Discriminating Gender on Chinese Microblog: A Study of Online Behaviour, Writing Style and Preferred Vocabulary

Li Li, Maosong Sun, Zhiyuan Liu
State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology,Tsinghua University, Beijing 100084, China
happylily0516@gmail.com, sms@tsinghua.edu.cn, lzy.thu@gmail.com

***Abstract*—As user attributes are useful for applications such as personalized recommendation, adverting and so on, user attribute predication on Twitter has attracted intensive attentions in recent years. Although Chinese micro-blogging services are different from Twitter on various aspects such as language, user behaviours and so on, few efforts have been made on Chinese micro-blogging services. In this paper, we propose a gender prediction model for Chinese microblog which exploits features including online behaviour, writing style, and preferred vocabulary. Experimental results on Sina Weibo, which is one of the most popular micro-blogging services in China, show that our model achieves the state-of-the-art accuracy 94.3%. We also find significant distinctions between male and female microblog users on online behaviour, writing style and preferred vocabulary, which would be helpful for improving personalized applications.**

***Keywords*—gender prediction, Chinese microblog, user behaviour analysis.**

I. Introduction

As micro-blogging services, e.g. Twitter[1], limit the maximum length of each message, they make people to post and share messages more frequently and easily. As a result, micro-blogging services have revolutionized the way people access and share information and interact with each other [1]. Since microblog messages and user data are useful for improving applications such as personalized recommendation, adverting, it has attracted intensive attentions on microblog related tasks, including POS tagging [2], entity linking [3], text normalization [1], opinion analysis [4], [5], and user attribute prediction [6], [7], [8], [9], [10].

However, most of existing efforts only focus on English micro-blogging services. It is not straightforward to apply algorithms and models on Chinese micro-blogging services. Despite that Chinese and English micro-blogging services are similar in certain aspects, there are significant distinctions between them. An obvious distinction is the major languages used by the users. Taking the most popular English micro-blogging service Twitter and the most Chinese micro-blogging service Sina Weibo[2] as examples. Although both of them limit the maximum length of each message to 140 characters, generally a Sina Weibo message takes more information than a Twitter message because a single Chinese character is more informative than a single English character. [11]. As a result, the user behaviour on Chinese and English micro-blogging services is different [12].

We focus on the task of gender prediction in this work. As gender is not included in Twitter user profile, it is difficult to build Twitter training dataset. One common method for resolving this problem is human annotation [7], [9], e.g. using Amazon Mechanical Turk[3]. Other researchers detect gender information of users based on their screennames [13], [14]. Obviously, neither of these two types of methods can achieve an accuracy of 100% for a large corpus. Luckily, gender is included in the Sina Weibo user profile, and the personal information for certificated accounts is also available. Thus, the correctness of gender information for certificated accounts can be guaranteed on Sina Weibo. Therefore, the accuracy of our gender prediction model on Sina Weibo may be a good reference for similar tasks on Twitter to some extent.

In this work, we propose a gender classifier for Chinese micro-blogging services. To our knowledge, it is the first time to conduct gender prediction on Chinese micro-blogging services. Three kind of features are considered in our classifier, i.e. online behaviour, writing style and preferred vocabulary of microblog users. Experimental results on Sina Weibo dataset show that our method achieves an accuracy of 94.3%, which is state-of-the-art. Further analysis reveals that there are significant distinctions between the behaviour of male and female microblog users. For example, males are more likely to forward messages than females on Sina Weibo. These findings will be helpful for improving personalized applications, psychological research and sociological research.

II. Related Work

Inferring attributes of online users has already attracted intensive attentions during the glorious period of blog. Schler et al. [15] obtained dataset including over 71,000 blogs from blogger.com. By analysis of dataset, they found significant distinctions in writing style and content between males and females, then they took them as features to predict gender and age information of users. MacKinnon et al. [16] predicted gender and geographic information of a user using the gender

[1] www.twitter.com

[2] www.weibo.com

[3] www.mturk.com

TABLE I. SINA WEIBO USER PROFILE OF KAIFU LI. MOST OF THE FIELDS ARE OPTIONAL.

Field	Value
ID	1197161814
Screen Name	Kaifu Li
Gender	Male
# Fans	9,343,276
# Attentions	259
# Messages	3,764
# Tags	Venture Capital, Education, Innovation Works and etc.
Personal Description	CEO of Innovation Works
Certificated Information	Chairman and CEO of Innovation Works

and geographic information provided by the friends of the user. Goswami et al. [17] predicted gender and age information employing unified features, including non-dictionary words, content words and average length of sentences.

Nowadays, along with the development of microblog, inferring user attributes of microblogers has also attracted intensive attentions. Rao et al. [18] predicted gender, age, regional origin and political orientation of users on Twitter. The features they extracted from tweets included sociolinguistic features (such as emoticons) and ngram features. In [6], the authors used word-level and character-level ngrams features extracted from user profiles, including screen name, full name, description and tweets. Besides, because the gender information is latent on Twitter, they use gender information provided by blog to annotate their dataset. That is, they followed the personal URLs, which were provided in profiles of Twitter users and linked to blog sites, to determine the gender of Twitter users. Ciot et al. [7] were the first to do gender inference in non-English contexts. They used rich features, including words, bigrams, trigrams, hashtags, mentions, together with tweet/retweet/hashtag/link/mention frequencies and out/in-neighborhood size. However, their experimental corpus contains only thousands of twitter users while Chinese language is not included in the non-English language set.

III. SINA WEIBO

Micro-blogging services are also very popular in China. According to the report of China Internet Network Information Center (CNNIC) [19], there were more than 250 million registered microblog users, which are nearly 50% of all Internet users in China. Sina Weibo is one of the most famous and popular micro-blogging services in China. It is reported that there are more than 200 million messages per day on Sina Weibo [20]. As a result, we conduct our experiments on Sina Weibo dataset.

As a typical micro-bloging service, Sina Weibo permits users to write no more than 140 characters in each message. And users can insert pictures, emoticons, URLs, hashtags (labeled by ##) to messages. Sina Weibo provides three ways for users to interact with each other:

- Commenting others' messages;
- Forwarding others' messages;
- Mentioning another user by "@username".

Sina Weibo also supports one user to follow other users, which is similar to Twitter. We refer to the users who follow a specific user as fans of the user.

On Sina Weibo, user can post original messages and forward others' messages. We will refer to original message as **OM** and forwarded message as **FM**. Preliminary observation reveals that OMs tend to reflect the daily life of a user and FMs tend to reflect the focus or interests of a user. Therefore, it is necessary to taking OMs and FMs as two kinds of contexts.

Each Sina Weibo user has a profile. The profile has several fields, including id, screen name, gender, fans, attention, messages, tags, personal description, etc. Most of the fields are optional. Table I shows the profile of Kaifu Li. Sina Weibo also provides a certification service. For example, Kaifu Li is certificated as the chairman and CEO of Innovation Works, which a famous company in China. We will refer to this kind of users as **certificated users**. Certificated users tend to provide more information in their profile, and the information is more reliable.

IV. GENDER PREDICTION

In this paper, we treat gender prediction as a binary classification problem, and use SVM-based (Support Vector Machine) method to build the classification model. Three kinds of features are used in our classifier, i.e. online behaviour features, writing style features, and preferred vocabulary features.

A. Online Behaviour Features

Online behaviour features mainly reflect user behaviours on Sina Weibo, including influence, activity, habit of posting messages and interaction with others. The detailed contents of online behaviour are shown in Table II. For example, "# OMs/ # FMs" reflects users' tendency to post original messages or forward others' messages.

TABLE II. ONLINE BEHAVIOUR FEATURES, WHICH REFLECT USER BEHAVIOURS ON SINA WEIBO.

Features	Explanation
# Fans	Number of fans
# Attentions	Number of attentions
# Messages	Number of Messages
#OMs/#FMs	Number of OMs divides number of FMs
# Comments	Number of comments per OM
# Forward	Number of forward per OM

B. Writing Style Features

Writing style features mainly reflect the user styles when posting messages on Sina Weibo, including the favor of inserting emoticons, hashtags, URLs, pictures, and interaction with others (# @username), etc. "Sen. Length" reflects preference of posting long or short messages. Table III shows the details of writing style. Moreover, as such features of OMs reflect writing style of users themselves while FMs reflect writing style of the authors of the FMs, we calculate writing style features of OMs and FMs separately.

TABLE III. WRITING STYLE FEATURES.

Features	Explanation
# Emoticons	Number of emoticons per message
# Hashtags	Number of hashtags per message
# URLs	Number of URLs per message
# Pictures	Number of pictures per message
# @username	Number of @username per message
# Sen. Length	Average sentence length of messages

C. *Preferred Vocabulary Features*

Preferred vocabulary features are lexical features extracted from tags, personal description, certificated information, OMs and FMs. We employ bigrams other than results of Chinese Word Segmentation (CWS) after comparing their performances on gender prediction. Besides, we choose $TF - IDF$ (term frequency-inverse document frequency) to weight these bigram features after comparing performances of TF (term-frequency), DF (document frequency) and $TF - IDF$. As χ^2 method has been proved to be one of the best methods to do feature selection [21], we use χ^2 method to do feature selection in our gender classifier.

In our experiment, we calculate $TF - IDF$ value of word for each user u and the formula is shown in (1). There $tf_{w,u}$ means the times w appearing in u's messages. D_u means the number of messages of u. $df_{w,u}$ means the number of messages containing w of u.

$$TFIDF_{w,u} = \frac{tf_{w,u}}{\sum_w tf_{w,u}} \times \log(D_u/df_{w,u}) \qquad (1)$$

For our normal two classification problem, the formula of χ^2 is shown in (2). There $\mathbb{D}$ means the training dataset, t means term while c means class. Thus, the subscript e_t means the appearing or not of term (1 means appear, 0 means not) while the subscript e_c means the type of class (1 means female, 0 means male). N means the document frequency in $\mathbb{D}$ while E means the expectation.

$$\chi^2(\mathbb{D}, t, c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}} \qquad (2)$$

The number of features of each kind are summarized in Table IV.

TABLE IV. FEATURES SIZES

Feature	Number
Online behaviour	6
Writing style	6 from OMs 6 from FMs
Preferred vocabulary	100 from Tags 100 from Personal Description 100 from Certificated Information 10,000 from OMs 10,000 from FMs
Total	20,318

V. EXPERIMENTS

A. *Dataset*

As stated in Section III, the information provided by certificated users are more reliable, we get the profiles and messages of 24,950 randomly chosen certificated users using the APIs provided by Sina Weibo as our dataset. The dataset is split into training and test set randomly. The training set consists of the profiles of 20,000 users and their 46,841,545 messages. The rest 4,950 profiles and 9,051,364 messages are used as test set.

TABLE V. THE PERFORMANCE OF GENDER PREDICTION MODEL. A-ACCURACY.

Classifier	A(Train)	A(Test)
NB	0.814	0.815
C4.5	0.990	0.810
LR	0.947	0.925
SVM	**0.990**	**0.943**
Human annotation	-	0.802

TABLE VI. THE DETAILED PERFORMANCE OF SVM-BASED METHOD. A-ACCURACY.

Field	A(Train)	A(Test)
Online behaviour	0.555	0.558
Writing style	0.699	0.720
Preferred vocabulary	0.983	0.942
All three fields	**0.990**	**0.943**

B. *Evaluation Results*

Table V shows the performance of our classifier on training and test dataset. We execute our experiment on Weka [22] and LIBLINEAR [23] to compare the performances of typical binary classification, including NB (Naive Bayes), C4.5 (Decision Tree), LR (Logistic Regression) and SVM. Finally, we find that SVM-based method gets the highest accuracy of 94.3% on this task. To our knowledge, the best result of gender prediction on Twitter is 91.9%, which is also lower than ours.

Besides, as we are the first to predict gender attribute on Chinese micro-blogging services, we take human annotation as the baseline, which is shown in the last line in Table V. As it is time consuming to annotate all the data, we randomly select 101 certificated users from test dataset and ask two professional students to label the gender information. Finally, they gain an accuracy of 80.2% with a Pearson correlation coefficient 0.785. Thus, we can come to a conclusion that our prediction model is superior to human annotation. This phenomenon shows that small amount of typical lexical features is inadequate for predicting gender attribute of Sina Weibo User, while handling large amount of features is the preponderance of machine learning.

Moreover,We show the detailed performance of SVM-based method to compare the effectiveness of three categories of feature in Table VI. There, "A" also is the abbreviation of accuracy. We show the separate performance of each feature category in the first three lines, and combine all the three feature categories in "All three fields". We can see that using preferred vocabulary features achieves the best performance, which is closed to the accuracy of combining these three kinds of features.

VI. ANALYSIS

We analyze gender behaviour differences in this section. We find that male and female behaviours are quite different with respect to online behaviour, writing style and preferred vocabulary.

A. *Online Behaviour*

Figure 1 shows the gender behaviour differences on online behaviour, where M/F refers to the feature value of male users divides that of female users. From observation of Figure 1, we can see that, for all the features on online behaviour, males'

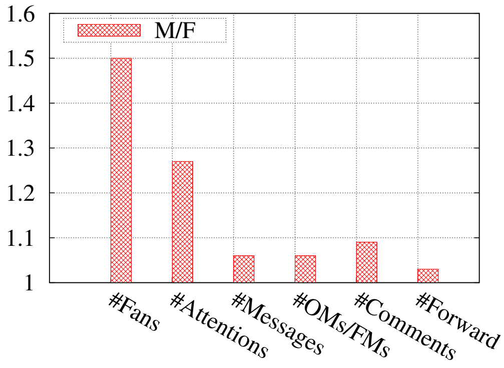

Fig. 1. Gender behaviour differences on online behaviour. M/F refers to the feature value of male users divides that of female users.

value is higher than females', especially in # Fans. Thus, we can come to the following three conclusions.

1) According to the bias to males in # Fans, # Attention and # Messages, males' activity and influence are greater than females' on Sina Weibo.
2) According to the bias to males in # OMs/# FMs, males are more likely to post original messages than females.
3) Considering the bias to males in # Comments and # Forward, the messages of males are more likely to be forwarded than females'.

B. Writing Style

Writing style feature quantitatively calculate the writing style of users on Sina Weibo, and as such features of OMs reflect writing style of users themselves while FMs reflect writing style of the authors of the FMs, we calculate writing style of OMs and FMs separately. Figure 2 shows the gender behaviour differences on Writing style of both OMs and FMs.

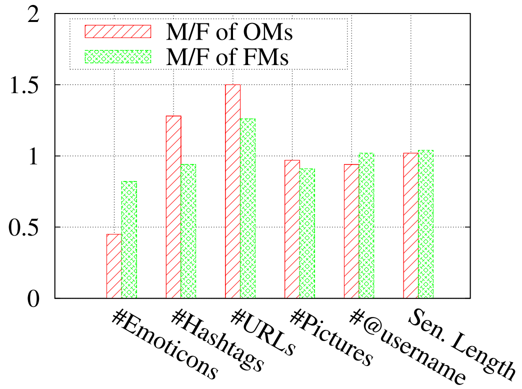

Fig. 2. Gender behaviour differences on writing style. M/F refers to the feature value of male users divides that of female users.

From gender behaviour differences on writing style of OMs, we can see that the values are biased to females for features including # Emoticons and # @username. However, for features including # URLs and # Hashtags, the results are inverse. Thus, we come to the following four conclusions.

1) Females insert more emoticons into their messages, thus style of OMs of females is more perceptual and lively than males.
2) Females are more preferred to interact with others by way of @username than males.
3) Hashtags usually appear as topics of messages on Sina Weibo [5], thus the bias to males in # Hashtags indicates that males are more likely to directly display topics of their messages.
4) Males are more likely to share information in messages in the format of URL.

Moreover, to summarize the gender behaviour differences on writing style of FMs, we come to similar conclusions of OMs. On the one hand, females are preferred to forward messages with more emoticons, thus the style of FMs of females are more perceptual and lively than that of males. On the other hand, males are more likely to forward messages sharing information in the format of URL.

Thus, we find a phenomenon that the conclusions of writing style of OMs and FMs are similar. Thus we suppose whether males are more likely to forward messages from males while females are more likely to forward messages from females. To demonstrate this hypothesis, we quantitatively calculate the forward behaviour of Sina Weibo users. That is, we calculate the distribution of forwarded messages on Sina Weibo. Table VII shows the fact is consistent with our assumption.

TABLE VII. THE DISTRIBUTION OF FMS.

Gender	# FMs from Males	# FMs from Females
Male	88.3%	11.7%
Female	38.7%	61.3%

C. Preferred Vocabulary

At last, we want to analyze gender behaviour differences in terms of preferred vocabulary from tags, personal description, certificated information, OMs and FMs. In order to observe the differences intuitively and plainly, we don't use the bigrams but words after CWS as preferred vocabulary. As most results of bigram are not words and it is hard to understand the meaning behind them. Moreover, as tags are already separated by marks in our dataset, there is no need to conduct CWS on tags any more, Thus we only use THULAC [24] to conduct CWS on personal description, certificated information, OMs and FMs.

Before introducing the definition of preferred vocabulary, we need to explain the definition of CR(coverage rate), whose calculation is based on UF(user frequency). Let's take males' tags information as an example of calculation of CR. N is the total number of unique words (word's length ≥ 2) of tags of males, and the words are w_1,w_2,...,w_N. $UF_M(w_i)$ is the user frequency in males for tags. $T_M = \sum_i UF_M(w_i)$. Then for a subset of vocabulary, $S = w_{i1}, w_{i2}, ..., w_{in}$, the CR is defined as

$$CR(S) = (UF_M(W_{i1})+UF_M(W_{i2})+...+UF_M(W_{in})))/T_M \quad (3)$$

For each CR value β, the corresponding size of vocabulary is the number of words of the smallest subset that satisfy $CR(S) \geq \beta$. According to such definition, we get Table VIII.

TABLE VIII. SIZE OF VOCABULARY FOR TYPICAL CR.

CR	Gender	Tags	Personal Description	Certificated Information	OMs	FMs
100%	Male	30,323	22,884	17,914	2,652,710	2,854,354
	Female	32,279	20,177	16,511	2,053,393	2,741,016
	Common	6,187	8,545	4,602	730,546	1,378,082
80%	Male	14,413	7,165	3,789	39,066	46,670
	Female	17,101	6,367	3,209	30,229	42,867
	Common	3,755	3,958	1,900	28,941	39,717
50%	Male	821	957	216	7,051	10,288
	Female	1,249	908	188	5,679	9,487
	Common	570	673	165	5,299	8,749
20%	Male	27	100	27	1,278	2,214
	Female	28	88	22	1,056	2,046
	Common	18	59	18	975	1900

From Table VIII, we can see that the size of vocabulary of males' tags is bigger than males for the same CR. However, for personal description, certificated information, OMs and FMs, the conclusion is reverse. This indicates that females' wording of tags is more diverse than that of males'. However, for personal description, certificated information, OMs and FMs, males' wording is more diverse than females'. Besides, from all the typical CR listed in Table VIII, we can see that males and females have a certain number of non-overlapping words. And we suppose these non-overlapping words reflect gender behaviour differences on preferred vocabulary to some extent. Thus, we visualize these top 20 (rank according to UF) non-overlapping words of males and females from tags (green), personal description (red), certificated information (blue), OMs (purple) and FMs (yellow) in the format of word cloud in Figure 3 and Figure 4. In these two figures, the sizes of words change along with UF while the color shades of words change along with correlation getting from gender prediction model.

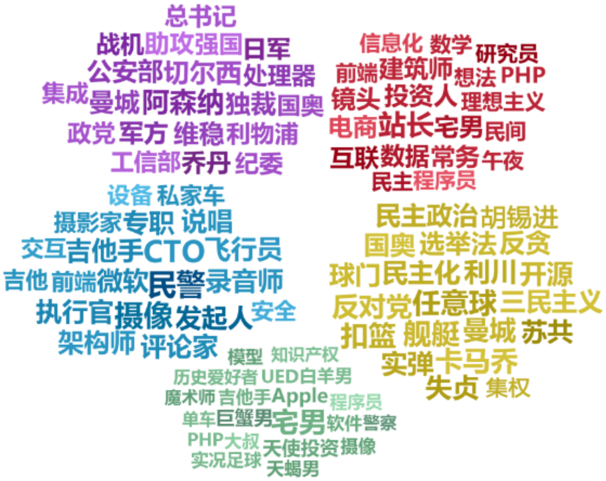

Fig. 3. Preferred vocabulary of males.

From Figure 3 and Figure 4, we can see that there are obvious gender behaviour differences on preferred vocabulary. Let's start with tags tags, personal description. As they are mainly personal statement in common. For tags, males use words like "知识产权" (intellectual property),"吉他手" (guitar),"魔术师" (magician),"程序员" (programmer),"警察" (police),"投资" (investment),"摄像" (camera shooting)," 天使投资" (angel investment), etc while females use "粉红控" (pink complex),"包包控" (bag complex),"hellokitty控" (hellokitty complex),"高跟鞋控" (high-heeled shoes complex),

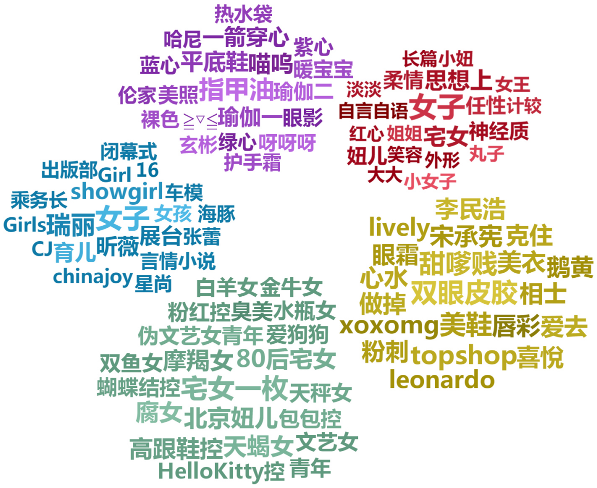

Fig. 4. Preferred vocabulary of females.

etc. For person description, males use words like "投资人" (investor), "建筑师" (architect), "研究员" (researcher),"站长" (webmaster),"程序员"(programmer),"前端"(front end),"数学" (math),"PHP","电商" (e-business),"互联"(internet), etc while females use words like "女王" (queen),"女子"(female),"自言自语" (soliloquize),"神经质" (neurotic),"任性" (capricious), etc. We can see that males usually show their position or industry in personal statement while females more likely to display their character.

OMs and FMs mainly reflect user's interest on Sina Weibo. in OMs, males use words like "总书记" (the general secretary),"独裁" (dictatorship),"政党" (political party),"军方" (the military),"维稳" (stability maintenance),"纪委" (the commission of discipline),"助攻" (assist),"曼城" (Manchester United),"阿森纳" (Arsenal),"切尔西" (Chelsea), etc while females user "指甲油" (nail polish),"平底鞋" (flat shoes),"瑜伽" (yoga),"眼影" (eye shadow),"护手霜" (hand cream),"玄彬" (Hyun Bin), etc. In FMs, the phenomenon is similar. Males use words like "民主" (democratic),"政治" (politics),"选举法" (the electoral law)," 反贪" (anti-corruption),"集权" (centralization of power),"反对党" (opposition),"球门" (goal), "扣篮" (dunk),"曼城" (Manchester United),"卡马乔" (Camacho), etc while females use words like "眼霜" (eye cream)、"粉刺" (acne),"美鞋" (shoes),"双眼皮胶" (double eyelid),"唇彩" (lip gloss),"李民浩" (Lee Minho),"宋承宪" (Song Seung Heon),"leonardo", etc. Theses distinctions indicate that males are interested in politics and football games while females are interested in beauty and stars.

Then we compare the differences on preferred vocabulary for verified information. For Sina Weibo users, verified information is the description of positions , thus distinctions of verified information between males and females indicate the differences of industry directly. Males use words like “摄影家” (photographer),“吉他手” (guitar),“ 飞行员” (pilot),“执行官”(CTO),“架构师” (architect),“评论家” (commentator), etc while females use words like “ 乘务长” (purser),“车模” (car model),“昕薇” (Xi Vi),“瑞丽” (RAYLI),“展台” (exhibition hall), etc. These distinctions reflect that, for Sina Weibo verified users, males usually work for IT, music, etc while females are usually employed in fashion.

VII. Conclusion

In this paper, we propose a gender classifier for Chinese micro-blogging service. The classifier uses three kinds of features, i.e. online behaviour features, writing style features, and preferred vocabulary features. We have build a dataset with 24,950 certificated users on Sina Weibo. Experimental results show that our classifier achieves an accuracy of 94.3%, which is superior to human. Moreover, we analyze user gender behaviour differences from these three aspects in details, and find obvious distinctions between males and females. Such as the writing style of females is more lively and perceptual than males. These distinctions are significant for personalized recommendation and personalized advertisements.

Acknowledgment

This research is supported by the National Natural Science Foundation of China (General Program, Grant No. 61170196), and the National Social Science Foundation of China (Major Program, Grant No. 13&ZD190).

References

[1] R. Yan, M. Lapata, and X. Li, “Tweet recommendation with graph co-ranking,” in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, 2012, pp. 516–525.

[2] K. Gimpel, N. Schneider, B. O’Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith, “Part-of-speech tagging for twitter: Annotation, features, and experiments,” DTIC Document, Tech. Rep., 2010.

[3] Y. Guo, B. Qin, T. Liu, and S. Li, “Microblog entity linking by leveraging extra posts,” in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 863–868.

[4] S. Feng, L. Zhang, B. Li, D. Wang, G. Yu, and K.-F. Wong, “Is Twitter a better corpus for measuring sentiment similarity?” in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 897–902.

[5] X. Zhou, X. Wan, and J. Xiao, “Collective opinion target extraction in Chinese microblogs,” in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1840–1850.

[6] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella, “Discriminating gender on twitter,” in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 2011, pp. 1301–1309.

[7] M. Ciot, M. Sonderegger, and D. Ruths, “Gender inference of Twitter users in non-English contexts,” in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1136–1145.

[8] M. Marchetti-Bowick and N. Chambers, “Learning for microblogs with distant supervision: Political forecasting with twitter,” in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 2012, pp. 603–612.

[9] D. Nguyen, R. Gravel, D. Trieschnigg, and T. Meder, “ “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*, 2013.

[10] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman *et al.*, “Personality, gender, and age in the language of social media: The open-vocabulary approach,” *PloS one*, vol. 8, no. 9, 2013.

[11] S. Chen, H. Zhang, M. Lin, and S. Lv, “Comparision of microblogging service between sina weibo and twitter,” in *Computer Science and Network Technology (ICCSNT), 2011 International Conference on*, vol. 4, Dec 2011, pp. 2259–2263.

[12] Q. Gao, F. Abel, G.-J. Houben, and Y. Yu, “A comparative study of users’ microblogging behavior on sina weibo and twitter,” in *User modeling, adaptation, and personalization*. Springer, 2012, pp. 88–101.

[13] C. Fink, J. Kopecky, and M. Morawski, “Inferring gender from the content of tweets: A region specific example.” in *ICWSM*, 2012.

[14] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist, “Understanding the demographics of twitter users,” in *ICWSM*, 2011.

[15] J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker, “Effects of age and gender on blogging.” in *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 2006, pp. 199–205.

[16] I. MacKinnon and R. H. Warren, “Age and geographic inferences of the livejournal social network,” in *Statistical Network Analysis: Models, Issues, and New Directions*, 2007, pp. 176–178.

[17] S. Goswami, S. Sarkar, and M. Rustagi, “Stylometric analysis of bloggers’ age and gender,” in *Third International AAAI Conference on Weblogs and Social Media*, 2009.

[18] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, “Classifying latent user attributes in twitter,” in *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, 2010, pp. 37–44.

[19] CNNIC, “29th statistical survey report on the internet development in china,” 2012. [Online]. Available: http://www.cnnic.cn/research/bgxz/tjbg/201201/t20120116_23668.html

[20] L. Xu, K. Liu, S. Lai, Y. Chen, and J. Zhao, “Mining opinion words and opinion targets in a two-stage framework.”

[21] Y. Yang and J. O. Pedersen, “A comparative study on feature selection in text categorization,” in *ICML*, vol. 97, 1997, pp. 412–420.

[22] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The weka data mining software: an update,” *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[23] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, “LIBLINEAR: A library for large linear classification,” *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[24] K. Zhang and M. Sun, “A stacked model based on word lattice for chinese word segmentation and part-of-speechtagging.” [Online]. Available: http://nlp.csai.tsinghua.edu.cn/thulac