Twitter Homophily: Network Based Prediction of User's Occupation

Jiaqi Pan*
University of Electronic Science and Technology of China
jiaqi.pan1019@gmail.com

Rishabh Bhardwaj*
National University of Singapore
rishabhbhardwaj15@gmail.com

Wei Lu
Singapore University of Technology and Design
luwei@sutd.edu.sg

Hai Leong Chieu
DSO National Laboratories
chaileon@dso.org.sg

Xinghao Pan
DSO National Laboratories
pxinghao@dso.org.sg

Ni Yi Puay
DSO National Laboratories
pniyi@dso.org.sg

Abstract

In this paper, we investigate the importance of social network information compared to content information in the prediction of a Twitter user's occupational class. We show that the content information of a user's tweets, the profile descriptions of a user's follower/following community, and the user's social network provide useful information for classifying a user's occupational group. In our study, we extend an existing dataset for this problem, and we achieve significantly better performance by using social network homophily that has not been fully exploited in previous work. In our analysis, we found that by using the graph convolutional network to exploit social homophily, we can achieve competitive performance on this dataset with just a small fraction of the training data.

1 Introduction

Twitter (http://twitter.com) is a microblogging service launched in 2006, where, a user can publish messages with up to 280 characters, called "tweets". Unlike many other social networking platforms, such as Facebook and LinkedIn, Twitter does not provide structured fields for users to fill in personal information. However, a user can write a 160-character-long small public summary about itself called a "Bio". Besides linguistic information from tweets and Bios, online social media is a rich source of network information. People's personal networks are homogeneous, i.e., friends share more attributes such as race, ethnicity, religion, and occupation–known as the homophily principle (McPherson et al., 2001). Such network information has been utilized in friend recommendation (Guy et al., 2010), community detection (Yang and Leskovec, 2013), etc. Figure 1 shows two users connected on Twitter. By looking at their Bio and tweets, it can be inferred that these users share the same occupational interest.

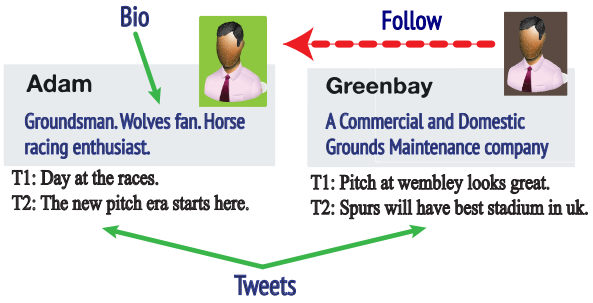


Figure 1: User and Network information on Twitter Microblog.

Profiling users can enhance service quality and improve product recommendation, and hence is a widely studied problem. User occupational class prediction is an important component of user profiling and a sub-task of user demographic feature prediction. Existing approaches to predicting Twitter users' demographic attributes explore, select, and combine various features generated from text and network to achieve the best predictive performances in respective classification tasks (Han et al., 2013; Miller et al., 2012; Preoţiuc-Pietro et al., 2015; Huang et al., 2015; Aletras and Chamberlain, 2018). The three categories of features are: account level features, tweet text features, and network based features. Past research have shown the distinctive usage of language across gender, age, location, etc. in tweets (Sloan et al., 2015; Cheng et al., 2010; Burger et al., 2011; Rao et al., 2010), which makes content based prediction effective.

As for user occupational class prediction, Preoţiuc-Pietro et al. (2015) built a dataset where

*Equal Contribution; work performed while both authors were visiting Singapore University of Technology and Design (SUTD).

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2633–2638
Florence, Italy, July 28 - August 2, 2019.

users are assigned to hierarchical job categories. They used word cluster distribution features of content information to predict a user's occupational group. Aletras and Chamberlain (2018) constructed a user's followings connections to learn the user embedding as a feature input to the classification models. Considering the regional disparities of economic development stages, the major job categories may vary significantly across regions. Sloan et al. (2015) summarized occupation distribution of Twitter users in the UK by looking into their profiles.

In this paper, we analyze the usefulness of a user's network information over the user's tweets for predicting its occupational group. We extend the existing dataset for occupation classification (Preoţiuc-Pietro et al. (2015)) by introducing the network information about a user, i.e. follower/following IDs together with their Bio descriptions, and we construct a user-centric network to extract useful community and text based features. The acquired features from the network are then exploited using a graph neural network. The obtained results show the importance of a network information over tweet information from a user for such a task.

2 Graph Convolutional Network

A Graph Convolutional Network (GCN) (Kipf and Welling, 2017) defines a graph-based neural network model $f(X, A)$ with layer-wise propagation rules:

$$\hat{A} = \tilde{D}^{-1/2}(A + \lambda I)\tilde{D}^{-1/2} \quad (1)$$

$$X^{(l+1)} = \sigma(\hat{A}X^{(l)}W^{(l)} + b^{(l)}) \quad (2)$$

where X is the feature matrix for all the nodes with $X^{(0)}$ being the initial feature input of size $d_{nodes} \times d_{features}$, A is the adjacency matrix of dimension $d_{nodes} \times d_{nodes}$, $\tilde{D}$ is the degree matrix of $A + \lambda I$, λ is a hyperparameter controlling the weight of a node against its neighbourhood, and $W^{(l)}$ and $b^{(l)}$ are trainable weights and bias for the l-th layer, respectively. In each layer of GCN, a node aggregates its direct neighbours' features according to $\hat{A}$ and linearly transforms the representation using W and b. A nonlinear activation function σ (e.g., ReLu) is then applied. The number of layers of GCN decides the number of hops away that the neighbours' features will be smoothed over for each node.

Gr	SOC	Users
1	Managers, Directors, Senior Officials	461
2	Professional Occ.	1,611
3	Associate Profess., Technical Occ.	926
4	Administrative Secretarial Occ.	162
5	Skilled Trades Occ.	768
6	Caring, Leisure, Other Service Occ.	259
7	Sales and Customer Service Occ.	58
8	Process, Plant, Machine Operatives	188
9	Elementary Occ.	124

Table 1: The table shows the major groups (left column) and categorized jobs with different sub-major groups (middle column) by SOC. The right-most column shows the number of main users in the data.

3 Experimental Setup

3.1 Data

We base our work on a publicly available Twitter dataset that maps 5,191 users to 9 major occupational classes (Preoţiuc-Pietro et al., 2015). The dataset contains user IDs (we call these users the *main* users henceforth) and the bag-of-words from tweets. The hierarchical structure of occupational classes in the data was defined based on the Standard Occupation Classification (SOC) from the UK[1].

To explore the role of network information in occupational class prediction, we extend the above dataset by crawling follower/following IDs (henceforth referred to as follow IDs) for each main ID (IDs corresponding to main users). For the crawled follow IDs, we further crawl their Bio descriptions. We refer to the extended dataset as ED. ED contains 4,557 main users with both followers and followings information. The remaining Twitter accounts could not be scrapped because of various reasons such as account suspension and protected tweets.

Table 1 shows the occupational class distribution of the main users in the ED. In all our work, we discard the Bio information of the main users as these were used to annotate this dataset. We tokenize the Bio text of the follow IDs using the Glove Twitter pre-processing guidelines[2]. As for social network construction, we consider each follower/following relationship as an undirected edge. Based on the reasoning that the social network information is passed between main IDs

[1] http://www.ons.gov.uk/

[2] https://nlp.stanford.edu/projects/glove/preprocess-twitter.rb

mainly through some common follow IDs, the follow IDs that only connect to very few main IDs will have minimum functionality in information flow.

Thus, we decide to filter the graph by keeping the follow IDs with more than 10 connections to the main IDs. All connections between main IDs are retained. The filtering step results in 29 main IDs losing all their connections. For all such isolated main IDs, we retrieve all its follow IDs having at least one other main ID connection. After all these operations, we are able to construct an un-weighted graph in which all the main IDs are connected. The filtered graph contains 34,630 unique users (including 4,557 main IDs) and 586,303 edges. Although the main users are not collected to be connected to each other – only 2,550 main IDs have at least one direct connection to another main ID, we find that they often share common follow IDs which allows us to retrieve their social representations.

To compare with previous works, we also construct a partial network dataset that contains only following IDs of all the 4,557 main IDs. We refer to this partial dataset as PD. PD adheres to the same network construction methodology as ED.

We divide the dataset into training, development, and test sets using stratified split with the splitting ratio of 80%, 10%, and 10%. All the experimental results are reported on the same test set. The split information and the processed dataset ED can be found together with code on github: https://github.com/jqnap/Twitter-Occupation-Prediction.

3.2 Features and Models

Node Embeddings: To encode user-user social relationship of main IDs with the follow network, we learn latent representations of all IDs (node embedding) which can be easily exploited for the prediction task. The embeddings are learned by forming node sequences using Deep Walk (Perozzi et al., 2014).

Based on the network processing strategy used in Aletras and Chamberlain (2018), we construct unweighted bipartite graphs using our filtered network. The two sides of a bipartite graph are follow IDs and main IDs respectively. Note that the main ID-main ID connections will break the bipartiteness. To resolve this, we duplicate the main ID nodes to the follow IDs' side and then link connections within main IDs. We construct for both ED and PD, and obtain a full graph (fG) and a partial graph (pG) respectively.

Next, we performed 10 random walks starting from each main ID, alternating between main ID and followers/followings with a walk length of 80. For each node, the walk sequence is used to generate embeddings using a similar approach to word2vec (Mikolov et al., 2013). We use the same hyper-parameters as in Aletras and Chamberlain (2018).

Text Features: To have a valid comparison with existing approaches, we construct two sets of text features: (1) bag-of-clusters (Preoţiuc-Pietro et al., 2015): we assign each word that appears in each main ID's concatenated tweets document to its corresponding word cluster, where the word clusters are obtained by applying spectral clustering (Ng et al., 2002; Shi and Malik, 2000) to word embeddings. Next, we calculate the cluster assigning frequencies for each main ID. (2) bag-of-words (BOW): since the initial dataset used the Bio information of the main users to annotate their occupations, we remove all the Bio information of main users. We kept only the most frequent 5,000 words from the Bio (of other users) and another 5,000 words from tweets text as the dictionary of separate BOW vectors to the model. We feed the obtained text features and node embedding features to both the Logistic Regression (LR) classifier and the Support Vector Machine (SVM) classifier [3]. Both classifiers are trained following the one-vs-all approach for the 9-way classification task. ℓ_2 regularization is used for LR, whose coefficient is tuned based on the development set. We use the RBF kernel for SVM, normalize the features before feeding them to SVM as inputs, and tune the regularization coefficient C using the development set.

GCN: In the case of GCN (as shown in Figure 2), we use its transductive semi-supervised setting. The inputs are the adjacency matrix of all the network IDs and a feature matrix of the Bio's bag-of-words. Specifically, we keep the input feature vectors corresponding to the main IDs as null (all zeros), since their Bios were discarded. We experiment GCN with 2, 3 and 4 convolutional layers. The 3-layer GCN slightly outperformed the

[3] We use the scikit-learn implementations of LR and SVM classifiers: https://scikit-learn.org/

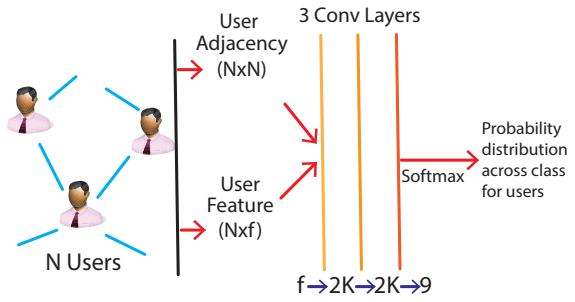


Figure 2: GCN architecture for occupational class prediction. 2K is the best performing hidden size.

	LR	SVM
Word Clusters (200)*	49.8	52.6
Clusters+DeepWalk-pG (200 + 32)*	51.3	55.0
Main ID tweets BOW (5, 000)	53.7	54.6
F-Bio (5, 000)	56.6	56.3
DeepWalk-fG (32)	51.5	55.3
DeepWalk-fG + F-Bio (32 + 5, 000)	56.6	57.5
	GCN	
Bio BOW (34, 630 × 5, 000)	59.9	
Adjacency (34, 630 × 34, 630)	**61.0**	

Table 2: Performance in terms of accuracy percentage comparison of logistic regression (LR), support vector machines (SVM), and graph convolutional networks (GCN). The first two rows (marked with ∗) are existing approaches from Preoţiuc-Pietro et al. (2015) and Aletras and Chamberlain (2018). The number in brackets are the dimension of the feature space. pG and fG refer to partial graph and full graph respectively. We use F-Bio to denote "Follower Bio BOW".

2-layer GCN and is on-par with the 4-layer GCN. We also test another setting where we do not use the Bio information: we keep the feature as a matrix of one-hot encoded vectors corresponding to all 34,630 IDs. For all the experiments, we set λ to 1 in Equation 1.

4 Results and Discussion

4.1 Text Features and Node Embeddings

As shown in Table 2, we compare our results using network information with existing methods: bag-of-clusters (Preoţiuc-Pietro et al., 2015) and Deepwalk on the followings graph concatenated with bag-of-clusters (Aletras and Chamberlain, 2018).

We first conduct experiments on our collected ED dataset with 4,557 main users using existing methods. The better accuracy among existing methods is given by the concatenated bag-of-clusters and Deepwalk embeddings: 55.0%.

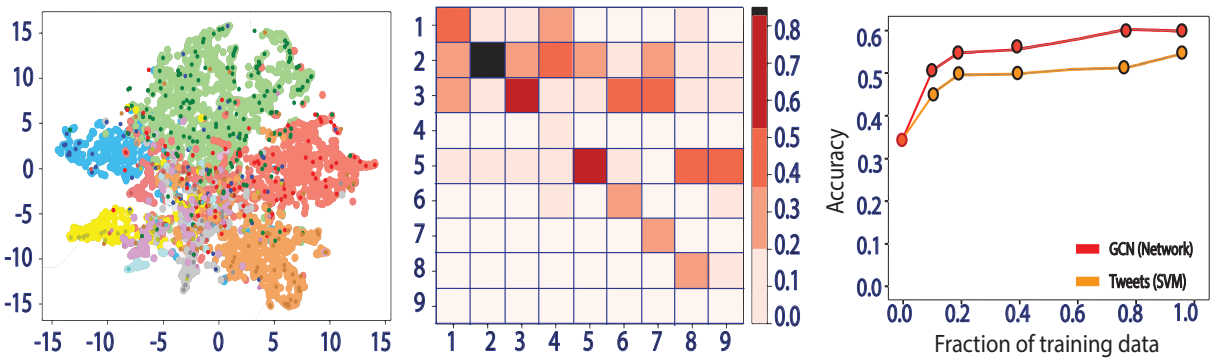
Next, we investigate the performance of bag-of-words features from main ID tweets and follow Bios using logistic regression (LR) and support vector machines (SVM). From the experiments on tweets, we find that using the bag-of-words features achieve comparable performance to using the bag-of-clusters features. Thus we opt for the bag-of-words representation in subsequent experiments. The optimized model using Bio text features outperforms using tweet content. It can be inferred that the Bio descriptions of follow accounts provide more useful information compared to tweets. The reason could be the higher noise in tweets, while people are comparatively more careful while writing their Bios.

The next set of results uses follow network features. Based on Aletras and Chamberlain (2018), we perform deep walk with 32-dim learned node representations, and used it as input to LR and SVM. We achieve higher accuracy (55.3%) as compared to tweets BOW (54.6%). However, the model is less effective than using follow Bio BOW. Combining both node representations and follow Bio BOW features further boosts the accuracy to 57.5%.

4.2 GCN

To analyze the importance of Bios in conjunction with social network information, we exploit graph convolutional networks. With an accuracy of 59.9%, the model exceedingly outperforms existing approaches on tweets and partial network information. Our best result 61.0% accuracy is achieved by using GCN with one-hot encoding for nodes, which is significantly higher than existing methods. This shows that GCN is able to exploit the rich topological information of network to learn social representations for users. We postulate that the GCN with Bio did not do better than just a one-hot encoding for nodes because the main users do not have Bios: so all the labeled nodes in the GCN have no Bios, which makes learning difficult.

We visualize the GCN final layer representations of training set (big ovals) and test set (dark colored dots) in Figure 3a. It can be observed that many test data samples are mapped to the correct group of occupation, showing the capability of



(a) 2D t-SNE Visualization (b) Confusion Matrix (c) Accuracy vs Training data

Figure 3: (a) A 2D t-SNE plot of final layer user represntations learned using GCN; (b) Confusion matrix of prediction made by GCN (rows and columns represent actual and predicted group, respectively); (c) Model performance vs fraction of training data used.

GCN utilizing Twitter network information for the prediction task. To analyze wrongly mapped test samples, we observed confusion matrix as shown in Figure 3b. We see that group 4 is predicted as belonging to group 1 or 2. When we compare the jobs lying in groups 1, 2, and 4, we found that they contain similar types of sub-occupations, such as "financial account managers" and "finance officers", or "engineers" and "engineering technicians". The same phenomenon can be seen for group 9 and group 5.

Figure 3c compares the performance of two models, using tweet only features (LR–tweets) and follow network features (GCN–Bio), based on a fraction of training samples used for model learning. Even with 10% of the labeled training data, GCN with Bio-BOW features achieves comparable accuracy to existing models as well as models trained on tweet BOW with all the training set. This shows the significance of a user's network information.

We analyze the predictions on test samples made by GCN with Bio feature input and GCN with the one-hot encoded input. We find that 11% of the test set's main IDs are correctly classified by only one of the two GCNs. This suggests that Bio features provide complementary information to the one-hot encoded input. In this work, the acquired network is dense. In cases when network is sparse, one-hot representation of an ID seems infeasible while BOW may generalize for the larger graph.

While occupational class prediction could be used to improve service quality, we note that the use of network information might result in unintended consequences such as racial and ethnicity based segregation in online spaces. To alleviate such concerns, it would be useful in future to incorporate explainable predictions with work such as (Xie and Lu, 2019), to further mitigate such risks involved.

5 Conclusion and Future Work

Previous works have used tweets or a fraction of the network information to extract features for occupation classification. To analyze the importance of network information, we extended an existing Twitter dataset for a user's social media connections (follow information). We showed that by using only follow information as an input to graph convolutional networks, one can achieve a significantly higher accuracy on the prediction task as compared to the existing approaches utilizing tweet-only information or partial network structure.

Directions of future research include adaptation of our methods to a large scale, sparsely connected social network. One might also want to investigate the inductive settings of GCN (Hamilton et al., 2017) to predict demographic information of a user from outside the black network.

Acknowledgments

We would like to thank the reviewers for their helpful comments on our work. This work is supported by DSO grant DSOCL17061.

References

Nikolaos Aletras and Benjamin Paul Chamberlain. 2018. Predicting twitter user socioeconomic at-

tributes with network and language information. In *Proc. of Hypertext and Social Media*.

John D Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on twitter. In *Proc. of EMNLP*.

Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proc. of CIKM*.

Ido Guy, Naama Zwerdling, Inbal Ronen, David Carmel, and Erel Uziel. 2010. Social media recommendation based on people and tags. In *Proc. of SIGIR*.

Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, pages 1024–1034.

Bo Han, Paul Cook, and Timothy Baldwin. 2013. A stacking-based approach to twitter user geolocation prediction. In *Proc. of ACL (System Demonstrations)*.

Yanxiang Huang, Lele Yu, Xiang Wang, and Bin Cui. 2015. A multi-source integration framework for user occupation inference in social media systems. *World Wide Web*, 18(5):1247–1267.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proc. of ICLR*.

Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Zachary Miller, Brian Dickinson, and Wei Hu. 2012. Gender prediction on twitter using stream algorithms with n-gram character features. *International Journal of Intelligence Science*, 2(04):143.

Andrew Y Ng, Michael I Jordan, and Yair Weiss. 2002. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proc. of KDD*.

Daniel Preoţiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. 2015. An analysis of the user occupational class through twitter content. In *Proc. of ACL-IJCNLP*.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proc. of the 2nd international workshop on Search and mining user-generated contents*.

Jianbo Shi and Jitendra Malik. 2000. Normalized cuts and image segmentation. *Departmental Papers (CIS)*, page 107.

Luke Sloan, Jeffrey Morgan, Pete Burnap, and Matthew Williams. 2015. Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data. *PloS one*, 10(3):e0115545.

Shangsheng Xie and Mingming Lu. 2019. Interpreting and understanding graph convolutional neural network using gradient-based attribution methods. *CoRR*, abs/1903.03768.

Jaewon Yang and Jure Leskovec. 2013. Overlapping community detection at scale: a nonnegative matrix factorization approach. In *Proc. of WSDM*.