Detecting User Occupations on Microblogging Platforms: An Experimental Study

Xia Lv [1], Peiquan Jin [1,2], Lin Mu [1], Shouhong Wan [1,2], Lihua Yue [1,2]

[1] School of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, China

[2] Key Laboratory of Electromagnetic Space Information, Chinese Academy of Sciences, Hefei 230027, China

jpq@ustc.edu.cn

Abstract. User occupation refers to the professional position of a user in real world. It is very helpful for a number of applications, e.g., personalized recommendation and targeted advertising. However, because of the risk of privacy leaks, many users do not provide their occupation information on microblogging platforms. This makes it hard to detect user occupations on microblogging platforms. In this paper, we conduct an experimental study on this issue. Particularly, we propose an experimental framework of detecting user occupations on microblogging platforms. We first implement a number of classification models and devise various sets of features for user occupation detection. Then, we propose to construct an occupation-oriented lexicon, which is collected by an iterative extension algorithm considering semantic similarity and importance between words. We combine the lexicon with the word embedding approach to detect user occupations. We conduct comprehensive experiments and present a set of experimental results. The results show that the lexicon-based word embedding method achieves higher accuracy compared with traditional feature-base classification models.

Keywords: Occupation detection; Feature extraction; Word embedding.

1 Introduction

Microblog platforms have been an importance source for information extraction [1]. Users' information on microblog platforms is very helpful for many applications such as personalized recommendation and targeted advertising, due to the great number of microblog users. There are many aspects regarding user information, among which user occupation information is of particular business values for commercial applications [2]. However, because of privacy considerations, users usually do not provide their occupations on microblogging platforms. Thus, it is a challenging issue to automatically detect user occupation on microblogging platforms.

There are already some efforts concentrating on mining users' profile information [3-6, 12, 14-15, 18, 22], such as detection of gender, age, and political orientation. A

multi-source integration framework concentrates more on building feature set of content model and network information, which conducts extensive empirical studies for user occupation industry inference [24], and latent feature representation such as word clusters and embedding is used to classify occupations, but they only generate simple semantic features without comparison to classical models [25]. However, predicting user occupations based on microblog platforms is still a new problem. So far, only a few research works are focused on it, and most of them are related to "*occupation field*" [10, 23-25]. An occupation field indicates an area, which is less specific than an occupation. Thus, it is not sufficient for real applications. For example, "*entertainment*" is an "*occupation field*", but we may wonder whether a user is an actor/actress or a singer.

In this paper, we focus on detecting user occupation on microblogging platforms like Sina Weibo. We address it by leveraging available information such as observable digit contents, user behavior, custom tags, and linguistic messages of users. We experimentally compare several classical models over different feature sets to measure the performance on user occupation detection. Further, we propose a lexicon-based feature selection method and combine it with the word embedding method for user occupation detection. Briefly, we make the following contributions in this paper:

(1) We build a framework for detecting microblog user occupations. It takes advantages of two kinds of information resources, namely the observable digit information and the linguistic messages of users.

(2) We extract features from message contents according to three linguistic models, i.e., *BOW*(*bag-of-words)*, *n-gram*, and *topic model*. Then, we apply these feature sets into a number of classification models including *the logistic regression model* (LR), SVM, and *the random forest model*.

(3) We propose to construct an occupation-oriented lexicon that adapts semantic similarity and word importance to refine the feature set. The occupation-oriented lexicon is used to simplify the feature selection work and reduce the dimension of features. We integrate the lexicon with the word embedding method to detect user occupations and the experimental results suggest the effectiveness of this design.

(4) We conduct experiments on a real data set from Sina Weibo. The performance of different classification models over different feature sets is compared, and the lexicon-based method is also evaluated in terms of different settings.

2 The Proposed Framework for User Occupation Detection

Figure 1 shows the framework of detecting user occupations on microblogging platforms. It consists of occupation identification, feature extraction, occupation-oriented lexicon integration and classification. In the first stage of occupation identification, we screen the signed "*V*" users from huge and mixed unlabeled datasets, and then choose 8 occupations to label the users via manual and rule-based methods. We collect microblogs and custom tags by a web crawler according to selected users IDs, finally forming the whole occupation labeled dataset. In the

feature-extraction stage, we divide the dataset into digit contents and message contents. The digit contents consist of basic user profiles, social influence, and user behavior features, while the message contents contain features described by three linguistic models including *BOW*(*bag-of-words)*, *n-gram*, and *topic model*. Next, we utilize message contents to generate an occupation-oriented lexicon, and remove irrelevant words. The remaining words are represented by the word embedding method. Finally, we perform feature-based extraction for user occupations through a number of classification models.

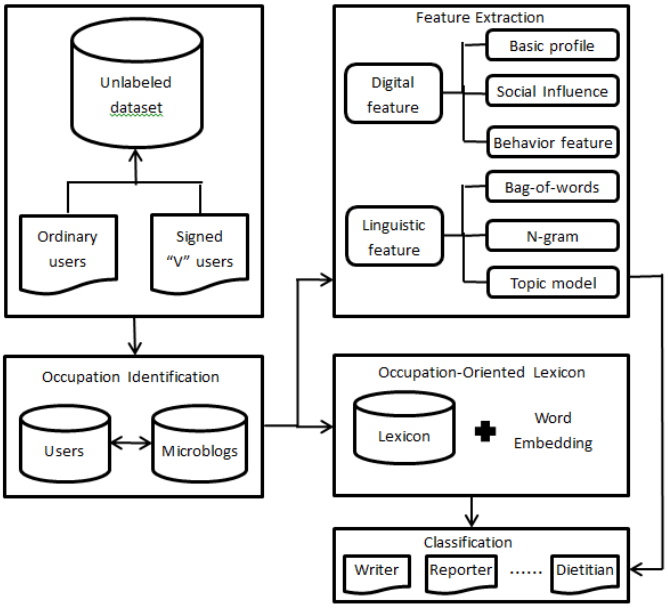

Fig.1. Framework of Microblog User Occupation Detection

2.1 Occupation Identification

Sina Weibo offers APIs for developers to collect data. However, due to the limitation of the APIs, we can only obtain a set of "*verified users*", whose occupations are mandatorily labeled. Then, we extract the most popular occupations from the set of verified users to form the occupation candidates for the experiments. In this paper, we finally prepare eight occupations which are *writer, reporter, lawyer, photographer, actor, singer, doctor, and dietitian*. These occupations are used as the targets to be detected.

We randomly choose 8000 user IDs in the verified user set and use these user IDs to fetch their microblogs from Sina Weibo. For each user, we finally collect a set of microblogs, which contains 100 to 500 text messages. A user containing less than 100 microblogs are not considered.

2.2 Feature Extraction

In this part we describe in detail two types of information, namely digital information and linguistic messages, which can help characterize a user.

2.2.1 Digital Features

Most microblogging platforms provide basic user profile information such as user nickname, location, and a brief introduction. Sina Weibo also allows developer to get basic user information such like the count of a user's microblogs, friends and followers. In summary, we design 38 digital features in this paper, which are denoted by the symbol "*DIGITAL*" in the experiment. All these digital features can be classified as three groups.

Basic Profile Feature. Gender, province, messages count and favorites count are the only four basic profile features provided by Sina Weibo. Basic profile features are widely studied in previous works on mining user attributions [3-7, 12, 14-15, 18, 22].

Social Influence. Social influence of a user is evaluated by two kinds of features. The first kind of features includes the number of followers, the number of friends, and the number of mutual fans. The second one includes the average number of comments, the average number of retweets, and the average number of likes. These are regarded as quantitative indicators to determine the amount of information [21].

Behavior Feature. Generally, users use hashtag "#" to denote the topic of messages. The hashtags "【" and "】" are also used to surround the news title. To this end, we can find the behavior habit of a user. Posting behavior is described by a set of statistics capturing the usage habits of social media such as the average number of messages per day [6]. Such information is useful for constructing a model of a user intuitively [4]. In this paper, we consider the following behavior features: the hashtag count per message, the average number of topics and news, the average number of messages per day, and the average number of messages per hour within one day.

2.2.2 Linguistic Features

Linguistic content information contains user name, description, custom tag and messages posted by users. We concatenate the name, description and custom tag together as a short introduction for each user. In addition, we explore various linguistic content features based on three linguistic models, as detailed below.

Bag-of-Words (BOW). The bag-of-words model is a simplifying representation of a text. We always label one person as "*writer*" according to some keywords captured like "*new book*", "*publishing house*" and many book title marks when scanning one's home page, as in life, we also classify different types by typical keywords in classification task, so a bag of words could represents a text in general. Rao and colleagues manually built a list of words to characterize sociolinguistic behaviors, but it's difficult to translate into strong class-indicative because of much manual effort. Instead, we use *chi-square (CHI)* to select feature words, which measures the degree of the independence between the feature and categories. In addition, we use *term frequency–inverse document frequency (tfidf)* which reflects how important a word is

to a document in a collection or corpus to represent words [9]. Three bags of words are extracted, we name them "*EMOTION*", "*ENGLISH*" and "*WORDS*", while emotion and English characters are all replaced by specific characters in "*WORDS*".

N-gram. N-gram is a contiguous sequence of n items from a given sequence of text or speech. An n-gram of size 1 is referred to as a "*unigram*"; size 2 is a "*bigram*". Rao and colleagues uses a mixture of sociolinguistic features and n-gram models to represent twitters [4]. In this paper we will also utilized this as an approach for our task by deriving the unigram and bigrams of the text. We use "*UNIGRAM*" and "*BIGRAM*" to express this feature set.

Topic Feature. Because of the short-text property of microblogs, the method of representing a microblog by single words cannot reflect the topics of the microblog. For this reason, we use topic models to extract the topic features of microblogs. As the *Latent Dirichlet Allocation (LDA)* model [6, 13] and the *Biterm Topic Model (BTM)* model [17] are commonly used as topic models, we consider these two models in our work. Specially, for the LDS model, we concatenate all users' messages as the input and each user is represented as a multinomial distribution over different number of topics. For the BTM model, we concatenate user name, description and custom tag into a short text introduction, and use the BTM model to enhance the topic learning on the new combined short introduction. The combined feature set of LDA and BTM is named "*T-DISTRIBUTION*" in Table 1. In addition to topic distribution, topic related words are also indispensable when judging user occupations. Thus, we use *word2vec* to obtain topic related words and to produce word embedding. Through word2vec, a relationship lexicon responding to preset occupations can be obtained, which is named "*T-WORDS*" in Table 1.

Table 1. Feature Sets

Feature		Dimension	Description
DIGITAL		38	Basic profile, social influence, behavior feature
BOW	*EMOTION*	1501	Emotion feature. E.g. [cool],[orz]
	ENGLISH	1746	Capital English character. E.g. TFBOYS
	WORDS	9030	Ordinary words
N-gram	*UNIGRAM*	2276	Unigram model. E.g. a, happy, ending
	BIGRAM	7132	Bigram model. E.g. a happy, happy ending
Topic	*T-WORDS*	2333	Topic related words
	T-DISTRIBTION	2**k*	Topic distribution(BTM+LDA), *k* is topic number

2.3 Classification Models

There are a number of classification models proposed by the machine learning area. In this paper, we compare three classification models: a linear SVC classifier with L2 regularized logistic regression [23], a Support Vector Machine (SVM) classifier with linear kernel [4], and an ensemble classification of Random Forest.

3 Occupation-Oriented Lexicon Integration

Traditional classification focuses much on feature extraction. Most of the existing classification works on social network platforms have limitation because of data sparseness and noise. Classical models may produce many features, but semantic features are likely to be ignored. To solve this problem, we propose to build an occupation-oriented lexicon to improve the performance of user occupation detection. This is motivated by the fact that people usually conjecture their occupations by some typically words; thus we can combine occupation-oriented lexicon with word embedding to represent user features.

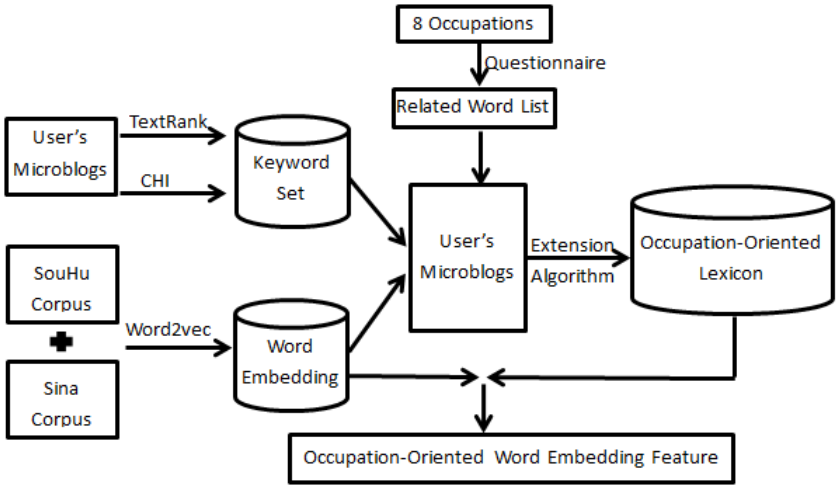

Fig. 2. Flow of Occupation-Oriented Lexicon Integration

As shown in Fig. 2, we first use Word2vec to train the corpus. Then, we get a model of word embedding, which expresses similarity between words. We use TextRank and CHI to generate the keyword set, which contains important words in users' microblogs. Moreover, we invite several persons to conduct questionnaires, during which each involved person is asked to give the words related with different occupations. These related words are used as an input of our lexicon extension algorithm. Finally, we combine lexicon with word embedding to get occupation-oriented word embedding features.

There are three tools used in Fig. 2, namely word2vec, CHI, and TextRank. As described in the above section, word2vec can learn the vector representations of words in the high dimensional vector space; it can find the semantic relationships by computing the cosine distances between words. Thus, we apply its word embedding to compute similarity and express words. CHI aims to select feature words, which measures the degree of the independence between the feature and categories. In addition, TextRank was proposed to solve keyword extraction and it is tasked with the automatic identification of terms that best describe the subject of a document. Therefore, we use these two methods to generate keyword set.

The lexicon extension algorithm in Fig. 2 is an iterative algorithm that integrates important and occupation-oriented lexicon. Algorithm 1 shows its detailed process.

Algorithm 1 Occupation-Oriented Lexicon Extension Algorithm

Input: OList: occupation list; W_{w2v}: the similar word dictionary of word2vec training model;
K_{dic}: the keyword dictionary of TextRank and CHI generating; R: the related words list of different occupations from questionnaire survey; m: number of words extract in

Output: OOL: occupation-oriented lexicon

```
t = 0;
While(1):
    t = t + 1;
    For each occupation occ in OL do
        If OOL_occ = ∅ then
            For each word w in R_occ do
                If w in K_dic and w in W_w2c then
                    OOL_occ.add(w);
        Candidate Words List : CWList = {};
        For each word w in OOL_occ do
            If w in W_w2c then
                Similarity Map: SMap = {};
                For W_sim in SimilarityList(w) do
                    If W_sim in K_dic and W_sim in W_w2c then
                        V_sim = Sum(Distance(W_sim, each word of R_occ));
                SMap.put(W_sim, V_sim);
                Sort(SMap);
                If SMap.size > m then
                    CWList.add(SMap.get(top m));
                else CWList.add(SMap);
        OOL_occ.add(CWList);
        Distinct(OOL_occ);
    Lexicon_t = Sum(OOL_occ);
    If Acc(Lexicon_t) < Acc(Lexicon_{t-1}) then
        break;
Return Lexicon_{t-1}
```

As shown in Algorithm 1, we utilize keywords and related words to filter out noise to identify occupation. The functions *SimilarityList* and *Distance* are computed by word2vec. We combine word embedding and lexicon to predict, and the accuracy whether improved or not guides to stop iterating. Stated another way, we execute the

cyclical function continue until the experiment accuracy of lexicon integrated this round is lower compared to last round.

4 Experimental Results

4.1 Experimental Settings

We use the eight occupations discussed in Section 2.1 to label the 8000 verified users. Each 1000 users are randomly divided into a training set (70% users) and a testing set (30% users). Before implementing linguistic models, we use ICTCLAS to segment the text words. Classification is performed using *Scikit-learn* [8, 26], which is a Python module integrating a wide range of state-of-the-art machine learning algorithms. We compare three machine learning methods: a linear SVC classifier with L2 regularized logistic regression, an SVM classifier with linear kernel and an ensemble classification of Random Forest on 3-fold cross validation of the training set. The SVM model has 4 parameters of "*C*" and the Random Forest has 23 candidate parameters of tree number (*n_estimators*) which are between 40 and 500.

The corpus of word2vec is the combination of SouHu news and microblogs, which contains 69.3 billion effective words. We adapt the *Skip-gram* model and set the window size to 8 to train the corpus with different word representation sizes. The TextRank and CHI are used on the microblogs of the labeled 8000 users, in which the window size in extracting keywords is set to 5. Finally, we get 22.75 thousand keywords and 9030 words selected by CHI.

We use *accuracy*, *macro-averaging precision*, *recall*, and *F-measure* as the metrics [6]. Let U be the user set, N be the number of occupations, if we detect $U_{correct}$ users correctly from U_{test} users, the accuracy is computed as Formula 1. The macro-averaging precision and recall for each occupation are expressed by Formula 2 and 3, respectively, where U_k is the user number of occupation k, and $U_{k,predict}$ is the number of users that are detected to have occupation k.

$$\text{accuracy}=\frac{|U_{correct}|}{|U_{test}|} \tag{1}$$

$$\text{precision} = \text{avg}\left(\frac{|U_{k,correct}|}{|U_{k,predict}|}\right)\ k = 1,2,\dots,N \tag{2}$$

$$\text{recall} = \text{avg}\left(\frac{|U_{k,correct}|}{|U_k|}\right)\ k = 1,2,\dots,N \tag{3}$$

4.2 Performance of Topic Model

First of all, the prediction performances for various numbers of topics are compared based on the topic model. We use unified Random Forest with the same parameter to predict. The results are shown in Table 2.

Table 2. Performance of *T-DISTRIBTION* with different topic sizes (%)

Topic Size / Metric	10	15	20	25	30	35	40	45
Precision	63.71	68.79	70.54	71.92	75.54	74.62	73.67	73.92
Recall	64.40	69.53	71.26	72.54	75.09	74.97	74.29	74.33
F-measure	64.05	69.16	70.90	72.23	75.31	74.80	73.98	74.12

As shown in Table 2, with the increasing of the topic size, the precision of prediction keeps growing until the topic size is 30, and then has a slight decreasing trend. Thus we use 30 topics for the next experiments.

4.3 Performance on Different Feature Sets

In this section, we compare different feature sets by 3 methods. Table 3 shows the results on occupation prediction with different features. In this table, we use "*B-ALL*" to express all bag-of-words features including "*EMOTION*", "*ENGLISH*" and "*WORDS*", and we use *tfidf* to represent word feature. Similarly, we use "*N-ALL*" to express the combination of unigram and bigram, and "*T-ALL*" indicates the topic feature involving topic distribution and topic related words. We choose the best performance with various parameter candidates. Then, we list the accuracy of each occupation among the best global performance in Table 4. In the Table 3 we use "*P*" represents precision, "*R*" represents recall, and "*F*" represents F-measure).

Table 3. Performance on different feature sets (%)

Features \ Classifier		LR			SVM			Random Forest		
		P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
DIGITAL		34.56	34.58	34.57	34.08	33.66	33.87	45.52	45.62	45.57
BOW	26.49	26.54	26.52	37.04	36.31	36.67	40.36	40.33	40.34	40.34
	35.67	35.58	35.63	42.92	44.79	43.83	45.58	45.62	45.60	45.60
	73.26	72.25	72.75	74.71	75.45	75.08	77.18	76.79	76.98	76.98
	72.66	72.08	72.37	74.29	75.03	74.66	77.45	77.08	77.27	77.27
N-gram	62.95	62.29	62.62	69.71	70.38	70.04	73.08	69.71	70.04	70.04
	72.52	71.62	72.07	74.96	75.66	75.31	77.43	77.04	77.24	77.24
	73.17	72.50	72.83	74.83	75.58	75.21	77.18	76.58	76.88	76.88
Topic	65.15	64.50	64.82	71.58	72.71	72.14	77.00	76.46	76.73	76.73
	72.81	71.88	72.34	68.04	69.47	68.75	75.77	75.29	75.53	75.53
	67.05	66.46	66.75	74.42	75.46	74.94	77.56	77.08	77.32	77.32
ALL		75.42	74.67	75.04	75.95	75.21	75.58	78.42	77.92	78.17

According to Table 3, we can see that the topic feature on the fewer feature dimensions reaches higher accuracy except some high dimensional feature set like "*WORDS*" and "*BIGRAM*". The features reflecting the main theme of the text, for instance, "*EMOTION*" and "*ENGLISH*", can only reflect small crowd's behavior. The

Random Forest model performs best among the three classifiers when a fit parameter is used.

4.4 Performance on Different Feature Units

Table 4. Top 10 important features of *DIGITAL* (%)

Feature Unit	Contribution Value
avgForward	0.0482728
avgBookmark	0.0472507
favouritesCount	0.0421770
avgTopic	0.0412609
time-1	0.0357934
avgNews	0.0353702
avgZan	0.0352574
followerCount	0.0314369
statusesCount	0.0299582
biFollowerCount	0.0294038

Table 5. Most contributive topic words of *T-WORDS* (%)

Occupation	Most Contributive Topic Words
writer	波德莱尔(Baudelaire) 小说家(novelist) 文学史(history of literature) 本书(this book) 文坛(the literary world) 副刊(supplement) 主人公(leading character in a novel) 作品集(Portfolio)
reporter	新闻办(Information Office) 采访时(in the/an interview) 郭伟(Wei Guo) 本刊(this newspaper)
lawyer	律师事务所(law office) 许兰亭(LanTing Xu) 案件(case) 法援(legal aid) 案子(case) 世联(GlobalLink) 事务所(office) 诉讼费(legal fare)
photographer	摄像(camera shooting) 摄像机(vidicon) 拍照(take photos) 相片(photograph) 杂志(magazine) 照相(take pictures)
actor	演出(performance) 演员(actor) 丁军(Jun Ding) 童谣(Yao Tong) 新人(a new people) 米学东(XueDong Mi) 嘉宾(guest) 应采儿(CaiEr Ying) 星运(luck of star) 饰(portray)
singer	演唱(sing) 歌手集(sings) 民谣(balled) 作词(write lyrics) 歌词(lyric) 演唱会(vocal concert) 个人专辑(personal albums) 二胡(Erhu)
doctor	医疗(medical) 治疗学(acology) 患儿(Children patient) 医学部(Department of Medicine) 手术刀(scalpel) 医生(doctor) 医疗保险(medical insurance) 门诊部(out-patient department)
dietitian	营养学(nutriology) 食品(food) 吃富(eat something) 一杯(one cup) 餐单(menu) 食用(edible) 瘦体(thin body) 一罐(one tin)

There is a feasible method provided by *Scikit-learn*, which can compute the contribution of each feature unit in the detection process. Therefore, we use the parameter "*feature_importances*" of the model trained by *Random Forest*, and then

get the contribution of each feature. We list top 10 features of "*DIGITAL"* in Table 4, and show several greatest contributive topic words of "*T-WORDS"* in Table 5.

Table 4 shows that popularity and speech recognition play an important role in the detection, which are indicated by the feature "*avgForward*" and "*avgZan*". We can also identify a reporter due to the high rate of reprinting, reflected by the feature "*avgTopic*" and "*avgNews*". In addition, "*time-1*" is a special feature that means the number of the microblogs posted between twelve and one o'clock, indicating that the user sleeps much later. The results in Table 5 show the most contributive topic words of "*T-WORDS*". With this mechanism, we can build thesaurus for various occupations and select useful topic-related words.

4.5 Impact of the Lexicon Size

To verify the impact of the lexicon extension, we apply 300 dimensions of word embedding to represent remaining words, and integrate all vector representations by computing the sum of each dimension. What's more, in order to utilize more information such as *tf* (term-frequency) and *tfidf* (term frequency and inverse document frequency), we test three methods (as shown in Table 6): word-embedding, *tf* * word-embedding, and *tfidf* * word-embedding. In this experiment, we use the Logistic Regression model as the classifier. The results are shown in Table 6.

Table 6. Accuracy of different lexicon size (%)

#Words	Method	Accuracy
1537	word-embedding	77.54
	tf * word-embedding	75.08
	tfidf * word-embedding	73.46
3530	word-embedding	82.50
	tf * word-embedding	78.92
	tfidf * word-embedding	78.21
6480	word-embedding	80.62
	tf * word-embedding	78.04
	tfidf * word-embedding	75.83

As shown in Table 6, *tf* * word-embedding, and *tfidf* * word-embedding both lead to the decreasing of accuracy. We give two possible explanations. First, *tfidf* is helpful to improve the weight of rare words, because we have removed unimportant words previously. Second, combing the word-embedding feature with *tf* or *tfidf* is likely to result in the loss of some semantics.

The results in Table also show that the accuracy of classification increases with the increasing of the number of words. Thus, we can know that only two rounds of extension steps can extract appropriate lexicon words. Further, when we use the lexicon in user occupation detection, it is helpful to filter noise and improve the effectiveness of detection.

4.6 Lexicon-Based Word Embedding

In this section, we show the performance of word embedding with different dimensions using the lexicon consisting of 3530 words (see Table 6). We compare three classifiers and show the best result of each in Table 7.

As shown in Table 7, with the increasing of the word dimension, we get better performance of classification. It shows that combining the occupation-oriented lexicon with the word embedding method is able to achieve higher accuracy compared with traditional classification models. *LR* and *SVM* produce similar accuracy. Specially, *SVM* achieves the best accuracy when it is set to "*linear*". To this end, the linear model is more suitable for user occupation classification.

The best result for different user occupations are shown in Table 8. The "*lawyer*" gets highest accurate prediction, while the "*reporter*" is the hardest one to identify. It indicates that the lawyer and the doctor group incline to post messages related to their occupation, while others make less mention of occupation. Specially, reporters usually reprint various kind of news involving other field, thus making much noise of messages and increasing the difficulty to identify.

Table 7. Performance on different word dimension (%)

Classifier / Dimension	LR	SVM	Random Forest
100	76.79	76.46	64.04
200	80.21	79.96	66.21
300	82.50	82.33	67.04
400	83.62	83.58	68.21
500	84.25	84.62	67.50
600	84.58	85.46	67.29
700	85.42	86.29	68.67
800	85.42	86.21	68.62
900	85.12	86.54	68.08
1000	86.08	87.12	67.46

Table 8. Accuracy of each occupation among best performance (%)

Occup. / Metric	*writer*	*reporter*	*lawyer*	*photographer*	*actor*	*singer*	*doctor*	*dietitian*
Precision	78.33	77.33	92.67	82.00	79.67	83.00	86.00	90.00
Recall	78.20	78.24	92.21	82.14	81.02	82.04	87.46	87.52
F-measure	78.31	77.78	92.44	82.07	80.34	82.52	86.72	88.74

5 Related Work

Many previous work has been done to mine users' attributions on social media such as Twitter and Sina Weibo, including age [14, 19, 22], language [14], gender [18], location of origin [4], and political orientation [4]. Such previous work used a mixture of sociolinguistic features and n-gram models. Most attributes are inferred from the messages posted by users. In [24], the authors proposed a multi-source integration framework concentrating on building a feature set of contents and network information. Latent feature representation such as word clusters and embedding was used in [25] to classify occupations. In [10], the researchers proposed a classification method for detecting user occupations in microblogs. It uses the domain-specific features from the text, user behavior and social network. The work in [23] considered personal information, community structure and unlabeled data together to identify professions.

Topic model (e.g. LDA and PLSA) is very helpful in microblog analysis [20]. Some researchers adopted the Single-pass Clustering technique by using LDA to extract the hidden microblog topics information [11], and other researchers proposed a model named TweetLDA for Twitter classification tasks [13]. A bi-term topic model (BTM) was proposed for modeling topics in short texts [17]. This model is demonstrated to be helpful for Twitter analysis. Based on topic modeling, topic clustering was also studied in [16].

The work of this paper focuses on microblog user occupation detection and differs from previous researches on microblog user information extraction. We conduct an experimental study to evaluate the traditional feature-based classification models on user occupation detection. We fuse aggregated features according to user basic profile information and user messages, and adopt three classical linguistic models to extract features. Furthermore, we propose to build an occupation lexicon to improve the effectiveness of user occupation detection.

6 Conclusion

Detecting user occupation from microblog platform is an important issue for information extraction on short texts. This paper presents a framework for classifying eight occupations based on Sina Weibo, and addresses the task through three classifiers and the word embedding method. The result indicates that topic features perform well and we get the best results when using the Random Forest model. We also propose an occupation-oriented lexicon and integrate it with word embedding. The experimental results show that the lexicon-based features with lower dimension achieve higher accuracy compared with traditional models.

In future work, we will investigate deep learning for user occupation extraction on microblogging platforms. As deep learning and multi-layered neutral networks have been proven to be very effective in many applications such as image classification and retrieval [27], they are likely to have high performance on user occupation detection.

Acknowledgements

This work is supported by the National Science Foundation of China (61379037 and 71273010). Peiquan Jin is the corresponding author.

References

1. Zheng, L., Jin, P., Zhao, J., Yue, L.: A fine-grained approach for extracting events on microblogs, In: Decker, H., Lhotská, L. et al. (eds.) DEXA 2014, LNCS, vol. 8644, pp. 275–283. Springer, Heidelberg (2014)
2. Lv X, Jin P, Yue L. User occupation prediction on microblogs. In: Li, F., Shim, K., et al. (eds.) APWeb 2016, LNCS, vol. 9932, pp. 497–501. Springer, Heidelberg (2014)
3. Mislove, A., Viswanath, B., Gummadi, K., Druschel, P.:. You are who you know: inferring user profiles in online social networks. In: Third International Conference on Web Search and Web Data Mining (WSDM), pp. 251–260. ACM, New York, NY, USA (2010).
4. Rao, D., Yarowsky, D., Shreevats, A., Gupta, M.: Classifying latent user attributes in twitter. In: 2nd International Workshop on Search and Mining User-Generated Contents, pp. 37–44, ACM, Toronto, ON, Canada (2010)
5. Burger, J., Henderson, J., Kim, G., et al.: Discriminating gender on Twitter. In: 2011 Conference on Empirical Methods in Natural Language Processing, pp. 1301–1309. ACL, Stroudsburg, PA, USA (2011)
6. Pennacchiotti, M., Popescu, A.: A machine learning approach to Twitter user classification. In: Fifth International Conference on Weblogs and Social Media, pp. 281–288. The AAAI Press, Barcelona, Catalonia, Spain (2011).
7. Golbeck, J., Robles, C., Turner, K.: Predicting personality with social media. In CHI'11 Extended Abstracts on Human Factors in Computing Systems. pp. 253–262. ACM, Vancouver, BC, Canada (2011).
8. Pedregosa, F., Varoquaux, G., Gramfort, A., et al.: Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research 12, 2825–2830 (2011)
9. Rajaraman, A., Ullman, J.: Mining of massive datasets. Cambridge University Press. Cambridge, UK (2012).
10. Zhou, M., Xu, Y., Zhao, X.: Study of feature extract on microblog user occupation Classification. In: Fourth International Symposium on Information Science and Engineering, pp. 20–23. IEEE CS, Shanghai, China (2012)
11. Huang, B., Yang, Y., Mahmood, A., et al.: Microblog topic detection based on LDA model and single-pass clustering. In: Yao, J., et al. (eds.) RSCTC 2012. LNCS, vol. 7413, pp. 166–171. Springer, Heidelberg (2012).
12. Tinati, R., Carr, L., Hall, W., et al.: Identifying communicator roles in twitter. In: 21st International Conference on World Wide Web, pp. 1161–1168. ACM, Lyon, France (2012).
13. Quercia, D., Askham, H., Crowcroft, J.: TweetLDA: supervised topic classification and link prediction in Twitter. In: 4th Annual ACM Web Science Conference, pp. 247–250. ACM, Evanston, Illinois (2012).
14. Nguyen, D., Gravel, R., Trieschnigg, D., et al.: " How Old Do You Think I Am?" A study of language and age in Twitter. In: Seventh International AAAI Conference on Weblogs and Social Media. pp. 439–448. The AAAI Press, Cambridge, Massachusetts, USA (2013).

15. Schwartz, H., Eichstaedt, J., Kern, M., et al.: Personality, gender, and age in the language of social media: The open-vocabulary approach. PLoS One 8(9), e73791 (2013)
16. Zhao, J., Li, X., Jin, P.: A time-enhanced topic clustering approach for news web search, International Journal of Database Theory and Application 5(4), 1–10 (2012)
17. Yan, X., Guo, J., Lan, Y., et al.: A biterm topic model for short texts. In: 22nd International conference on World Wide Web, pp. 1445–1456. ACM, Rio de Janeiro, Brazil (2013).
18. Huang, F., Li, C., Lin, L.: Identifying gender of microblog users based on message mining. In: Li, F., Li, G., et al. (eds.) WAIM 2014. LNCS, vol. 8485, pp. 488–493. Springer, Cham (2014)
19. Li, Y., Liu, T., Liu, H., et al.: Predicting microblog user's age based on text information. In: Lin, X., Manolopoulos, Y., et al. (eds.) WISE 2013. LNCS, vol. 8180. pp. 510–515. Springer, Berlin, Heidelberg (2014)
20. Yang, S., Kolcz, A., Schlaikjer, A., et al.: Large-scale high-precision topic modeling on twitter. In: 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1907–1916. ACM, New York, USA (2014).
21. Wu, X., Wang, J.: Micro-blog in China: identify influential users and automatically classify posts on Sina micro-blog. Journal of Ambient Intelligence and Humanized Computing 5(1), 51-63 (2014)
22. Chen, L., Qian, T., Wang, F., et al.: Age detection for Chinese users in Weibo. In: Dong, X., Yu, X., et al. (eds.) WAIM 2015. LNCS, vol. 9098, pp. 83–95. Springer, Cham (2015).
23. Tu, C., Liu, Z., Sun, M.: PRISM: Profession identification in social media with personal information and community structure. In: Zhang, X., Sun, M., et al. (eds.) CNCSMP 2015. CCIS, vol. 568, pp. 15–27, Springer, Singapore (2015).
24. Huang, Y., Yu, L., Wang, X., et al.: A multi-source integration framework for user occupation inference in social media systems. World Wide Web 18(5), 1247–1267 (2015).
25. Preoţiuc-Pietro, D., Lampos, V., Aletras, N.: An analysis of the user occupational class through Twitter content. In: 53rd Annual Meeting of the Association for Computational Linguistics, pp. 1754–1764. ACL, Beijing, China (2015).
26. Scikit-Learn, http://scikit-learn.org/stable/, last accessed 2017/04/21.
27. Wan, S., Jin, P., Yue, L.: An approach for image retrieval based on visual saliency, In: 2009 International Conference on Image Analysis and Signal Processing, pp. 172–175. IEEE CS, Linhai, China (2009).