# Multi-task Learning for Author Profiling with Hierarchical Features

Zhile Jiang[†], Shuai Yu[‡], Qiang Qu[‡], Min Yang[‡1], Junyu Luo[†], Juncheng Liu[†]

[†] College of Computer Science, Sichuan University
[‡] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

## ABSTRACT

Author profiling is an important but challenging task. In this paper, we propose a novel Multi-Task learning framework for Author Profiling (*MTAP*), in which a document modeling module is shared across three different author profiling tasks (i.e., age, gender and job classification tasks). To further boost author profiling, we integrate hierarchical features learned by different models. Concretely, we employ CNN, LSTM and topic model to learn the character-level, word-level and topic-level features, respectively. *MTAP* thus leverages the benefits of supervised deep neural neural networks as well as an unsupervised probabilistic generative model to enhance the document representation learning. Experimental results on a real-life blog dataset show that *MTAP* has robust superiority over competitors and sets state-of-the-art for all the three author profiling tasks[2].

## KEYWORDS

Author profiling, Multi-task learning, Hierarchical features

## 1 INTRODUCTION

Author Profiling is to ascertain various author characteristics like age, gender, native country, job by analyzing their written documents. It is of growing importance in the fields of forensics, security, and marketing. For example, from a forensic linguistics perspective one would like to know the linguistic profile of the author of a harassing text message on the web to narrow down the investigation. Recently, numerous author profiling methods have been developed [1, 2, 7].

Despite the remarkable progress of previous methods, we argue that the author profiling of informal texts still remains challenging in real-world. (i) Most previous approaches consider the profiling tasks independent from each other and train the tasks separately. However, the profiling tasks are often mutually dependent; (ii) The principle behind author profiling is that persons with different characteristics write in different ways, thus everyone's writing styles are different. The information about word morphology and shape is crucial for capturing authors' characteristics. (iii) The prior studies are often based on the bag-of-word assumption. They neglect the semantics of words though these factors are important in identifying author's characteristics, thus inevitably losing the distinguishable information; (iv) The authors with different characteristics have their own special interests, as in [8], the interest topics of authors should be taken into account to further improve the performance of author profiling.

To alleviate the aforementioned limitations, *MTAP* simultaneously optimizes three coupled profiling tasks: age classification, gender classification and job classification. In addition, *MTAP* integrates hierarchical features learned by different models to further boost the performance of author profiling. First, long short-term memory (LSTM) is used to encode the character-level features. Second, a convolutional neural network (CNN) model with 1-D convolutional filters is employed to focus on learning useful word-level n-gram features. Meanwhile, we employ Latent Dirichlet Allocation (LDA) to obtain topic-level features. Some word distributions learned by LDA correspond to authorship style as reflected by the vocabulary and interests of authors.

The main contribution of our approach is threefold: (1) To the best of our knowledge, we are the first to use the multi-task learning to jointly train three typical author profiling tasks. (2) *MTAP* integrates hierarchical features learned by different models, it can be regarded as an ensemble classifier that improves the robustness and generalization performance of a set of classifiers; (3) The experimental results show that *MTAP* consistently outperforms competitive methods on different profiling tasks.

## 2 OUR METHOD

We use $x = \{w_1, w_2, \ldots, w_N\}$ to denote the sequence of input document $x$ , where $N$ is the length of the sequence. To prevent conceptual confusion, we use superscripts "$chr$", "$word$", "$topic$" to indicate the variables that are related to character-level, word-level and topic-level features, respectively.

### 2.1 Character-level feature representation

We compute the character-level feature representation of each word with LSTM model. Given a word $w$ composed of $M$ characters $c_1, \ldots, c_M$, we first transform each characters $c_m$ into a character embedding $r^{chr}_{c_m}$ that is encoded by column vectors in the character embedding matrix $W^{chr} \in \mathbf{R}^{d^{chr} \times |V^{chr}|}$: $r^{chr}_{c_m} = W^{chr} v^{chr}_{c_m}$ ($v^{chr}_{c_m}$: the one-hot representation of character $c_m$; $d^{chr}$: the size of the character embedding; $|V^{chr}|$: the size of character vocabulary). Then, we use a LSTM layer to convert the sequence of character

[1] Min Yang is the corresponding author (min.yang@siat.ac.cn). The work was partially supported by the CAS Pioneer Hundred Talents Program.
[2] Data and codes are available at:https://goo.gl/vsSxGt

*WWW '18 Companion, April 23–27, 2018, Lyon, France*

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3186926

embeddings $\{r_{c_1}^{chr}, \ldots, r_{c_M}^{chr}\}$ into to a sequence of hidden states $\{h_{c_1}^{chr}, \ldots, h_{c_M}^{chr}\}$. In general, we represent each word $w$ with the last hidden state, that is $h_w^{chr} = h_{c_M}^{chr}$. The sequence of word representations are fed into another LSTM and the last hidden state of the LSTM is the input of a non-linear layer, resulting a $T$-dimensional document representation $emb_x^{chr}$ that models the character-level features of document $x$.

## 2.2 Word-level feature representation

For each document $x$, we encode the the sequence of word representations with a CNN layer that can learn the abstract representations of n-grams effectively and tackle the sentences with variable lengths naturally. We use multiple filters with varying window sizes to obtain multiple features and then apply a max-over-time pooling operation over the features. These pooled features are passed to a fully connected non-linear layer whose output is $T$-dimensional word-level document representation $emb_x^{word}$.

## 2.3 Topic-level feature representation

We apply LDA to model the corpus. Under the assumptions of LDA, each document $x$ with $N$ tokens is generated by firstly choosing a document topic distribution $\theta^{topic}$. Then, each token $w$ in the document is generated by choosing a topic $z$ from the document topic distribution $\theta^{topic}$ and choosing a token from the topic word distribution $\phi_z^{topic}$. The model can be inferred from the corpus using Gibbs sampling, as in [6]. In this paper, we aim to obtain the document topic distributions $\theta^{topic}$, which is fed into a non-linear layer to produce a $T$-dimensional representation $emb_x^{topic}$, representing topic-level features.

Finally, for document $x$ we combine the character-level, word-level and topic-level features with simple Hadamard product, obtaining the final document representation $emb_x = emb_x^{chr} \circ emb_x^{word} \circ emb_x^{topic}$. The final document embedding $emb_x$ is referred as the ensemble embedding of all hierarchical features we extracted, and is then fed to a $softmax$ classifier to predict the author profile of the given document.

| Method | Gender | Age | Job |
|---|---|---|---|
| SVM | 66.7 | 62.3 | 28.5 |
| LDAH | 63.2 | 61.7 | 20.3 |
| BMR | 76.1 | 77.7 | - |
| LSTM | 70.3 | 65.9 | 29.7 |
| CNN | 71.1 | 66.0 | 29.2 |
| MTAP | **79.2** | **79.6** | **37.6** |
| w/o multitask | 77.4 | 77.2 | 36.4 |
| w/o chr-level features | 79.1 | 79.4 | 37.0 |
| w/o word-level features | 75.3 | 74.8 | 35.3 |
| w/o topic-level features | 74.2 | 72.1 | 34.9 |

**Table 1: The accuracies of author profiling tasks on Blog dataset**

## 2.4 Multi-task learning

*MTAP* consists of three subtasks, each has its own training objective. We use two task-specific fully connect layers (i.e., *softmax* layer) to predict the gender, age, or job of the author of a set of documents. Each classifier is trained by minimizing the cross-entropy between the predicted distribution and the ground truth distribution. We denote the objective functions of gender, age, and job as $L^{gender}$, $L^{age}$ and $L^{job}$, respectively. For the purpose of improving the document learning, we train these three related tasks simultaneously. The joint multi-task objective function is minimized by: $L = L^{gender} + L^{age} + L^{job}$. We use a minibatch stochastic gradient descent (SGD) algorithm to update the parameters of the model.

# 3 EXPERIMENTS

## 3.1 Datasets

We conduct experiments on *Blog* dataset, which consists of 678,161 blog posts by 19,320 bloggers. Each blog contains the self-reported gender, age, industry and astrological sign. The statistics of the dataset can be found in [5]. For each task, we use 80% documents of each author as the training data, 10% documents of each author as the validation data, and the remaining are used for testing.

## 3.2 Experimental results

We compare our approach with several strong baseline methods including SVM [4], LDAH [6], BRM [3], LSTM, and CNN.

In the experiments, the results are evaluated using classification accuracy. The experimental results are summarized in Table 1. We observe that our method substantially outperforms the competitive methods and gets state-of-the-art on all the three tasks. This verifies the effectiveness of the proposed model.

We also report the ablation test of our model in terms of discarding multi-task learning, character-level features, word-level features and topic-level features, respectively (bottom four rows of Table 1). Generally, all four kinds of factors contribute, and topic-level features contribute most. This is within our expectation since the word distributions correspond to authorship style reflect authors' vocabulary and interests.

# REFERENCES

[1] Kholoud Alsmearat, Mahmoud Al-Ayyoub, and Riyad Al-Shalabi. 2014. An extensive study of the bag-of-words approach for gender identification of arabic articles. In *AICCSA*. IEEE, 601–608.

[2] Emad AlSukhni and Qasem Alequr. 2016. Investigating the Use of Machine Learning Algorithms in Detecting Gender of the Arabic Tweet Author. *IJACSA* 7, 7 (2016), 319–328.

[3] Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. 2009. Automatically profiling the author of an anonymous text. *Commun. ACM* 52, 2 (2009), 119–123.

[4] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *ACL*. 79–86.

[5] J Schler, M Koppel, S Argamon, and J Pennebaker. 2006. Effects of Age and Gender on Blogging. In *AAAI*.

[6] Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. 2011. Authorship Attribution with Latent Dirichlet Allocation. In *CoNLL*.

[7] Min Yang, Xiaojun Chen, Wenting Tu, Ziyu Lu, Jia Zhu, and Qiang Qu. 2018. A Topic Drift Model for authorship attribution. *Neurocomputing* 273 (2018), 133–140.

[8] Min Yang, Jincheng Mei, Fei Xu, Wenting Tu, and Ziyu Lu. 2016. Discovering author interest evolution in topic modeling. In *SIGIR*. ACM, 801–804.

[3] http://u.cs.biu.ac.il/ koppel/BlogCorpus.htm