would be 1.

B. Correlation Learning Layer

The separated embedding layer eliminate the interferences among multiple tasks, but the informative correlations between different tasks also have been ignored. Thus, we design a correlation learning layer to learn these hidden correlation features. The key components of this correlation learning layer are the transformation network and the bilinear mixer.

The transformation network is designed as a full connection network with attention mechanism. Full connection network is widely used in hidden feature learning on account of their excellent learning ability and desirable scalability. Commonly, in multi-task prediction application, some tasks are more strongly associated with a certain task. For example, gender and work type are more relevant than gender and age in walking scenario. Therefore, we adopt attention mechanism to assist the full connection layer for extracting the optimal correlation features by assigning the important hidden elements the relatively large weights. Then the correlation feature for i-th task could be presented as:

$$\mathbf{w}_i' = softmax(tanh(\mathbf{v}_i'\bar{\mathbf{T}}_i + \mathbf{b}_i'))$$

$$\mathbf{C}_i = \mathcal{F}_{\phi_i}(\mathbf{w}_i'\bar{\mathbf{T}}_i)$$

where $\mathcal{F}$ is the trainable model of full connection network, ϕ_t is the trainable parameters, $\mathbf{C}_i$ represents correlation feature for i-th task, $\bar{\mathbf{T}}_i$ is the concatenation of other task-specific features. Moreover, $\mathbf{v}_i'$ and $\mathbf{b}_i'$ are trainable parameters of the attention model. Attention weights $\mathbf{w}_i'$ describe the importance assigned to each hidden elements of the input. When we neglect the difference of importance in every elements and abandon the attention mechanism, all the attention weights would be 1.

We combine the i-th task-specific feature and the captured correlation feature in a bilinear mixer. Bilinear mixer is a two-factor model with the mathematical property of separability: their outputs are linear in either factor when the others held constant, which has been demonstrated that the influences of two factors can be efficiently separated and combined in a flexible representation [30]. The combination function can be formulated as:

$$\mathbf{M}_i = \mathbf{T}_i\mathbf{W}_i\mathbf{C}_i$$

where $\mathbf{W}_i$ is the trainable parameters of bilinear model, $\mathbf{T}_i$ is task-specific feature of i-th task , and $\mathbf{C}_i$ is the correlation feature extracted for i-th task.

C. Multi-Task Prediction Layer

With the separation representation and correlation representation obtained by the previous two layers, we obtain the prediction probability for the demographic attribute of a given user by:

$$q(\mathbf{y}_i|\tilde{\mathbf{x}}) = softmax(\mathbf{O}_i\mathbf{M}_i)$$

where $\mathbf{M}_i$ is the refined task-specific feature, $\mathbf{O}_i$ is the trainable parameter that is responsible for converting the refined task-specific feature into predictions through linear transformation.

The goal of demographic prediction is to infer all demographic attributes of users from their pedometer profiles. For i-th task, we minimize the sum of the negative log-likelihoods defined as:

$$Loss_i = -\sum_{j=1}^{N} logq(\mathbf{y}_{i,j}|\tilde{\mathbf{x}}_j)$$

where N is the total number of users. $\tilde{\mathbf{x}}_j$ and $\mathbf{y}_{i,j}$ are the input of j-th user's resampled daily walking step count sequence and he/her inferred attribute class of i-th task.

Combined with all these task-specific losses, the full multi-task loss function is:

$$Loss = \sum_{i=1}^{t} \lambda_i Loss_i$$

where the hyper-parameter λ controls the trade-off between all of t task-specific losses. Considering that all tasks are equal important in our experiments, we set all λ to be 1.

VI. Experiments

In this section, we present the details of the experiments and analyze the effectiveness of SECL.

A. Dataset

We build a real world pedometer dataset came from WeChat (https://weixin.qq.com/), a popular mobile application with over one billion active users. Nowadays, most mobile phones have embedded pedometers, so WeChat has the chance to develop a subfunction called WeChat Sport that collects and ranks users' as well as their net friends' daily walking step counts online. Everyone register in WeChat can see their and their friends' daily walking step counts on the ranking list provided by WeChat Sport. It is easy for the WeChat provider to get these data, but it's difficult for researchers to obtain such a high quality pedometer data. To get this data, we launched 168 volunteers and spent nearly one year to record their and their friends' daily step count through WeChat platform based on the pedometer embedded in smartphone. Then we spent a lot of time cleaning the data and asking these volunteers to annotate the data. Our dataset contains 39,246 users' 300-days walking step counts during the period from 2018.6.11 to 2019.4.6. All of the users are annotated with their demographic attributes: gender, age, and work type. To guarantee the reliability of the data, we have already removed those unsuitable users who have more than 150 days of zero

records. We make the dataset publicly available [1], and all the users in the dataset has been anonymized for the privacy issue. The detailed distribution of users' attributes are listed in Table I.

As we described in Section *Problem Formalization*, we conduct experiments in two different problem settings. For the partial label prediction problem setting, we randomly set certain attributes as observed. Each attribute of a user has a 50% chance to be observed and used in training. The remaining unknown attributes are used in the evaluation. For experiments on new user prediction, we split our dataset into non overlapping sets. We choose 8:1:1 as training, validation and testing split ratio. Following previous work [17], to minimize noise due to randomness, we create 10 different splits, and then average the results of 10 datasets and report them in this paper.

TABLE I
DISTRIBUTION OF USERS' ATTRIBUTES.

Attributes	Value	Users	Distribution
Gender	male	22134	56%
	female	17112	44%
Age	young	8635	22%
	adult	19230	49%
	middle age	7064	18%
	old	4317	11%
Work Type	physical	14128	36%
	mental	17660	45%
	other	7458	19%

B. Evaluation Metrics

Following previous work [17], [31], we employ F-measure such as macro F1 score (denoted mF1) and weighted F1 score (denoted wF1) to evaluate our model. F-measure is a widely used measure method in multi-task learning, and it is also the most popular evaluation metrics for demographic prediction. F1 score is calculated as the harmonic mean of precision (denoted mP or wP) and recall (denoted mR or wR). Macro F1 calculate metrics for each label, and find their unweighted mean. This does not take label imbalance into account. Weighted F1 calculates metrics for each label, and find their average weighted by support (the number of true instances for each label). This alters macro F1 to account for label imbalance.

C. Baseline Models

We compare our models with several state-of-the-art baseline models on demographic prediction. The description of these baselines are listed below:

- **POP** is a naive method that always predicts the given sample as the majority classes in training set. It is a popular baseline that ignores characteristics of users to verify the prediction performance of proposed dataset without any machine learning in previous work [17], [31].
- **SVD**, Singular Value Decomposition, is widely used in demographic prediction [14], [34]. It employs an effective matrix decomposition method to obtain low dimensional representations of users. Logistic models are trained for each demographic attribute separately.
- **JNE**, Joint Neural Embedding [31], maps users' all day walking histories into latent vectors. These vectors are processed by average pooling and then fed into a linear prediction layer for each task.
- **SNE**, Structured Neural Embedding [31], has similar structure with JNE. The only difference between SNE and JNE is that the loss of SNE is computed via a log-bilinear model with structured prediction.
- **ETN**, Embedding Transformation Network [17], adopt a shared embedding just as SNE. The shared embedding is fed into an embedding transformation layer to obtain the transformed representation. Then the transformed representation is directly fed into the prediction layer.
- **ETNA**, Embedding Transformation Network with Attention [17], is an improved version of ETN. The transformed representation produced by embedding transformed layer is fed to a task-specific attention layer to take into account the importance of each element in user profile.

D. Experimental Settings

When compared with the above baseline models, we adopt the most simple architectures for our proposed model. Specifically, we use only one layer of bidirectional RNN with 128 LSTM units in each separated embedding branch. And the full connection networks in correlation learning layer is set to be the shallow architectures using only one hidden layer with 128 sigmoid units. We use random values drawn from the Gaussian distribution with 0 mean and 0.01 standard deviation to initialize the weight matrices in LSTM, full connected layer, and prediction layer. Learning from [16], the forget gate bias are initialized to be 5 to let the forget gate close to 1, namely no forgetting. Thus, long-range dependencies can be better learned at the beginning of training. All other bias, the cell as well as hidden states of LSTMs in our work are initialized at 0. Adam [18] is used as the optimization algorithm and the mini-batch size is 128. The learning rate is set to be $1e^{-5}$. After each epoch, we shuffle the training data to make different mini-batches. In our experiments, all the input data is pre-processed by the non-linear resampling ($\mu = 0$, $\sigma^2 = 0.3$) with $k = 64$, except for specifically explained.

E. Performance Comparison

Table II shows the experimental results on new user prediction task and partial label prediction task. Based on these results, we have the following findings:

(1) As POP using the most simple strategy of always predicting the given sample as the majority classes in training set, all other learning models outperform this naive method. It means that machine learning can extract valuable information for demographic attributes from pedometer record, which

[1] https://github.com/deepdeed/SECL

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT MODELS.

Model Name	Partial Label						New User					
	mP	mR	mF1	wP	wR	wF1	mP	mR	mF1	wP	wR	wF1
POP [17]	0.079	0.152	0.104	0.276	0.505	0.357	0.015	0.059	0.024	0.081	0.282	0.126
SVD [14]	0.262	0.215	0.236	0.481	0.547	0.512	0.121	0.106	0.113	0.268	0.337	0.299
JNE [31]	0.316	0.221	0.260	0.509	0.551	0.529	0.175	0.109	0.134	0.313	0.335	0.324
SNE [31]	0.319	0.227	0.265	0.512	0.554	0.532	0.179	0.117	0.142	0.315	0.343	0.328
ETN [17]	0.341	0.255	0.292	0.532	0.560	0.546	0.188	0.139	0.160	0.318	0.361	0.338
ETNA [17]	0.355	0.274	0.309	0.548	0.573	0.560	0.211	0.147	0.173	0.327	0.373	0.348
SECL [a]	0.367	0.285	0.321	0.561	0.582	0.571	0.226	0.158	0.186	0.339	0.387	0.361
SECL [b]	0.393	0.307	0.345	0.585	0.607	0.596	0.238	0.171	0.199	0.353	0.402	0.376
SECL [c]	0.387	0.301	0.339	0.573	0.597	0.585	0.236	0.169	0.197	0.348	0.399	0.372
SECL	**0.410**	**0.329**	**0.365**	**0.597**	**0.612**	**0.604**	**0.253**	**0.188**	**0.216**	**0.363**	**0.417**	**0.388**

[a]abandon all attention mechanisms in both separated embedding layer and correlation learning layer.
[b]abandon attention mechanism in separated embedding layer.
[c]abandon attention mechanism in correlation learning layer.

further proves the significant prospect for the research of data mining on widely existed sensor data.

(2) As we emphasized in this paper, the ability to disentangle task-specific features and learn optimal correlations between different tasks are significant for demographic prediction with multi-task learning. The baseline models such as JNE, SNE, ETN and ETAN use the shared embedding that implicitly leverage these correlations, but they ignore the interferences among multiple tasks. We first adopt the separated embedding to avoid such interference, and then employ the correlation learning to obtain the correlations. Although we use the most simple architectures and abandon all attention mechanism of our model (SECL[a]) for comparison with baselines, it still outperform all the state-of-the-art baseline models, whether the general baseline model (SVD) or the special models for demographic prediction (ETN et al.). In following Section *Effectiveness Verification*, we verify the effectiveness of separated embedding and correlation learning more carefully.

(3) The attention mechanism is helpful for demographic prediction on pedometer data. As shown in table II, our model employing attention in correlation learning layer (SECL[b]) or in separated embedding layer (SECL[c]) is better than which abandon all attention mechanism (SECL[a]). This founding is similar to the previous work [17], where ETNA is better than ETN on transaction history, as well as here on pedometer records. In following Section *Visualization of Attention*, we will analyze this phenomenon in more detail.

F. Visualization of Attention

To further analyze the impact of attention mechanism in our model, we provide visualization of weights calculated by attention mechanism. We picked example that provide insights for users' pedometer records from Sunday to Saturday during 20 weeks with the average attention scores in each task. Based on the attention weights from our model, we draw heatmap in Figure 4.

First thing to notice in this example is that Saturday and Sunday obtain highest attention in gender prediction task. Usually, women prefer to go shopping on the weekends, and most women work like men on weekdays. The fact that weekend get relatively higher score and weekday get relatively lower score in gender prediction task, which fit the gap between male and female in Figure 5.1, and it also accords with our intuition.

In age prediction, our model give relatively higher attention scores to weekdays but lower scores to weekends. Intuitively, young people are more energetic and active, and adults tend to stay at their desks. But on the weekends, parents and children may go out to play together. Figure 5.2 empirically demonstrates this from a statistical point of view, as the gap between adult and young is relatively large on weekdays but small on the weekends.

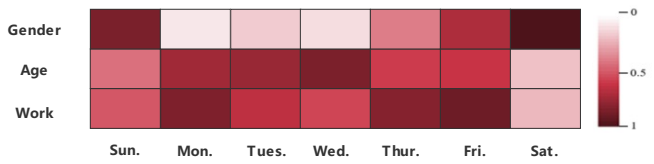

Fig. 4. Comparison of attention weights calculated by separated embedding layer.

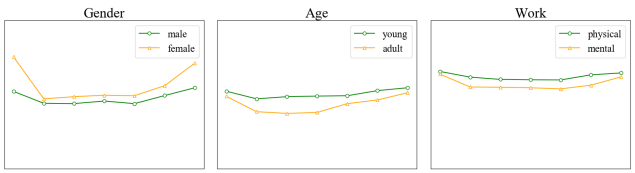

Fig. 5. Mean value (y-axis) of pedometer records from Sunday to Saturday (x-axis).

Similar to age prediction, we find that the attention scores given to weekdays are also larger then weekends in work type prediction. It is easy to understand that during the weekdays,

mental workers use their brains more than manual workers do with their hands and feet. The statistics of pedometer records in Figure 5.3 support this view.

In addition to observing attention weight in separated embedding layer, we also made the visualization analysis of attention weight in correlation learning layer. Figure 6 illustrate the heatmap of average attention scores between different tasks.

In gender prediction, we find that the attention scores of gender to work is larger than gender to age. This is in line with our view that the degrees of relevance between multiple tasks are different. Intuitively speaking, men are more likely to engage in physical work than women, and gender should be relatively more balanced in all age groups. This view is accords to the statistic results in Figure 7.1 and 7.2.

In age prediction, the attention scores of age to work is larger than age to gender. It is not hard to understand that middle-aged people tend to give up physical work due to physical capability decline, and energetic adults are more likely to be competent to do physical work. This view is supported by Figure 7.2 and 7.3.

In work type prediction, the attention scores of work to gender is larger than work to age. One of the reason is that the complementarity between gender to work is more significant than age to work. Figure 7.1 and 7.3 prove this view.

Note that the attention scores of gender to work is similar with work to gender. And this symmetry also exists between gender and age, as well as between age and work. This symmetry indicates that the complementarity between a pair of tasks is not directional but symmetrical. It indicates that we can go a further step to simplify our model by using shared attention in correlation learning layer. We will try this in future work.

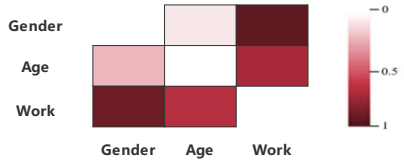

Fig. 6. Comparison of attention weights calculated by correlation learning layer.

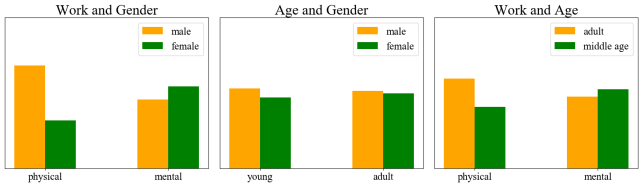

Fig. 7. Ratio (y-axis) of different attributes distributed to other attributes (x-axis).

G. Effectiveness of Correlation Learning

To verify the effectiveness of our model more carefully, we include the experiments of ETNA and SECL with variant structures in Table III. ETNACL is a model of ETNA integrating correlation learning layer. SE is a model of SECL abandoning correlation learning layer.

TABLE III
PERFORMANCE TEST FOR CORRELATION LEARNING.

Model Name	Partial Label		New User	
	mF1	wF1	mF1	wF1
ETNA	0.309	0.560	0.173	0.348
ETNACL	0.337	0.582	0.192	0.367
SE	0.315	0.566	0.181	0.353
SECL	0.365	0.604	0.216	0.388

ETNA first capture the correlation information with shared embedding in a implicate way, and then it get the task-specific feature via attention-based feature transformation. Contrary to ETNA, our SECL first capture the task-specific feature directly using attention-based separated embedding, and then obtain the optimal correlation feature in attention-based correlation learning layer. Obviously, our method is more reasonable and explainable.

SECL outperform SE in both tasks of partial label prediction and new user prediction. This directly prove the superior of correlation learning. Furthermore, ETNACL also outperform ETNA,which demonstrate that the correlation learning have strong generalization ability.

H. Impact of Network Depth

Another advantage of SECL is it can be more easily extended into a deep structure. This is because that SECL employ recurrent neural network (Bi-LSTM) and full connection (FC) as the backbone in separated embedding layer and correlation learning layer. Commonly, a deep structure could obtain better representation ability than a shallow one. Table IV presents the experimental results of SECL with different network depth.

TABLE IV
RESULTS OF SECL WITH DIFFERENT NETWORK DEPTH.

Network Architecture	Partial Label		New User	
	mF1	wF1	mF1	wF1
$[128]-[128]$	0.365	0.604	0.216	0.388
$[128,256]-[128]$	0.368	0.605	0.220	0.391
$[128,256,512]-[128]$	0.369	0.605	0.223	0.392
$[128,256,512]-[128,256]$	0.375	0.610	0.232	0.397
$[128,256,512]-[128,256,512]$	0.381	0.615	0.240	0.403

The model architectures could be represented by a general form as:

$$[L_1,...,L_A]-[F_1,...,F_B]$$

where L_i is the dimension for the hidden states of i-th Bi-LSTM in separated embedding layer, and F_j is the dimension for the hidden states of j-th FC in correlation learning layer.

According to the experimental results, we find that stacking more layers of Bi-LSTM can indeed improve all evaluated metrics. However, the improvements are not significant and the benefits will also vanish when more layers been stacked,

in that the recurrent units maintain activation for each time-step which have already make the network to be extremely deep. Therefore, stacking more recurrent layers in separated embedding layer will not bring too much additional discrimination ability to the model.

Nevertheless, stacking more FC in correlation learning layer could bring more improvement, for the deeper FC network has stronger representation ability to learn the hidden correlations between multi-tasks. But it should be noted that adding the depth of FC network will lead to a sharp increase in the number of trainable parameters. In future work, we will try more network structures to avoid parameter inflation.

I. Hyper-parameters Tuning

The hyper-parameters of resampling in data pre-processing (see Section *Data Pre-Processing*) are significant for noise tolerance on fallible sensor data.

We test SECL with different k from 8 to 1024 with step 8. We find SECL perform stably with little fluctuation when k is limited in 32 to 128, and the best k is 64. we also find that SECL suffers from performance degradation in some extremely cases such as $k = 1024$ and $k = 8$. It is because an extremely large k weakens the noise tolerance, while an extremely small k treats most of the information as noise.

We also test our model with different σ^2 from 0.01 to 0.99 with step 0.01. We find SECL perform well when σ^2 is limited in 0.26 to 0.37, and the best σ^2 is 0.30. When σ^2 get much bigger, the non-linear function tend to be linear, which cause many data points aggregating toward the mean value. Contrarily, an extremely small σ^2 makes many data points far away from mean value. Both of these situations lead to a problem of data sparsity.

Based on the above experiments, we suggest that the appropriate interval should be set for the hyper-parameters to avoid too large or too small values. Visualization analysis similar to Figure 1 can be performed for verifying the effectiveness of the hyper-parameter setting.

J. Experiments on Transaction Data

We also conduct experiments on transaction dataset used in [17]. This dataset is the first public dataset containing both transaction records and demographic information. It consists of purchasing histories of 56,028 users and contains the gender, age, and marital status of all the users. Table V reports the experimental results.

Results show that SECL also outperforms all baseline models with impressive improvement on transaction data, which prove the strong generalization ability of our model. According to the observation of separated embedding layer, we find that cosmetics and perfumes obtain higher attention score in gender prediction, it is because females purchase are actively in duty free stores as they are generally more interested in those items, which is already proved in previous work [17]. Beside, in correlation learning layer, we also discover the attention weights given to correlation of age and marital status are relatively higher. It accords with our intuition of gender should be more balanced in all age groups whether married or not. The statistics results on transaction data also prove this view.

TABLE V
RESULTS ON TRANSACTION DATA.

Model Name	Partial Label		New User	
	mF1	wF1	mF1	wF1
POP	0.108	0.370	0.028	0.134
SVD	0.247	0.524	0.118	0.306
JNE	0.269	0.539	0.139	0.334
SNE	0.271	0.542	0.137	0.321
ETN	0.300	0.557	0.165	0.336
ETNA	0.317	0.569	0.182	0.360
SECL	**0.379**	**0.617**	**0.246**	**0.419**

VII. CONCLUSION

In this paper, we extend the sight to the ubiquitous mobile and sensor device to bridge the gap between sensor data and users' demographic attributes. We release a new dataset with pedomoter records and demographic annotations. Furthermore, we proposed a Separated Embedding and Correlation Learning (SECL) model, which first disentangle task-specific features and then learn the correlation features between multiple tasks. Compared with previous models, our model is more reasonable and explainable.

REFERENCES

[1] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. [Online]. Available: http://arxiv.org/abs/1409.0473

[2] B. Ballinger, J. Hsieh, A. Singh, N. Sohoni, J. Wang, G. H. Tison, G. M. Marcus, J. M. Sanchez, C. Maguire, J. E. Olgin, and M. J. Pletcher, "Deepheart: Semi-supervised sequence learning for cardiovascular risk prediction," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, 2018, pp. 2079–2086. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16967

[3] D. R. Bassett, L. P. Toth, S. R. LaMunion, and S. E. Crouter, "Step counting: A review of measurement considerations and health-related applications," *Sports Medicine*, vol. 47, no. 7, pp. 1303–1315, Jul 2017. [Online]. Available: https://doi.org/10.1007/s40279-016-0663-1

[4] B. Bi, M. Shokouhi, M. Kosinski, and T. Graepel, "Inferring the demographics of search users: social data meets search queries," in *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013*, 2013, pp. 131–140. [Online]. Available: https://doi.org/10.1145/2488388.2488401

[5] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella, "Discriminating gender on twitter," in *Conference on Empirical Methods in Natural Language Processing*, 2011, pp. 1301–1309.

[6] V. Chandola, O. A. Omitaomu, A. R. Ganguly, R. R. Vatsavai, N. V. Chawla, J. Gama, and M. M. Gaber, "Knowledge discovery from sensor data (sensorkdd)," *SIGKDD Explor. Newsl.*, vol. 12, no. 2, pp. 50–53, Mar. 2011. [Online]. Available: http://doi.acm.org/10.1145/1964897.1964911

[7] J. Chen, H. Zhang, X. He, L. Nie, W. Liu, and T. Chua, "Attentive collaborative filtering: Multimedia recommendation with item- and component-level attention," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, 2017, pp. 335–344. [Online]. Available: https://doi.org/10.1145/3077136.3080797

[8] A. Culotta, N. R. Kumar, and J. Cutler, "Predicting the demographics of twitter users from website traffic data," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, 2015, pp. 72–78. [Online]. Available: http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9945

[9] N. Dong and N. A. Smith, "Author age prediction from text using linear regression," in *Acl-Hlt Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 2011, pp. 115–123.

[10] Y. Dong, Y. Yang, J. Tang, Y. Yang, and N. V. Chawla, "Inferring user demographics and social strategies in mobile social networks," in *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, 2014, pp. 15–24. [Online]. Available: https://doi.org/10.1145/2623330.2623703

[11] K. Filippova, "User demographics and language in an implicit social network," in *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012, pp. 1478–1488.

[12] N. Garera and D. Yarowsky, "Modeling latent biographic attributes in conversational genres," in *Joint Conference of the Meeting of the ACL and the International Joint Conference on Natural Language Processing of the Afnlp: Volume*, 2009, pp. 710–718.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[14] J. Hu, H. Zeng, H. Li, C. Niu, and Z. Chen, "Demographic prediction based on user's browsing behavior," in *Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007*, 2007, pp. 151–160. [Online]. Available: https://doi.org/10.1145/1242572.1242594

[15] R. Jones, R. Kumar, B. Pang, and A. Tomkins, ""i know what you did last summer":query logs and user privacy," in *Sixteenth ACM Conference on Conference on Information and Knowledge Management*, 2007, pp. 909–914.

[16] R. Józefowicz, W. Zaremba, and I. Sutskever, "An empirical exploration of recurrent network architectures," in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, 2015, pp. 2342–2350. [Online]. Available: http://jmlr.org/proceedings/papers/v37/jozefowicz15.html

[17] R. Kim, H. Kim, J. Lee, and J. Kang, "Predicting multiple demographic attributes with task specific embedding transformation and attention network," in *Proceedings of the 2019 SIAM International Conference on Data Mining, SDM 2019, Calgary, Alberta, Canada, May 2-4, 2019*, 2019, pp. 765–773. [Online]. Available: https://doi.org/10.1137/1.9781611975673.86

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[19] S. Li, J. Wang, G. Zhou, and H. Shi, "Interactive gender inference with integer linear programming," in *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, 2015, pp. 2341–2347. [Online]. Available: http://ijcai.org/Abstract/15/331

[20] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 2016, pp. 289–297.

[21] E. Malmi and I. Weber, "You are what apps you use: Demographic prediction based on user's apps," in *Proceedings of the Tenth International Conference on Web and Social Media, Cologne, Germany, May 17-20, 2016.*, 2016, pp. 635–638. [Online]. Available: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13047

[22] S. Narain, T. D. Vo-Huu, K. Block, and G. Noubir, "Inferring user routes and locations using zero-permission mobile sensors," in *2016 IEEE Symposium on Security and Privacy (SP)*, May 2016, pp. 397–413.

[23] O. A. Omitaomu, R. R. Vatsavai, A. R. Ganguly, N. V. Chawla, J. Gama, and M. M. Gaber, "Knowledge discovery from sensor data (sensorkdd)," *SIGKDD Explor. Newsl.*, vol. 11, no. 2, pp. 84–87, May 2010. [Online]. Available: http://doi.acm.org/10.1145/1809400.1809417

[24] C. Peersman, W. Daelemans, and L. V. Vaerenbergh, "Predicting age and gender in online social networks," in *International Workshop on Search and Mining User-Generated Contents*, 2011, pp. 37–44.

[25] D. Quercia, R. Lambiotte, D. Stillwell, M. Kosinski, and J. Crowcroft, "The personality of popular facebook users," in *Acm Conference on Computer Supported Cooperative Work*, 2012, pp. 955–964.

[26] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, "Classifying latent user attributes in twitter," in *International Workshop on Search and Mining User-Generated Contents*, 2010, pp. 37–44.

[27] Y. S. Resheff and M. Shahar, "Fusing multifaceted transaction data for user modeling and demographic prediction," *CoRR*, vol. abs/1712.07230, 2017. [Online]. Available: http://arxiv.org/abs/1712.07230

[28] J. Schler, "Effects of age and gender on blogging," in *Proc. of AAAI Symposium on Computational Approaches for Analyzing Weblogs*, 2006, pp. 199–205.

[29] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997. [Online]. Available: https://doi.org/10.1109/78.650093

[30] J. B. Tenenbaum and W. T. Freeman, "Separating style and content," in *Advances in Neural Information Processing Systems 9, NIPS, Denver, CO, USA, December 2-5, 1996*, 1996, pp. 662–668. [Online]. Available: http://papers.nips.cc/paper/1290-separating-style-and-content

[31] P. Wang, J. Guo, Y. Lan, J. Xu, and X. Cheng, "Your cart tells you: Inferring demographic attributes from purchase data," in *Wsdm*, 2016, pp. 173–182.

[32] Z. Yang, D. Yang, C. Dyer, X. He, A. J. Smola, and E. H. Hovy, "Hierarchical attention networks for document classification," in *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, 2016, pp. 1480–1489. [Online]. Available: http://aclweb.org/anthology/N/N16/N16-1174.pdf

[33] L. Zhang, X. Huang, T. Liu, A. Li, Z. Chen, and T. Zhu, "Using linguistic features to estimate suicide probability of chinese microblog users," *Lecture Notes in Computer Science*, vol. 8944, pp. 549–559, 2014.

[34] Y. Zhong, N. J. Yuan, W. Zhong, F. Zhang, and X. Xie, "You are where you go: Inferring demographic attributes from location check-ins," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM 2015, Shanghai, China, February 2-6, 2015*, 2015, pp. 295–304. [Online]. Available: https://doi.org/10.1145/2684822.2685287