via semantic analysis [5]. Some works are proposed to detect social impact of specified persons by using their personal information like employee information [6]. Social media data are also widely used for user portrait. In [7], micro-blog users are profiled by the classification results on the crawled set of messages, pictures and videos posted by social users. Apparently, structural information is a natural choice to represent user features for classification task. However, semi-structured or unstructured data contains more rich information to fully understand users' preferences or characteristics.

Conventional techniques, adopted to learn user portrait from text data, are generally based on natural language process techniques or machine learning methods. To acquire higher accuracy, researchers tried to propose various approaches for different scenarios. For instance, [8] proposed to profile location-sensitive users from labels aggregated via a crowdsourcing model. Similar approaches could be found in [9]. Text Attribution Tool (TAT) is proposed to extract numerical attributes of authors such as Arabic, emails [10]. Some approaches can also analyze social media data by proposing effective classifier built on top of entropy-based similarity measurement designed for short text classification [11]. Alternatively, some works focusing on feature engineering. In these works, features are categorized into three groups, i.e., content-based features, style- based features and topic-based features [3]. Then, regression models could be applied on these kinds of features and n-gram representation form could be employed to represent textual data [12]. Definitely, syntactic features could also be adopted for this task [13].

Nowadays, various deep learning-based approaches [14] have been proposed for natural language processing related tasks. In [15], word character representation based neural network with attention integrated is proposed which treated sequences of words rather than sequences of characters as the model input of the CNN [16]. Users' social posts like textual data can also be embedded by sub-words of the extracted n-grams [17]. A multi-task learning based framework is proposed in [4] for author profiling (MTAP). Some extended versions are also proposed in the literature like User-aware Sentiment Topic Model (USTM) [18]. However, none of existing works could be directly applied to our problem settings, i.e., learning an accurate user portrait from the collected social dating data. This task requires to estimate user portrait like age, education, characteristics from only one piece of user post. Apparently, these characteristics are correlated with each other to some extent.

3. THE PROPOSED PORTRAITAI

In this section, we will describe how the proposed approach could discover users' profile correctly. In the proposed multi-task neural network, we define five different objective functions with each loss function representing one kind of attributes, and different output layers correspond to each task. To save computational cost and improve performance, some textual features are shared among all five tasks.

3.1. Problem Formulation

The self-introduction written by users is essentially a sequence of words. Let X denote a self-introduction, and we have $x = \{w_1, w_2, w_3, \cdots, w_N\}$ where N denote the length of X with w_i as its i-th word. Each word wi is then represented using a word vector v which is a pre-trained word embedding matrix [19]. This word embedding matrix employs Skip Gram model [20] which can capture the relations between current word and its neighboring words. The objective is to maximize the following loss function, given as:

$$\mathcal{L}_{SG} = \frac{1}{|V|}\sum_{t=1}^{|V|}\sum_{0<|i|\leq c}\log\left(f(w_{t+i}, w_t)\right)$$

where V is the corpus set and $|V|$ is the size of V, w_{t+i} denotes the context words within a sliding window w_{t-c}^{t+c} and c is the window size. Moreover, $f(w_{t+i}, w_t) = p(w_{t+i|w_{t+i}})$, and the probability of the context words to be predicted can be calculated as:

$$p(w_{t+i|w_t}) = \frac{exp({v'}_{w_{t+1}}{}^{\top} v_{w_t})}{\sum_{w_{t+1}\in V} exp({v'}_{w_{t+1}}{}^{\top} v_{w_t})}$$

where V_{w_t} denotes the embedding vector of the t-th word w_t, and v and v' refer to the input and output embedding vector, respectively. This embedding matrix is trained on large-scale high-quality data set consisting of 8 million Chinese words and phrases. Herein, after the transformation, the new introduction is denoted as $S \in R^{e\times|V|}$, where e is the dimension of the word vectors of embedded words.

3.2. The Proposed Approach

Essentially, the estimation of user's attributes or characteristics could be considered as a classification task. Then, the estimation results on various user's attributes could be combined together to acquire user portrait. The mixture of single class classification task and several multiclass classification tasks could be learned under a multi-task framework as these tasks are correlated with each other.

The proposed multi-task model as well as the newly designed loss functions will be illustrated in the following paragraph. The framework of the proposed approach is plotted in Figure 1. Obviously, all subtasks have the same set of input data. Globally, different types of features are generated and global features containing high level information will be shared across subtasks circled on the right-hand side. Each subtask has its own feature extracted from textual data and is circled on the left-hand side. There are five subtasks in total. We employ five different FC layers,

i.e., sigmoid layer, softmax layer, to predict its class in each subtask. For simple task like gender prediction, we output either 'Male' or 'Female' for a user. For complicate classification task like characteristic profiling, we have four groups of classes and only one class from each group is to be output. Thus, the class in each group with a higher probability will be assigned as the class label of user's characteristic. At last, each user has four class labels as his/her characteristic labels.

For the loss function design, we have considered different cases. For each subtask, the output layer should be carefully designed. For gender and characteristics prediction subtask, the sigmoid function is chosen as the active function for the output layer. Its loss function is designed based on binary cross entropy, given as

$$L = \mathbb{E}\sum_{i=1}^{n} -w_n \left[y_n \cdot \log x_n + (1 - y_n) \cdot \log(1 - x_n) \right]$$

For the rest subtasks, we adopt softmax function as the active function and the corresponding loss function is based on category cross entropy, calculated as

$$\text{loss}(x, class) = -x[class] + \log\left(\sum_i \exp(x[i])\right)$$

As a result, we have five loss functions and each loss function corresponds to one subtask. Let L^{gender}, $L^{characteristic}$, $L^{education}$, L^{salary}, L^{height} and L^{age} respectively denote the loss function of gender, characteristic, education, salary and age prediction subtasks. These subtasks are trained simultaneously and the overall loss function for the proposed approach is given as:

$$L = L^{gender} + L^{characteristic} + L^{education} + L^{salary} + L^{height} + L^{age}$$

We adopt 'Adam' [21] to optimize the model and the bath size is set to 128.

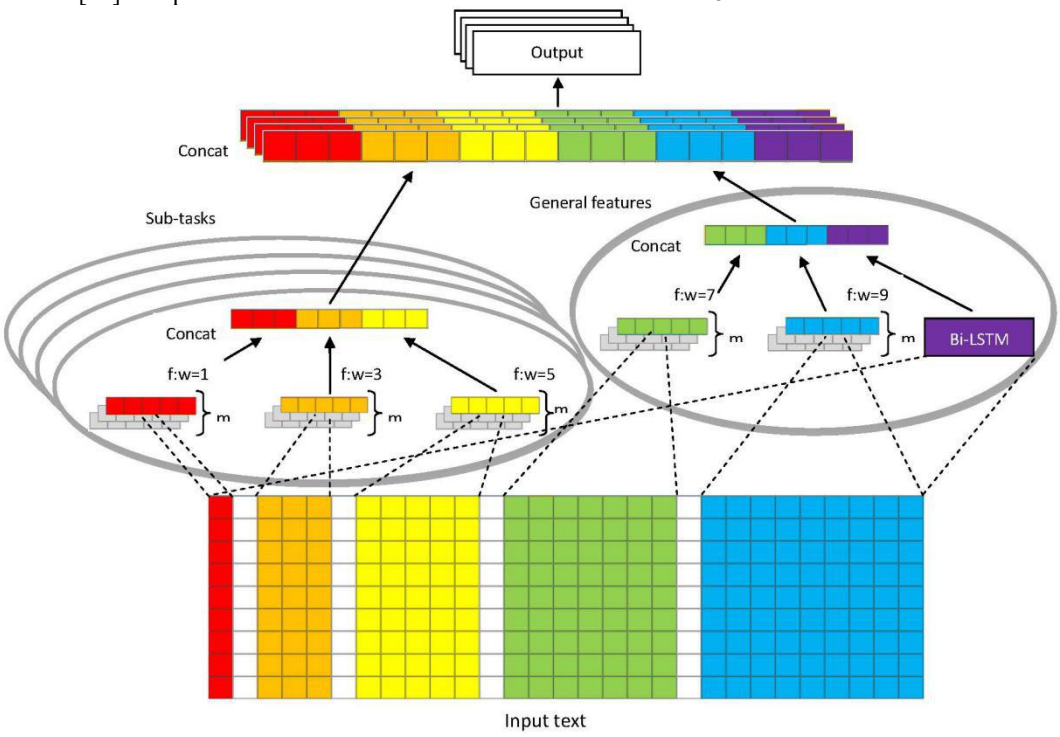

Figure 1 The proposed multi-task model for generating user portrait

4. EXPERIMENTS

In this section, we first briefly introduce the evaluation dataset used in the experiments. Then, we evaluate the proposed approach as well as the baseline methods. The experimental results on predicting user's attributes like age, gender, education level, salary and characteristics are reported for performance comparison.

4.1. Experimental Dataset and Preprocessing

Experiments were evaluated on a real-world dataset. It contains both user's self-introduction data as well as aforementioned attributes which is collected from Zhenai.com(https://www.zhenai.com/). This Website is one of the most popular online dating social networks in China. The collected data set contains 8,671 self-introductions by different users. The collected dataset is randomly partitioned into training set and testing set. The training and testing sets

contain 6,908 and 1,727 self-introductions, respectively.

4.2. Experimental Results

The evaluation results of the proposed approach as well as the compared three benchmarks models are reported in Table I. From this table, it is obvious that the proposed approach can achieve the best performance on gender, education, salary and age prediction results. The corresponding model performance is improved by 12.8%, 18.4%, 13.8% and 21.8%, respectively, when compared with the second-best approach. Note that the performance on predicting characteristics is slightly lower than the rest approaches by only 1%. The possible reason might be that the characteristics generally are very sparse and many missing values are replaced by 0 which might result in a consistent classification result. We will try to collect a larger dataset and figure out method to cope with the sparse characteristic classes. From these results, we can conclude that if we want to estimate the missing value or correct the fake value, the proposed approach is much better than the rest one. If we focus on predicting user's characteristics, the performance of the proposed approach is comparable to the rest approaches.

Table 1 Model performance comparison results with respect to accuracy criterion

Class	Gender	Characteristic	Education	Salary	Age
WCM	53.68	79.57	48.82	29.39	18.16
Multi-CNN	59.90	79.20	50.22	33.74	24.13
Multi-task	60.34	79.85	48.15	35.41	29.21
Portrait AI	68.10	78.83	57.00	40.31	35.57

5. CONCLUSIONS

With the flourish of various Web applications, user portrait already plays a very important role in understanding user's preferences and thus make the corresponding recommendation. Among these applications, online dating Website urgently needs a more accurate user portrait. Such Website usually has more than 100 million users and these users are reluctant to provide accurate information for some important but sensitive attributes. In the literature, this issue is seldom studied and thus we propose this multi-task deep neural network-based approach. The proposed approach first extracts three kinds of features from the self-introduction. Two of them are treated as global features and one is subtask feature extracted to discriminate itself from other subtasks. We have designed five sub loss functions and sum them up as the global loss function to simultaneously train the proposed model for the estimation of five attributes. Evaluations are performed on the collected dataset and some data preprocessing steps are also involved. The promising results have demonstrated that the proposed approach is superior to the benchmark methods on predicting age, gender, education and salary, but is slightly lower than the rest approaches when predicts characteristics. In the future, more data will be collected and released as public dataset for similar tasks.

ACKNOWLEDGMENTS

This paper is partially supported by Shenzhen Science and Technology Program under Grant No. JCYJ20170811153507788, and the Guangdong Province Science and Technology Department Project under Grant NO.20178090901022. This work is also partially supported by the National Science Foundation of China under grant No.61872108.

REFERENCES

[1] C. Basu, H. Hirsh, W. Cohen et al., "Recommendation as classification: Using social and content-based information in recommendation," in Aaai/iaai, 1998, pp. 714–720.
[2] Z. Wang, J. Liao, Q. Cao, H. Qi, and Z. Wang, "Friend book: a semantic-based friend recommendation system for social networks," IEEE transactions on mobile computing, vol. 14, no. 3, pp. 538–551, 2015.
[3] K. Santosh, R. Bansal, M. Shekhar, and V. Varma, "Author profiling: Predicting age and gender from blogs," Notebook for PAN at CLEF, vol. 2013, 2013.
[4] Z. Jiang, S. Yu, Q. Qu, M. Yang, J. Luo, and J. Liu, "Multi-task learning for author profiling with hierarchical features," in Proceedings of the International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2018, pp. 55–56.
[5] B. George, M. Setayesh, T. P. McCandless, P. Pichardo, K. A. Wolfe, R. Parang, M. F. Diaz-Arscott, J. Condit, B. Culler, N. Horton et al., "Method and system for implementing a cloud-based social media marketing method and system," Apr. 25 2017, uS Patent 9,633,399.
[6] D. Chandrasekaran, D. Costello, and P. Stubbs, "Social media profiling," Nov. 7 2013, uS Patent App. 13/465,335.
[7] M. Pinocchio and A.-M. Popescu, "Automatic profiling of social media users," Jan. 17 2013, uS Patent App. 13/183,260.
[8] W. Niu, J. Caverlee, and H. Lu, "Location-sensitive user profiling using crowdsourced labels," in Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[9] R. Li, S. Wang, H. Deng, R. Wang, and K. C.-C. Chang, "Towards social user profiling: unified and discriminative influence model for inferring home locations," in Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012, pp. 1023–1031.
[10] D. Estival, T. Gaustad, S. B. Pham, W. Radford, and B. Hutchinson, "Tat: an author profiling tool with application to Arabic emails," in Proceedings of the Australasian Language Technology Workshop 2007, 2007, pp. 21–30.
[11] F. Claude, R. Konow, and S. Ladra, "Fast compressed-based strategies for author profiling of social media texts," in Proceedings of the 4th Spanish Conference on Information Retrieval. ACM, 2016, p. 14.
[12] M. Martinc, I. Skrjanec, K. Zupan, and S. Pollak, "Pan 2017: Author profiling-gender and language variety prediction." in CLEF (Working Notes), 2017.
[13] J.-P. Posadas-Dura´n, I. Markov, H. Go´mez-Adorno, G.

Sidorov, I. Batyrshin, A. Gelbukh, and O. Pichardo-Lagunas, "Syntactic n-grams as features for the author profiling task," Proceeding of the CLEF, 2015.
[14] D. Kodiyan, F. Hardegger, S. Neuhaus, and M. Cieliebak, "Author profiling with bidirectional rnns using attention with grus: Notebook for pan at clef 2017," in CLEF 2017 Evaluation Labs and Workshop–Working Notes Papers, Dublin, Ireland, 11-14 September 2017, vol. 1866. RWTH Aachen, 2017.
[15] Y. Miura, T. Taniguchi, M. Taniguchi, and T. Ohkuma, "Author profiling with word+ character neural attention network." in CLEF, 2017.
[16] S. Sierra, M. Montes-y Go´mez, T. Solorio, and F. A. Gonza´lez, "Convolutional neural networks for author profiling," CLEF, 2017.
[17] M. Franco-Salvador, N. Plotnikova, N. Pawar, and Y. Benajiba, "Subword-based deep averaging networks for author profiling in social media." in CLEF (Working Notes), 2017.
[18] Z. Yang, A. Kotov, A. Mohan, and S. Lu, "Parametric and non-parametric user-aware sentiment topic models," in Proceedings of the 38th International
[19] ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2015, pp. 413–422.
[20] Y. Song, S. Shi, J. Li, and H. Zhang, "Directional skip-gram: Explicitly distinguishing left and right context for word embeddings," in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). Association for Computational Linguistics, 2018, pp. 175–180. [Online]. Available: http://aclweb.org/anthology/N18-2028
[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013.
[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.