Semi-supervised Gender Classification with Joint Textual and Social Modeling

Shoushan Li, Bin Dai, Zhengxian Gong, Guodong Zhou*
Natural Language Processing Lab
School of Computer Science and Technology, Soochow University, China
lishoushan@suda.edu.cn, bdai@stu.suda.edu.cn
zhxgong@suda.edu.cn, gdzhou@suda.edu.cn

Abstract

In gender classification, labeled data is often limited while unlabeled data is ample. This motivates semi-supervised learning for gender classification to improve the performance by exploring the knowledge in both labeled and unlabeled data. In this paper, we propose a semi-supervised approach to gender classification by leveraging textual features and a specific kind of indirect links among the users which we call "s*ame-interest*" links. Specifically, we propose a factor graph, namely Textual and Social Factor Graph (TSFG), to model both the textual and the "*same-interest*" link information. Empirical studies demonstrate the effectiveness of the proposed approach to semi-supervised gender classification.

1 Introduction

Gender classification is a fundamental task with regard to infer user's gender from the user-generated data. Recently, this task is getting increasingly more attention in some prevailing research fields, such as social network analysis and natural language processing. Applications developed from gender classification have enormous commercial value in personalization, marketing and judicial investigation (Mukherjee and Liu, 2010; Burger *et al.*, 2001; Volkova *et al.*, 2013).

In social media, conventional methods handle gender classification as a supervised learning problem over the past decade (Corney *et al.*, 2002; Ciot *et al.*, 2013). In supervised learning approaches, both user-generated textual and user social link features are verified to be effective for gender classification. For instance, in Figure 1, it is easy to infer *User* ***c*** to be a *female* through analyzing her saying "*I'm gonna be a* ***mom****!!* " Meanwhile, it is also possible to infer *User* ***c*** is more likely to be a *female* through analyzing her social link since she follows a cosmetic-selling *User* "***Dior***".

Although supervised methods have achieved remarkable success for gender classification, their good performances always depend on a large amount of labeled data, which often need expensive labor costs and long production time. How to learn a classification model with low dependence on the large-scale labeled data becomes an important and challenging problem in gender classification.

In this paper, we propose a semi-supervised learning approach to alleviate the above problem in supervised gender classification. Instead of using a large scale of labeled data, we exploit a small scale of labeled data and large amount of unlabeled data to train the model. Our semi-supervised approach employs both user-generated textual knowledge and user social link information. The basic motivation of our approach lies in the observation that social link information might be helpful to infer user gender. Specifically, we focus on the "*following*" link and think that two users who follow the same particular user could have the same gender. For instance, in Figure 1, *User* ***b***, *User* ***c*** and *User* ***d*** follow the same user named ***Dior*** and they are thought to be indirectly linked. Once *User* ***b*** and *User* ***c*** are correctly classified to be *female* with textual features, *User* ***d*** is more likely to be *female* since she is indirectly linked to *User* ***b*** and *User* ***c***.

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/

* Corresponding author

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2092–2100, Osaka, Japan, December 11-17 2016.

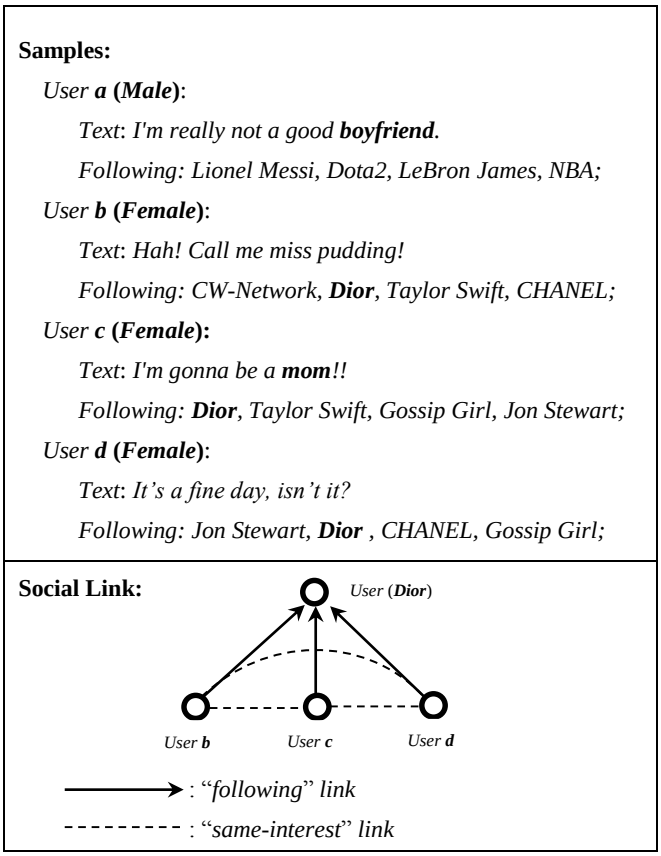

Figure 1: An example of Text and concerns in social media

Specifically, we propose a factor graph, namely Textual and Social Factor Graph (TSFG), to model both the textual and user social link information. Here, a social link between two users happens when the two users follow the same user. For instance, in Figure 1, *User **b*** and *User **c*** both follow the user named ***Dior***. These two users are thought to be linked with an indirect link, called "*same-interest*" link. In our TSFG approach, both the textual features and social links are modeled as various factor functions and the learning task aims to maximize the joint probability of all these factor functions. Empirical evaluation demonstrates the effectiveness of our TSFG approach to capture the inherent user social link. To the best of our knowledge, this work is the first attempt to incorporate both the textual and social information in semi-supervised gender classification.

The remainder of this paper is organized as follows. Section 2 overviews related work on gender classification. Section 3 introduces data collection and analysis. Section 4 describes our TSFG approach to gender classification. Section 5 presents the experimental results. Finally, Section 6 gives the conclusion and future work.

2 Related Work

In the last decade, gender classification has been studied in two main aspects: supervised learning and semi-supervised learning.

As for supervised learning, gender classification has been extensively studied in several textual styles, such as Blog (Nowson and Oberlander, 2006; Peersman *et al.*, 2011; Gianfortoni *et al.*, 2011), E-mail (Mohanmad *et al.*, 2011), YouTube (Filippova, 2012) and Micro-blog (Rao *et al.*, 2010; Liu et al., 2013). These studies mainly focus on employing various kinds of textual features such as character, word, POS features and their *n*-gram features to train the classifier. More recently, some studies focus on some specific application scenarios on supervised gender classification, such as multi-lingual gender classification (Ciot *et al.*, 2013; Alowibdi *et al.*, 2013) and interactive gender classification (Li *et al.*, 2015).

As for semi-supervised learning, gender classification has been studied with much less previous studies. 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. Wang *et al.* (2015) employ both non-interactive and interactive texts as two different views in their co-training approach for semi-supervised gender classification.

Unlike the studies above, our study focuses on both textual features and social links for semi-supervised gender classification.

3 Data Collection and Analysis

The data is collected from Sina Micro-blog[2], the most famous Micro-blogging platform in China. In this platform, local users publish short messages and are allowed to follow other users to listen to their messages. From the website, we crawl each user's homepage which contains the user information (e.g. Name, gender, and, verified type), messages and following users. The data collection process starts from some randomly selected users, and then iteratively gets the data of their followers and followings. We remove some unsuitable users that meet one of the following two conditions: (1) verified organizational users that are verified as organization; (2) the non-active users that have less than 50 followers or 50 followings.

In total, we obtain about 10000 user homepages, from which we randomly select a balanced data set containing 1000 male and 1000 female users. Let $Fo(u_i)$ denotes the set of u_i 's all "following" users; F_{male} denotes the set of all male users' "following" users; F_{female} denotes the set of all female users' "following" users. F_{male} and F_{female} can be calculated as following:

$$F_{male} = \bigcup_{u_i \in S_{male}} Fo(u_i) \tag{1}$$

$$F_{male} = \bigcup_{u_i \in S_{female}} Fo(u_i) \tag{2}$$

Where S_{male} and S_{female} denote the sets of male and female users respectively.

Table 1 shows the statistics about the numbers of "following" users of all male and female users. From this table, we can see that there are many users who are only followed by male users or female users. Specifically, in our data set, 143389 users are followed by only male users and 119504 users are followed by only female users. Thus, these gender-sensitive followings are good clues to infer each user's gender.

	#of "*following*" users
$\lvert F_{male} \rvert$	162116
$\lvert F_{female} \rvert$	138231
$\lvert F_{male} \cap F_{female} \rvert$	18727
$\lvert F_{male} \rvert - \lvert F_{male} \cap F_{female} \rvert$	143389
$\lvert F_{female} \rvert - \lvert F_{male} \cap F_{female} \rvert$	119504

Table 1: Statistics of the following users

4 Textual and Social Factor Graph Model

A factor graph consists of two layers of nodes, i.e., variable nodes and factor nodes, with links between them. The joint distribution over the whole set of variables can be factorized as a product of all factors.

[2] http://weibo.com/

4.1 Model Definition

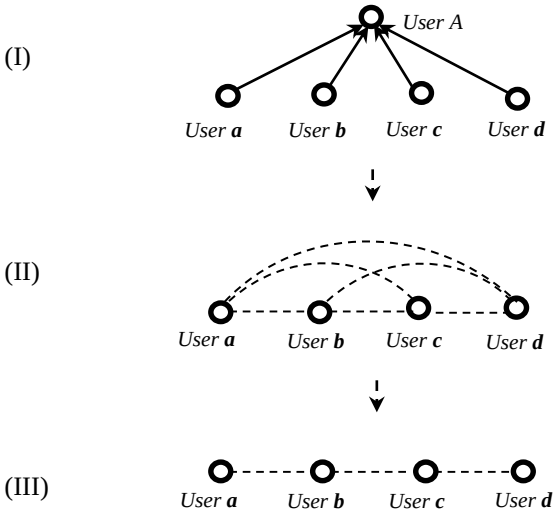

Figure 2: An example for illustrating the user links where

(I) shows the "following" links among all users;
(II) shows the "same-interest" links among the four users;
(III) shows the simplified four "same-interest" links among the four users.

Formally, let $G=(V,E)$ represent an instance network, where V denotes a set of the involved users in our data set. $E \subset V \times V$ is a set of relationships between users. Specifically, if a user u_i and a user u_j have the same following (i.e., the same-interest link), there is an edge $e_{ij}, e_{ij} \in E$, linking the two users u_i and u_j.

The "following" link: If a user u_i follows another user u_j, there is a "following" link between u_i and u_j. For instance, Figure 2(I) shows an example where four users, namely *User a*, *User b*, *User c*, and *User d*, are in our data set and each of them follows *User A*. Thus there are four "following" links among these five users.

The "same-interest" link: If a user u_i and a user u_j follows the same user, there is a "same-interest" link between u_i and u_j. For instance, Figure 2(II) shows six "same-interest" links among the four users, i.e., *User a*, *User b*, *User c*, and *User d*. The "same-interest" links derived from "following" links as showed in Figure 2(I).

Suppose that there are N users who have the same interest, the number of the same-interest links is C_N^2. However, when N is large, the number of the links is too large, which might make our factor graph model difficult to learn. Therefore, we simplify the link model by deleting $C_N^2-(N-1)$ links, only reserving a link line containing N-1 links, as shown in Figure 2(III).

We model the above network with a factor graph and our objective is to infer the gender categories of instances by learning the following joint distribution:

$$P(Y|G)=\prod_i \prod_k f(X_i,y_i) h_k(y_i,H(y_i)) \tag{3}$$

Where two kinds of factor functions are used.

1) Textual feature factor function: $f(X_i,y_i)$ denotes the traditional textual feature factor functions associated with each text representation of the user u_i, i.e., X_i. The textual feature factor function is instantiated as follows:

$$f(X_i,y_i)=\frac{1}{Z_1}\exp\left(\sum_j \alpha_j \Phi(x_{ij},y_i)\right) \tag{4}$$

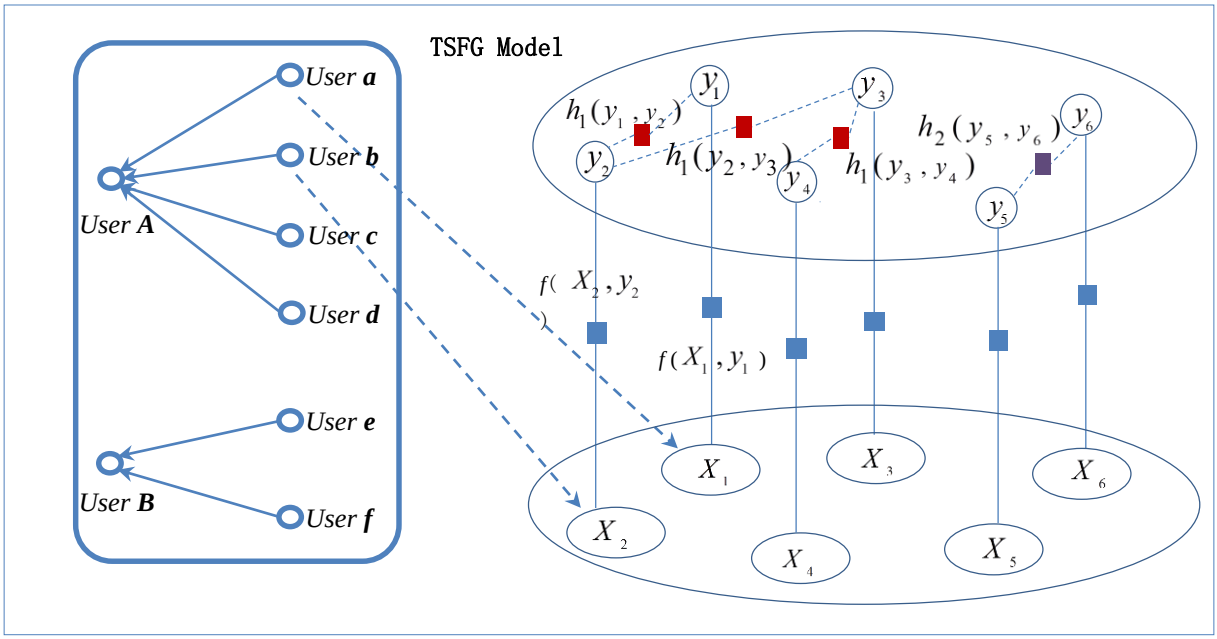

Figure 3: An example of TSFG where six instances are involved:

User ***a****, User* ***b****, User* ***c****, User* ***d****, User* ***e****, and User* ***f****.*

Note: each instance is represented as X_i . $f(\cdot)$ represents a factor function for modeling textual features. $h(\cdot)$ represents a factor function for modeling the "*same-interest*" link between two instances.

Where $\Phi(x_{ij}, y_i)$ is a feature function and x_{ij} represents a textual feature, i.e., a word feature in this study.

2) Social link factor function: $h_k(y_i, H(y_i))$ denotes the "same-interest" relationship among the users who follow the same user $u_k, u_k \in F_{male} \bigcup F_{female}$. $H(y_i)$ is the label set of the users linked to y_i. The social link factor function is instantiated as follows:

$$h_k(y_i, H(y_i)) = \frac{1}{Z_2} \exp\left\{ \sum_{y_i^l \in H(y_i)} \beta_{ikl} \left(y_i - y_i^l \right)^2 \right\} \tag{5}$$

Where β_{ikl} is the weight of the function, representing the degree if influence of the two instances y_i and y_i^l.

Figure 3 gives an example of our textual and social factor graph (TSFG) where six users, i.e., *User* ***a****, User* ***b****, User* ***c****, User* ***d****, User* ***e****,* and *User* ***f****,* are involved.

4.2 Model Learning

Learning the DFG model is to estimate the best parameter configuration $\theta = (\{\alpha\}, \{\beta\})$ to maximize the log-likelihood objective function $L(\theta) = \log P_\theta(Y|G)$, i.e.,

$$\theta^* = \arg\max L(\theta) \tag{6}$$

In this study, we employ the gradient decent method to optimize the objective function. For example, we can write the gradient of each α_j with regard to the objective function:

$$\frac{\partial L(\theta)}{\partial \alpha_j} = E\left[\Phi(x_{ij}, y_i)\right] - E_{P_{\alpha_j}(Y|G)}\left[\Phi(x_{ij}, y_i)\right] \tag{7}$$

Where $E\left[\Phi(x_{ij}, y_i)\right]$ is the expectation of feature function $\Phi(x_{ij}, y_i)$ given the data distribution. $E_{P_{\alpha_j}(Y|G)}\left[\Phi(x_{ij}, y_i)\right]$ is the expectation of feature function $\Phi(x_{ij}, y_i)$ under the distribution $P_{\alpha_j}(Y|G)$

given by the estimated model. Figure 4 illustrates the detailed algorithm for learning the parameter α. Note that LBP denotes the Loopy Belief Propagation (LBP) algorithm which is applied to approximately infer the marginal distribution in a factor graph (Frey and MacKay, 1998). A similar gradient can be derived for the other parameters.

Input: Learning rate η

Output: Estimated parameters α

Initialize $\alpha \leftarrow 0$

Repeat

1) Calculate $E\left[\Phi\left(x_{ij}, y_i\right)\right]$ using LBP
2) Calculate $E_{P_{\alpha_{ij}}(Y|G)}\left[\Phi\left(x_{ij}, y_i\right)\right]$ using LBP
3) Calculate the gradient of α according to Eq. (7)
4) Update parameter α with the learning rate η

$$\alpha_{\text{new}} = \alpha_{\text{old}} + \eta \frac{L(\alpha)}{\alpha}$$

Until *Convergence*

Figure 4: The learning algorithm for TSFG model

It is worth noting that we need to perform the LBP process twice for each iteration: One is to estimate the original distribution of unlabeled instances which are denoted as $y_i = ?$ and the other is to estimate the marginal distribution over all pairs. In this way, the algorithm essentially leverage both the labeled data and unlabeled data to optimize the complete network.

4.3 Model Prediction

With the learned parameter configuration θ, the prediction task is to find a Y^{T*} which optimizes the objective function, i.e.,

$$Y^{T*} = \arg\max P\left(Y^T \middle| Y^{L+U}, G, \theta\right) \tag{8}$$

Where Y^{T*} are the labels of the instances in the testing data and Y^{L+U} are the labels (or estimated labels) of the instances in the labeled and unlabeled data.

Again, we utilize LBP to calculate the marginal probability of each instance $P\left(y_i \middle| Y^{L+U}, G, \theta\right)$ and predict the label with the largest marginal probability. For all instances in the test data, the prediction indicated above is performed iteratively until converge.

5 Experimentation

We have systematically evaluated our TSFG approach to semi-supervised gender classification.

5.1 Experimental Settings

Data Setting: The data set contains 2000 users, as described in Section 3. From this data set, we select 200 users as initial labeled data, 1400 users as unlabeled data, and the remaining 400 users as the test data.

Features: Three types of textual features, including bag-of-words, f-measure, and POS pattern features, are adopted in our experiments. These features yield the state-of-the-art performance in gender classification (Mukherjee and Liu, 2010). To get word and POS features, we use the toolkit ICTCLAS[3] to perform word segmentation and POS tagging on the Chinese text.

[3] http://www.ictclas.org/ictclas_download.aspx

Classification Algorithm: For supervised learning, various of classification algorithms are available. As suggested by Li *et al.* (2015), we apply maximum entropy (ME) for supervised gender classification. Specifically, the ME algorithm is implemented with the Mallet Toolkit[4]. For semi-supervised learning, we implement our TSFG approach, together with some baselines.

Evaluation Measurement: The performances are evaluated using the standard *precision*, *recall*, and *F-score* in each gender category. For overall evaluation, we use macro-average *F-score* over both gender categories, which is denoted as F_{macro}.

Significance test: *T*-test is used to evaluate the significance of the performance difference between two approaches (Yang and Liu, 1999).

5.2 Experimental Results

For thorough comparison, several gender classification approaches are implemented including:

- **Baseline(Textual)**: employing ME classifier and textual features with only initial labeled data (without any unlabeled data).
- **Baseline(Textual+Social)**: employing ME classifier and both textual and social features with only initial labeled data (without any unlabeled data). Social features are extracted by considering each user ID of the followers of a user as a word.
- **Self-training(Textual)**: employing ME classifier and textual features with both labeled data and unlabeled data using self-training.
- **Self-training(Textual+Social)**: employing ME classifier and both textual and social features with both labeled data and unlabeled data using self-training.
- **Co-training(Textual)**: employing ME classifier and textual features with both labeled data and unlabeled data using co-training. We implement the co-training algorithm by randomly splitting the feature space into two disjoint feature subspaces as two views (Nigam and Ghani, 2000).
- **Co-training(Textual+Social)**: employing ME classifier and both textual and social features with both labeled data and unlabeled data using co-training. We implement the co-training algorithm by randomly splitting the feature space into two disjoint feature subspaces as two views (Nigam and Ghani, 2000).
- **TSFG:** our approach as described in Section 4.

	Male			*Female*			*Total*
Approach	*Precision*	*Recall*	*F-score*	*Precision*	*Recall*	*F-score*	F_{macro}
Baseline(Textual)	0.760	0.650	0.700	0.694	0.795	0.741	0.721
Baseline(Textual+Social)	0.800	0.700	0.747	0.733	0.825	0.776	0.762
Self-Training(Textual)	0.714	0.710	0.711	0.711	0.715	0.713	0.712
Self-Training(Textual+Social)	0.754	0.735	0.744	0.741	0.760	0.751	0.747
Co-Training(Textual)	0.725	0.700	0.712	0.710	0.735	0.722	0.717
Co-Training(Textual+Social)	0.784	**0.745**	0.764	0.757	0.795	0.776	0.770
TSFG	**0.961**	0.735	**0.833**	**0.785**	**0.970**	**0.868**	**0.851**

Table 2: Performance comparison of different approaches to semi-supervised gender classification

Table 2 shows the performance comparison of different approaches to gender classification. From this table, we can see that:

(1) Social BOW features are helpful in both supervised and semi-supervised learning approaches.
(2) Self-training fails to exploit unlabeled data to improve the performance and it performs even worse than the baseline approaches.

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

(3) Co-training is effective for semi-supervised gender classification when both textual and social features are employed. This result indicates that the use of social features in semi-supervised gender classification in co-training is beneficial, although the improvement is rather limited, about 1%.

(4) Our approach TSFG performs best among all semi-supervised learning approaches. Moreover, the improvement over the two baselines is remarkable, 13% higher than Baseline(Textual) and 8.9% higher than Baseline(Textual+Social). Significance test shows that our approach significantly outperforms co-training (p-value<0.01)

Figure 5 shows the performances of our approach and the two baseline approaches when varying the sizes of the initial labeled data. From this figure, we can see that social features are always helpful for gender classification and Baseline(Textual+Social) consistently outperforms Baseline(Textual). Our approach fails to take effect when the size of the initial labeled data is too small (10 labeled instances in each category). When the size of the initial data is larger than 20 instances in each category, our TSFG approaches consistently performs much better than the two baseline approaches. Significance test shows that our TSFG approach significantly outperforms both Baseline(Textual) and Baseline(Textual+Social) when the size of the initial labeled instance is larger than 20 in each gender category (p-value<0.01).

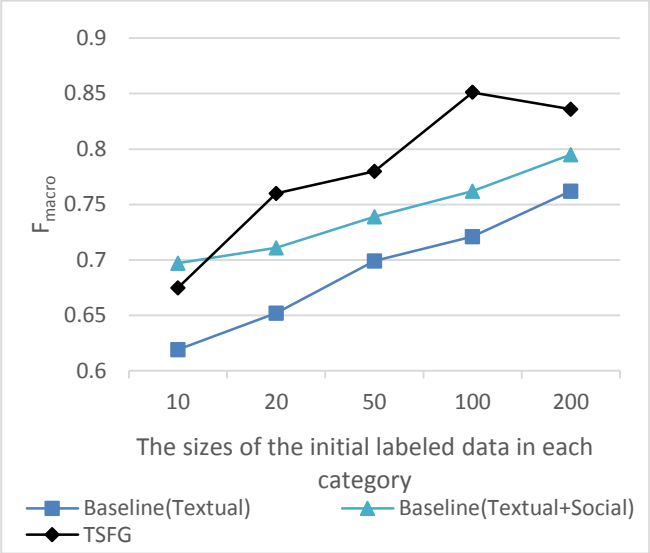

Figure 5: The performances of our approach and the two baseline approaches when varying the sizes of the initial labeled data.

6 Conclusion

In this paper, we propose a novel approach to semi-supervised gender classification in social media. In our approach, we first define a social link named "*same-interest*" link which models an indirect link between two users who follow the same user. Then, we propose a factor graph-based approach, namely Textual and Social Factor Graph (TSFG), where both the textual features and "*same-interest*" social links are modeled as various factor functions. Finally, we employ the graph to leverage both the labeled data and unlabeled data to optimize the complete network. Empirical studies show that our TSFG approach successfully exploits unlabeled data to improve the performance, remarkably outperforming other semi-supervised learning approaches.

In our future work, we would like to improve our semi-supervised learning approach by leveraging some other kinds of link information. Furthermore, we will apply our TSFG approach to some other NLP tasks where both textual and social features are available, such as user age prediction (Rosenthal and McKeown, 2011) and user occupation classification (Preotiuc-Pietro *et al.*, 2015).

Acknowledgements

This research work has been partially supported by five NSFC grants, No.61273320, No.61375073, No.61331011, No.61305088 and No.61373096.

Reference

Burger J. and J. Henderson and G. Kim and G. Zarrella. 2011. Discriminating Gender on Twitter. In *Proceedings of EMNLP-11*, pp. 1301–1309.

Corney M., O. Vel, A. Anderson and G. Mohay. 2002. Gender-Preferential Text Mining of E-mail Discourse. In *Proceedings of ACSAC-02*, pp. 282-289.

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.

Filippova K. 2012. User Demographics and Language in an Implicit Social Network. In *Proceedings of EMNLP-12*, pp. 1478–1488.

Frey B. and D. MacKay. 1998. A Revolution: Belief Propagation in Graphs with Cycles. In *Proceedings of NIPS-98*, pp.479–485.

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.

Ikeda D., H. Takamura and M. Okumura. 2008. Semi-Supervised Learning for Blog Classification. In *Proceedings of AAAI-08*, pp.1156-1161.

Li S., J. Wang, G. Zhou and H. Shi. 2015 Interactive Gender Inference with Linear Programming. In *Proceedings of IJCAI-15*, pp. 2341-2347.

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.

Mukherjee A. and B. Liu. 2010. Improving Gender Classification of Blog Authors. In *Proceedings of EMNLP-10*, pp. 207-217.

Mohammad S. and T. Yang. 2011. Tracking Sentiment in Mail: How Genders Differ on Emotional Axes. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis(2011)*, pp.70-79.

Nigam, K. and R. Ghani. 2000. Analyzing the Effectiveness and Applicability of Co-training. In *Proceedings of CIKM-2000*, 86-93.

Nowson S. and J. Oberlander. 2006. The Identity of Bloggers: Openness and Gender in Personal Weblogs. In *Proceeding of AAAI-06*, pp. 163-167.

Peersman C., W. Daelemans, L. Van Vaerenbergh. 2011. Predicting Age and Gender in Online Social Networks. In *Proceedings of SMUC-11*, pp. 37-44.

Preotiuc-Pietro D., V. Lampos and N. Aletras. 2015. An Analysis of the User Occupational Class through Twitter Content. In *Proceedings of ACL-15*, pp. 1754-1764.

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.

Rosenthal S. and K. McKeown. 2011. Age Prediction in Blogs: A Study of Style, Content, and Online Behavior in Pre- and Post-Social Media Generations. In *Proceedings of ACL-11*, pp.763-772.

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.

Wang J., Y. Xue, S. Li and G. Zhou. 2015 Leveraging Interactive Knowledge and Unlabeled Data in Gender Classification with Co-training. In *Proceedings of DASFAA-15*, pp. 246-251.

Yan X. and L. Yan. 2006. Gender inference of Weblog Authors. In *Proceedings of AAAI-06*, pp.228-230.

Yang Y. and X. Liu. 1999. A Re-Examination of Text Categorization Methods. In *Proceedings of SIGIR-99*, pp. 42-49.