Proceedings of the Thirteenth International AAAI Conference on Web and Social Media (ICWSM 2019)

A Co-Training Model with Label Propagation on a Bipartite Graph to Identify Online Users with Disabilities

Xing Yu, Sunandan Chakraborty, Erin Brady
School of Informatics and Computing, Indiana University-Indianapolis
535 W. Michigan St., Indianapolis, IN, USA, 46202
yu64@iu.edu, sunchak@iu.edu, brady@iupui.edu

Abstract

Collecting data from representative users with disabilities for accessibility research is time and resource consuming. With the proliferation of social media websites, many online spaces have emerged for people with disabilities. The information accumulated in such places is of great value for data collection and participant recruiting. However, there are also many active non-representative users in such online spaces such as medical practitioners, caretakers, or family members. In this work, we introduce a novel co-training model based on the *homophily phenomenon* observed among online users with the same disability. The model combines a variational label propagation algorithm and a naive Bayes classifier to identify online users who have the same disability. We evaluated this model on a dataset collected from Reddit and the results show improvements over traditional models.

Introduction

In accessibility studies, it is a common challenge to recruit and collect data from people with certain disabilities (Sears and Hanson 2012). This difficulty largely stems from the fact that people with a specific disability often comprise a relatively small portion of the general population. Researchers need to successfully identify representative users despite factors such as geographic sparsity, inaccessibility of an experimental environment, or lack of contact information.

To mitigate this problem, some existing research work has been carried out with participants that do not have the specific disability instead, who are referred to as *non-representative users* (Sears and Hanson 2012). One common scenario is called simulation, in which a non-representative user simulates a representative user in an experiment (*e.g.*, a sighted person wearing a blindfold in an attempt to simulate the experience of a visually impaired person using a new technology). However, as findings in related work (Heller 1989) pointed out, data and results derived in this manner are often misleading. These types of simulations can also negatively impact the participants' views of people with disabilities (Silverman, Gwinn, and Van Boven 2015). As a result, having an efficient way to gain access to robust, representative user-generated data would be valuable for accessibility researchers.

In this work, we present a co-training model that combines a label propagation algorithm and a naive Bayes classifier to identify representative users on social media websites. Our method was devised based on the assumption of homophily, which presumes that online users with the same disability are closely tied to each other via the disability-related posts in their online social networks. This phenomenon has already been observed among online users with disabilities in existing work (Wu and Adamic 2014; Yu and Brady 2017). The model we propose uses a variational label propagation algorithm to capture the social network information as well as a naive Bayes classifier to capture the textual information in online posts. We carried out experiments based on a dataset collected from Reddit.com to identify amputee users and present the results.

Method

Ideas and Challenges

The homophily principle predicts that users on social media websites are more likely to interact with each other if they share common characteristics (e.g., they are all amputees). To formalize this intuition, we assume an undirected graph $G(V, E)$ in which $v_i \in V$ represent online users and $e_{ij} \in E$ represents an edge that connects v_i and v_j on a social media website. W is a weight function that returns edge weights $w_{ij} = W(e_{ij})$, which is a quantitative measurement of the frequency of interactions that observed between two online users (e.g., exchange of replies). It is natural to assume that $W(e_{ab}) > W(e_{ac})$ if $\phi(v_a) = \phi(v_b)$ and $\phi(v_a) \neq \phi(v_c)$, where $\phi(v_i)$ is a function that returns the class label of v_i which represent the disabilities each user has.

However, the homophily principle also suggests that an individual's social network is complex and based on different levels on different homophily dimensions (McPherson, Smith-Lovin, and Cook 2001). Hence, the assumption that $W(e_{ab}) > W(e_{ac})$ does not always hold for $\phi(v_a) = \phi(v_b)$ and $\phi(v_a) \neq \phi(v_c)$. In a real social network, we may observe that $W(e_{ab}) < W(e_{ac})$ despite $\phi(v_a) = \phi(v_b)$ and $\phi(v_a) \neq \phi(v_c)$ in the scenario that v_a and v_c have strong

homophily on another dimension (*e.g.*, interests in cars). A method to let label information propagate only via desired dimensions could reduce misclassifications in the task.

In order to solve this problem, we introduce a co-training model that combines two algorithms, a variation of the label propagation algorithm and the naive Bayes algorithm, to classify users with disabilities focusing on connections via certain posts in a sparse online social networks derived from social media websites.

Graph Construction

A bipartite graph $G_B(V, P, A)$ is generated to represent the social networks of online users. In G_B, $v_i \in V$ represents online users and $p_i \in P$ represents online posts. V and P are sets of vertices in G_B and $a_{ij} \in A$ are the directed edges that connect the two types of vertices. Since online users V are connected via posts P in this graph, an edge a_{ij} always points from a v_i to a p_j.

Based on the graph G_B, we define two functions, $W_{v\to p}$ and $W_{p\to v}$, for calculating transition probabilities for edges. $W_{v\to p}(a_{ij}) = \frac{r_{v_i\to p_j}}{\sum_{k=1}^{|P|} r_{v_i\to p_k}}$, which returns the transition probability of an edge $a_{ij}(i \neq j)$ that points from v_i to p_j. $r_{v_i\to p_j}$ denotes the count of comments authored by v_i in post p_j. For example, if v_i left three comments in post p_j, then $r_{v_i\to p_j} = 3$. $|P|$ denotes the cardinality of P, which is the number of all post vertices in G_B. $W_{p\to v}(a_{ij}) = \frac{r_{v_i\to p_j}}{\sum_{k=1}^{|V|} r_{v_k\to p_j}}$, which returns the transition probability of the same edge a_{ij} from an opposite direction. $|V|$ denotes the cardinality of V, which is the number of all user vertices in G_B.

Based on the two functions, a normalized adjacency matrix $T \in \mathbb{R}^{(|V|+|P|)\times(|V|+|P|)}$ of G_B can be derived as following:

$$T_{ij} = \begin{cases} W_{v\to p}(a_{i(j-|V|)}) & \text{if } i \leq |V| \text{ and } j > |V| \\ W_{p\to v}(a_{(j-|V|)i}) & \text{if } i > |V| \text{ and } j \leq |V| \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

There are two things worth noting about the bipartite graph. First, each row of the adjacency matrix T is normalized, which is important in proving that the label propagation algorithm would converge in the next section. Second, based on the assumption of homophily, users with the same disability are more likely to be connected to each other via certain post-nodes. These post-nodes are typically discussions of disability related topics. User-nodes connected via posts on other topics may represent other dimensions of homophily, which we try to screen out in this classification task.

Label Propagation on Bipartite Graphs

The key component of our model is a new label propagation algorithm that we designed for the bipartite graph. In this new algorithm, we assume that there are two types of vertices, V and P, in G_B. They both have a labeled and an unlabeled set. The possible labels L for the two types of nodes are the same. The problem setting of our method is different from the version proposed in existing work (Rossi, Lopes, and Rezende 2014) as we include label information for both types of nodes instead of only one.

Given an adjacency matrix $T \in \mathbb{R}^{n\times n}$ and a label matrix $C \in \mathbb{R}^{n\times k}$, in which $n = |V| + |P|$ and $k = |L|$ that $L = \{l_1, l_2, l_3, ...l_k\}$ is the set of all possible labels. $C_{ij} = 1$ if v_i or p_i has label l_j and $C_{ij} = 0$ otherwise. For $v_i \in V$, the label vector C_i represents of the probability of having that corresponding disabilities for v_i. For $p_i \in P$, the label vector C_i represents the probabilities of users who participated in this post may have those disabilities.

The sets V and P are both separated into a labeled and an unlabeled set, which are denoted as V^l, V^u and P^l, P^u. The ultimate goal is to learn the labels of V^u in G_B. The label information C propagates as follows:

$$\begin{bmatrix} C_{V^l} \\ C_{V^u} \\ C_{P^l} \\ C_{P^u} \end{bmatrix} := \begin{bmatrix} T_{V^lV^l} & T_{V^lV^u} & T_{V^lP^l} & T_{V^lP^u} \\ T_{V^uV^l} & T_{V^uV^u} & T_{V^uP^l} & T_{V^uP^u} \\ T_{P^lV^l} & T_{P^lV^u} & T_{P^lP^l} & T_{P^lP^u} \\ T_{P^uV^l} & T_{P^uV^u} & T_{P^uP^l} & T_{P^uP^u} \end{bmatrix} \begin{bmatrix} C_{V^l} \\ C_{V^u} \\ C_{P^l} \\ C_{P^u} \end{bmatrix}$$

The sub-matrices $T_{V^lV^l}$, $T_{V^lV^u}$, $T_{V^uV^l}$, $T_{V^uV^u}$, $T_{P^lP^l}$, $T_{P^lP^u}$, $T_{P^uP^l}$, and $T_{P^uP^u}$ are matrices of 0s due to the fact that the same type of nodes do not have edges among themselves in the bipartite graph G_B. So, label information only propagates across the two types of vertices in each iteration:

$$C_{V^u} := T_{V^uP^l}C_{P^l} + T_{V^uP^u}C_{P^u}$$

$$C_{P^u} := T_{P^uV^l}C_{V^l} + T_{P^uV^u}C_{V^u}$$

By denoting the values of C as $C^{(i)}$, starting as $C^{(0)}$, at the ith iteration, at the nth iteration the value of $C^{(n)}$ can be written as below:

$$\begin{aligned} C_{V^u}^{(n)} = &\sum_{i=0}^{n-1} (T_{V^uP^u}T_{P^uV^u})^i (T_{V^uP^l}C_{P^l} + \\ &T_{V^uP^u}T_{P^uV^l}C_{V^l}) + (T_{V^uP^u}T_{P^uV^u})^n C_{V^u}^{(0)} \end{aligned} \quad (2)$$

$$\begin{aligned} C_{P^u}^{(n)} = &\sum_{i=0}^{n-1} (T_{P^uV^u}T_{V^uP^u})^i (T_{P^uV^l}C_{V^l} + \\ &T_{P^uV^u}T_{V^uP^l}C_{P^l}) + (T_{P^uV^u}T_{V^uP^u})^n C_{P^u}^{(0)} \end{aligned} \quad (3)$$

Since the adjacency matrix T is normalized by row, the row sum of all sub-matrices (*e.g.*, $T_{V^uP^u}$) is less or equal to a value γ smaller than 1. There exists a dot product $B = T_{V^uP^u} \cdot T_{P^uV^u}$ that satisfies the following constraint:

$$\sum_{j}^{|V^u|} B[i,j] \leq \gamma < 1, \ \forall i = 1, 2, ..., |V^u|$$

Based on this constraint, we can prove the following:

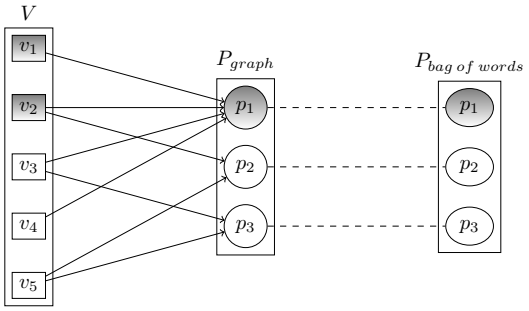

Figure 1: The Co-Training model. Rectangles nodes represent users, circle nodes represent posts. Elliptical nodes are textual information of posts. Shaded nodes belong to the labeled set. Unshaded nodes belong to the unlabeled set.

$$\begin{aligned}\sum_{j}^{|V^u|} B^n[i,j] &= \sum_{j}^{|V^u|}(A^{n-1}A)[i,j] \\ &= \sum_{j}^{|V^u|}\sum_{k}^{V^u} B^{n-1}[i,k]B[k,j] \\ &= \sum_{k}^{|V^u|} B^{n-1}[i,k]\sum_{j}^{|V^u|} B[k,j] \\ &\leq \sum_{k}^{|V^u|} B^{n-1}[i,k]\gamma \leq \gamma^n, \forall i = 1,2,3,...,|V^u|\end{aligned}$$

Thus, each row's summation of B approximates 0 when $n \to \infty$, the following terms become matrices of zeros:

$$\begin{aligned}\lim_{n\to\infty}(T_{V^uP^u}T_{P^uV^u})^n C^{(0)}_{V^u} = 0 \\ \lim_{n\to\infty}(T_{P^uV^u}T_{V^uP^u})^n C^{(0)}_{P^u} = 0\end{aligned} \tag{4}$$

It is clear, by plugging equations 4 back into equations 2 and 3, that the results of label propagation does not depend on the initial value of C_{V^u} and C_{P^u}. The algorithm will converge eventually based on the adjacency matrix T and the label matrix C as long as they are normalized by row.

We refer to this variation as *label propagation on a bipartite graph* (LPBG). LPBG returns the probability matrix C at termination, which can be used for class assignments.

Co-Training Process

In the co-training model, we assume each post $p_i \in P$ has two representations. x_1 is the bipartite graph G_B that represents the network information. x_2 is the bag-of-words representation of the linguistic information in each post.

The structure of the two representations in the co-training model is depicted in Figure 1. We use LPBG as the first classifier f_1 on representation x_1, P_{graph} in Figure 1, and a naive Bayes (NB) classifier, which is commonly applied in text classification (Lewis and Ringuette 1994), as the second classifier f_2 on representation x_2, $P_{bag\ of\ words}$ in Figure 1. The model trains classifiers f_1 and f_2 independently in each iteration. Each of the newly trained classifiers learns k most confident instances for each class from the unlabeled set P^u. Then $|L| \times k$ newly learned instances are removed from P^u and added into P^l for the next iteration. When sufficient label information is learned for P^u after certain number of iterations, the model uses LPBG to learn the final labels of online users.

This new design has two benefits. First, the LPBG algorithm restricts information propagation among user nodes through certain post-nodes. Second, the NB classifier assigns labels information to disconnected post-nodes to mitigate the problem of sparse graphs.

Experiment

We compiled a new dataset for the experiment by collecting all posts and comments from two amputee-relevant subreddits: [1]/r/amputee and [2]/r/prosthetics on Reddit. Two accessibility researchers manually annotated class labels (Cohen's κ = 0.93) for each online user. Finally, we derived 619 users that include 221 amputees and 398 non-amputees, and collected 614,256 posts from 2008 to 2018 authored by them.

The input of the Co-Training model includes a bipartite graph (denoted as G_B) and a corpus of online posts (denoted as D). Each document $d_i \in D$ contains all the words in the post p_i. We initialized the training process by generating V^l based on the degrees of user vertices. The process randomly samples (without replacement) $0.5 \times |V^l|$ representative users and $0.5 \times |V^l|$ unrepresentative users based on node degrees in the graph. Intuitively speaking, users who participated in more posts would have high chances of being sampled into the training set. Based on the chosen set of V^l, all $p_i \in P$ in G_B that are connected to $v_i \in V^l$ are collected. The set is used as P^l where each post has the same label as the majority label of the user-nodes connected to it.

An under-sampling process is carried out to balance the instances with different labels in P^l to improve the performance of the NB classifier in the co-training process. We extract $d_i \in D$ that match $p_i \in P^l$ to create a subset corpus D_{P^l}, which is used as the training data for the naive Bayes classifier. Each document is represented as a TF–IDF vector in the training process.

A third group of users in G_B participated in at least one post $p_i \in P$ but do not have any labels. These users are included when generating the graph to keep the edge weights and proportions of users in posts accurate, but they are removed in the co-training process.

We set the model to learn 150 instances for each class at each iteration. After 6 iterations, a final iteration of LPBG is carried out, which returns C_{V^u}. Then the label for $v_i \in V^u$ is derived by choosing the one class with the highest probability (Equation 5).

$$\phi(V_i) = L[\underset{j}{argmax}\ C_{V^U}[i,j]] \tag{5}$$

[1]https://www.reddit.com/r/amputee

[2]https://www.reddit.com/r/prosthetics

Set size	Methods	Precision	Recall	Accuracy
25%	**Co-Training**	**0.80**	**0.59**	**0.82**
	Baseline 1 (LP)	0.30	0.02	0.64
	Baseline 2 (NB)	0.69	0.65	0.65
	Baseline 3 (Topic)	0.53	0.64	0.63
50%	**Co-training**	**0.84**	**0.75**	**0.89**
	Baseline 1 (LP)	0.39	0.13	0.64
	Baseline 2 (NB)	0.51	0.66	0.66
	Baseline 3 (Topic)	0.62	0.57	0.57
75%	**Co-Training**	**0.82**	**0.89**	**0.90**
	Baseline 1 (LP)	0.57	0.05	0.65
	Baseline 2 (NB)	0.70	0.65	0.65
	Baseline 3 (Topic)	0.61	0.64	0.64

Table 1: Average metrics of 5 runs of the Co-Training model and baseline methods with different sizes of training sets.

Results

Table 1 shows the average results of five runs of the co-training process that were initiated with randomly sampled training sets. The results of different sizes of training sets (25%, 50%, and 75% with regard to the total number of users) are included for comparison.

Baseline 1 is the label propagation algorithm (LP) proposed in (Rossi, Lopes, and Rezende 2014) that we directly applied on the bipartite graph G_B. In baseline 2, we used a NB classifier based on TF-IDF vectors. In baseline 3, we used topic modeling to reduce the input dimensions. We first applied latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003) to learn topics distribution for each post. Then the topic distribution vectors are used to classify the users. The co-training model performs better than all the baselines. As a semi-supervised learning algorithm, it also demonstrates good results with a training set of 25% of the entire dataset.

Baseline 1 showed low recall, not surprising given the fact that real world social networks are very sparse/disconnected graphs. Performance fluctuates depending on the choice of initial nodes. From the results of baseline 2 and 3, we see that the performance of solely using language cues to classify users is not ideal. This is consistent with our observation that amputee users' post content is not that different from other users overall, since their online activities cover a wide range of topics besides disability-specific content.

Discussion

Iterations: The number of iterations in the co-training model affects its performance. We observed that the overall accuracy peaks after 6 iterations and drops afterwards. This problem is caused by the unbalanced classes in the dataset. Around 6 iterations, we observed that the NB classifier mostly generates negative instance as most positive instances are already identified, and the performance of label propagation drops when the graph is overflowed with negative instance. Hence, an optimal number of iterations should be chosen when applying the Co-Training model on unbalanced datasets.

Performance: The model is a fast semi-supervised learning algorithm. The choice of the second model is flexible (*e.g.*, a SVM with non-linear kernel based on word2vec representation). However, the choice of the second classifiers would affect the speed of the training process, which should be taken in consideration when designing the model.

Limitations: Although our newly proposed model performed better than the baselines and achieved a reasonable result in the classification task, more future work is necessary. First, we evaluated the model in a binary classification task in our experiment. However, the model is theoretically applicable for multi-class classification, for example identifying people with different disabilities from larger forums. Hence, a larger dataset with multiple classes could verify the model in future work. Second, it will be informative to compare this model with other node classification models to check its superiority.

Conclusions

In order to classify posts by online users with disabilities, we designed a variational label propagation algorithm and introduced a co-training model. We tested the new model in an experiment using a dataset collected from Reddit. Our model achieved an overall accuracy of 82% with 25% data as the training set. The accuracy reaches 90% when the training set size is increased to 75%. The results are significantly better than baseline methods. This efficient semi-supervised model brings potential solutions for accessibility researchers to carry out participant recruiting and data collection to compensate the difficulties they experience in recruiting.

References

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

Heller, M. A. 1989. Picture and pattern perception in the sighted and the blind: the advantage of the late blind. *Perception* 18(3):379–389.

Lewis, D. D., and Ringuette, M. 1994. A comparison of two learning algorithms for text categorization. In *Third annual symposium on document analysis and information retrieval*, volume 33, 81–93.

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

Rossi, R. G.; Lopes, A. A.; and Rezende, S. O. 2014. A parameter-free label propagation algorithm using bipartite heterogeneous networks for text classification. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, 79–84. ACM.

Sears, A., and Hanson, V. L. 2012. Representing users in accessibility research. *ACM Transactions on Accessible Computing (TACCESS)* 4(2):7.

Silverman, A. M.; Gwinn, J. D.; and Van Boven, L. 2015. Stumbling in their shoes: Disability simulations reduce judged capabilities of disabled people. *Social Psychological and Personality Science* 6(4):464–471.

Wu, S., and Adamic, L. A. 2014. Visually impaired users on an online social network. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 3133–3142. ACM.

Yu, X., and Brady, E. 2017. Understanding and classifying online amputee users on reddit. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, 17–22. ACM.