Leveraging Interactive Knowledge and Unlabeled Data in Gender Classification with Co-training

Jingjing Wang[1,2] Yunxia Xue[1,2] Shoushan Li[1,2*] Guodong Zhou[1,2]

[1] Natural Language Processing Lab, School of Computer Science and Technology, Soochow University, China
[2] Collaborative Innovation Center of Novel Software Technology and Industrialization

{djingwang,yunxia.xue,shoushan.li}@gmail.com
gdzhou@suda.edu.cn

Abstract. Conventional approaches to gender classification much rely on a large scale of labeled data, which is normally hard and expensive to obtain. In this paper, we propose a co-training approach to address this problem in gender classification. Specifically, we employ both non-interactive and interactive texts, i.e., the *message* and *comment* texts, as two different views in our co-training approach to well incorporate unlabeled data. Experimental results on a large data set from micro-blog demonstrate the appropriateness of leveraging interactive knowledge in gender classification and the effectiveness of the proposed co-training approach in gender classification.

Keywords: interactive knowledge, gender classification, co-training

1 Introduction

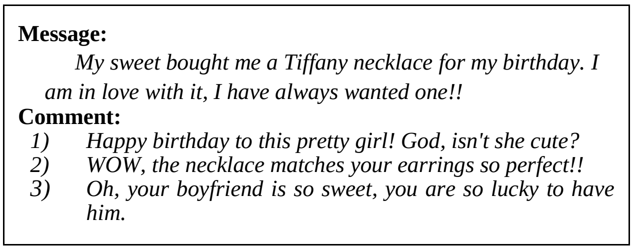
Gender classification, a fundamental task in social media analysis, aims to predict the user gender with the user-generated data. With the rapid growth of social media in recent years, gender classification has been drawing more and more attention for a wide range of real-life applications, such as intelligent marketing, personalization prediction, automatic advertising, and sentiment analysis (Mukherjee and Liu, 2010; Burger et al., 2001; Volkova et al., 2013).

On one hand, conventional approaches conceptualize gender classification as a supervised learning problem and rely on human-annotated data for model learning. Such supervised approaches have delivered reasonable performance. However, the reliance on labeled data, which is normally hard and expensive to produce, presents a major obstacle to the widespread application of gender classification.

On the other hand, we notice that most of previous studies in gender classification focus on exploring text knowledge to infer user's gender with a statistical text classifier (Nowson and Oberlander, 2006; Ciot et al., 2013). Although these studies have achieved certain success in gender classification, the utilized knowledge is always limited to non-interactive text, i.e., generated from the given user. In fact, as a well-established platform for user interaction, social media not only provides a public platform for a user to publish his own experiences and opinions, but also offers an effec-

[*] Corresponding author

tive channel for other users to feedback through some interactive mechanisms, e.g., allowing a user to write a *comment* as a feedback to certain *message*. Such interactive process normally generates some kinds of interactive text (e.g., the *comment*) which provides a complementary resource to infer the gender of the message-publishing user.



Message:

My sweet bought me a Tiffany necklace for my birthday. I am in love with it, I have always wanted one!!

Comment:

1) *Happy birthday to this pretty girl! God, isn't she cute?*
2) *WOW, the necklace matches your earrings so perfect!!*
3) *Oh, your boyfriend is so sweet, you are so lucky to have him.*

Figure. 1. An example from Twitter with the *message* text and the corresponding *comment* text

Figure 1 illustrates an example from Twitter with the *message* text and the *comment* text. From the *comment* text, a kind of interactive knowledge, it is easy to infer the message-publishing user to be *female* from these descriptions, such as *pretty girl*, *your earrings* and *your boyfriend*, while it may be not obvious from the *message* text.

In this paper, we focus on exploiting interactive knowledge, i.e., the *comment* text in social media, in gender classification. With the help of such knowledge, we propose a co-training approach to well incorporate unlabeled data and alleviate the high reliance of gender classification on labeled data by casting non-interactive and interactive texts as two views.

2 Related Work

Previous studies in gender classification mainly adopt supervised learning approaches on different text styles, such as Blog (Peersman et al., 2010; Gianfortoni et al., 2011), E-mail (Corney et al., 2002), YouTube (Filippova, 2012) and Micro-blog (Rao et al., 2010; Liu et al., 2013). Specifically, besides those standard features such as character, word, and POS features, these studies focus on exploring more effective features for gender classification. For example, Nowson and Oberlander (2006) validate that using context-based n-grams tends to be more accurate than using dictionary-based features in weblog. Mukherjee and Liu (2010) propose some POS pattern features to improve the classification performance.

In comparison, there are only a few previous studies in semi-supervised gender classification. Ikeda et al. (2008) propose a semi-supervised approach to gender classification in blog. Their main idea is to utilize a sub-classifier to measure the relative similarity between two blogs so as to capture the classification knowledge in the unlabeled data. More recently, Burger et al. (2011) mention the importance of using unlabeled data and directly apply a self-training approach to perform semi-supervised learning for gender classification.

Different from above studies, our study focuses on interactive knowledge for gender classification, which has not been explored before. Furthermore, our co-training approach to semi-supervised gender classification is based on two views. Experimental results show that our two-view co-training approach is much more effective than the single-view self-training approach.

3 Gender Classification with Co-training

In supervised learning, a predictor *f* is trained to map an input vector *x* into a class label *y*. In this paper, the input vector *x* is the feature representation, generated from either the *message* text or the *comment* text.

Formally, the objective of gender classification is illustrated as follows:

$$f(x) \to y \qquad (1)$$

$$\text{Where } \mathrm{y} \in Y \text{ and } Y = \{male,\ female\}$$

In the literature, a variety of effective features have been proposed for gender classification. In this paper, we adopt following features due to their good performance in previous studies, e.g., Mukherjee and Liu (2010).

(1) **Bag-of-words features**: These are basic word features. This kind of basic features is popularly utilized in not only gender classification but also many other NLP tasks where text knowledge is leveraged.

(2) **F-measure feature**: It is defined to capture the POS usage. As a unitary measure of text's relative contextuality, this feature explores the notion of implicitness of text (Heylighen and Dewaele, 2002).

(3) **POS sequence patterns**: These features are extracted from POS sequence of the text (Mukherjee and Liu, 2010) to capture the writing styles of different genders.

```
Input:  $L_{Mes}$: labeled message samples; $L_{Com}$: labeled comment samples
        $U_{Mes}$: unlabeled message samples; $U_{Com}$: unlabeled comment samples
Output: $L_{Mes}$: New labeled message samples; $L_{Com}$: New labeled comment samples
Procedure: Loop for $N$ iterations until $U_{Mes} = \varnothing$ or $U_{Com} = \varnothing$
(1).  Learn classifier $C_M$ with $L_{Mes}$
(2).  Use $C_M$ to label the samples from $U_{Mes}$
(3).  Choose $n_1$ positive and $n_1$ negative message samples $M_1$ most confidently predicted by $C_M$
(4).  Choose corresponding comment samples $S_1$ (the comment samples from the same users in $M_1$)
(5).  Learn classifier $C_C$ with $L_{Com}$
(6).  Use $C_C$ to label the samples from $U_{Com}$
(7).  Choose $n_2$ positive and $n_2$ negative comment samples $S_2$ most confidently predicted by $C_C$
(8).  Choose corresponding message samples $M_2$ (the message samples from the same users in $S_2$)
(9).  $L_{Mes} = L_{Mes} + M_1 + M_2$; $L_{Com} = L_{Com} + S_1 + S_2$
(10). $U_{Mes} = U_{Mes} - M_1 - M_2$; $U_{Com} = U_{Com} - S_1 - S_2$
```

Figure. 2. Co-training algorithm for gender classification

In semi-supervised learning, we employ the co-training algorithm (Blum and Mitchell, 1998) to leverage the two views from two kinds of text: one contains all the messages written by the user himself and the other contains the comments written by other users. Figure 2 illustrates the co-training algorithm for gender classification by using the *message* text and the *comment* text as two different views. In this figure, two different classifiers trained with the *message* text and the *comment* text are denoted as the *message* classifier C_M and the *comment* classifier C_{C}, respectively.

4 Experimentation

Data Setting

The data is collected from Sina Weibo[1], the most famous Micro-blog platform in China. In a total, 12,651 users are crawled. Since many of them are not active, we remove those users who have posted less than 10 messages or received less than 10 comments since their registration or who have less than 50 followers or 50 followings. Furthermore, verified organizational users are removed. As a result, we get 7658 users, from which we randomly select 2000 *male* and 2000 *female* users in our experimentation. In supervised learning, we use different sizes of labeled data as training data and 400 samples as test data. In semi-supervised learning, we select 400 samples as the initial labeled data, 400 samples as test data and the remaining as unlabeled data.

Features

The basic features are bag-of-words features. Apart from these basic features, we also consider F-measure and POS sequence pattern features, as described above. These features yield the state-of-the-art performance in gender classification (Mukherjee and Liu, 2010). To get the word and POS features in Chinese text, we use ICTCLAS (http://www.ictclas.org/ictclas_download.aspx) to perform word segmentation and POS tagging on the Chinese text.

Classification algorithm

The maximum entropy (ME) classifier implemented with the public tool, Mallet Toolkits[2].

Experimental Results

Table 1. Accuracy of supervised learning with different sizes of training data

	1600	2000	2400	2800	3200
Using *Message* Text	0.788	0.800	0.800	0.805	0.813
Using *Comment* Text	0.760	0.768	0.773	0.780	0.780
Using both *Message* and *Comment* Texts	**0.825**	**0.830**	**0.840**	**0.845**	**0.855**

[1] http://weibo.com/

[2] http://mallet.cs.umass.edu/

Table 1 shows the accuracy of supervised learning with *message*, *comment* and both texts when different sizes of training data are used. From the table, we can see that although merely using *comment* text performs a bit worse than using *message* text, the performance is greatly improved when both texts are employed.

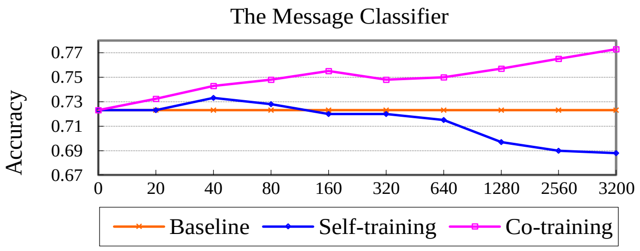


Figure. 3. Performance of the *message* classifier in self-training and co-training when different sizes of unlabeled data are added

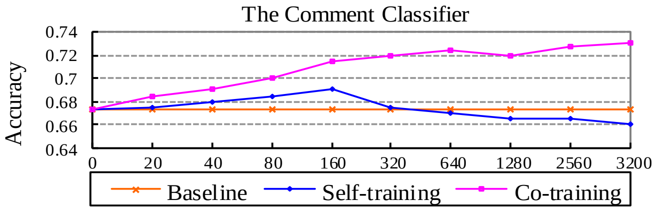


Figure. 4. Performance of the *comment* classifier in self-training and co-training when different sizes of unlabeled data are added

Figure 3 and Figure 4 show the performance of the *message* and *comment* classifiers when different numbers of unlabeled samples are added into the classifier with both self-training and co-training. Here, in each interaction, we pick 5 most confident *male* samples and *female* samples, i.e., $n_1 = n_2 = 5$. The baseline approach refers to the supervised classifier trained with only the initial labeled data (no unlabeled data is used) and the self-training approach refers to the single-view semi-supervised learning approach as applied in Burger et al. (2011).

From these figures, we can see that the performance of both the *message* and *comment* classifiers improves gradually when more and more unlabeled data are added. Finally, the *message* classifier obtains 5% improvements and the *comment* classifier obtains 5.7% improvements. From these figures, we can also see that self-training fails to effectively improve the performance and even performs worse than the baseline classifier when more and more unlabeled data are incorporated.

5 Conclusion

In this paper, we perform gender classification by exploiting the interactive knowledge from the *comment* text as a complement to the non-interactive knowledge from the *message* text. On the basis, a co-training approach is proposed to integrate both kinds of knowledge and unlabeled data. Evaluation demonstrates the complementarity of the interactive knowledge and the non-interactive knowledge in gender classification. It also demonstrates that the proposed co-training approach can well incorporate unlabeled data and considerably improve the performance, while the self-training approach fails.

Acknowledgments

This research work has been partially supported by three NSFC grants, No. 61273320, No.61375073, No.61331011, and Collaborative Innovation Center of Novel Software Technology and Industrialization.

Reference

1. Blum A. and T. Mitchell. Combing Labeled and Unlabeled Data with Co-Training. 1998. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pp. 92-100.
2. Corney M., O. Vel, A. Anderson and G. Mohay. 2002. Gender-Preferential Text Mining of E-mail Discourse. In *Proceedings of the 18th Annual Computer Security Applications Conference*, pp. 282-289.
3. Ciot M., M. Sonderegger and D. Ruths. 2013. Gender Inference of Twitter Users in Non-English Contexts. In *Proceedings of EMNLP-13*, pp. 1136–1145.
4. Gianfortoni P., D. Adamson and C. Rosé 2011. Modeling of Stylistic Variation in Social Media with Stretchy Patterns. In *Proceedings of EMNLP-11*, pp. 49–59.
5. Ikeda D., H. Takamura and M. Okumura. 2008. Semi-Supervised Learning for Blog Classification. In *Proceedings of AAAI-08*, pp.1156-1161.
6. Filippova K. 2012. User Demographics and Language in an Implicit Social Network. In *Proceedings of EMNLP-12*, pp. 1478-1488.
7. Heylighen F., and Dewaele, J. 2002. Variation in the contextuality of language: an empirical measure. *In Proceedings of Foundations of Science*, 7, 293–340.
8. Liu N., Y. He, Q. Chen, M. Peng and Y. Tian. 2013. A New Method for Micro-blog Platform Users Classification Based on Infinitesimal-time. *Journal of Information & Computantional Science*. 10:9 (2013) 2569–2579.
9. Mukherjee A. and B. Liu. 2010. Improving Gender Classification of Blog Authors. In *Proceedings of EMNLP-11*, pp. 207–217.
10. Nowson S. and J. Oberlander. 2006. The Identity of Bloggers: Openness and Gender in Personal Weblogs. In *Proceedings of AAAI-06*, pp. 163-167.
11. Peersman C., W. Daelemans and L. Vaerenbergh. 2010. Predicting Age and Gender in Online Social Networks. In *SMUC '10 Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pp. 37-44
12. Rao D., D. Yarowsky, A. Shreevats and M. Gupta. 2010. Classifying Latent User Attributes in Twitter. In *Proceeding SMUC '10 Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pp. 37-44.
13. Volkova S., T. Wilson and D. Yarowsky. 2013. Exploring Demographic Language Variations to Improve Multilingual Sentiment Analysis in Social Media. In *Proceedings of EMNLP-13*, pp. 1815–1827.