Clustered Model Adaption for Personalized Sentiment Analysis

Lin Gong, Benjamin Haines, Hongning Wang
Department of Computer Science
University of Virginia, Charlottesville VA, 22904 USA
{lg5bt,bmh5wx,hw5x}@virginia.edu

ABSTRACT

We propose to capture humans' variable and idiosyncratic sentiment via building personalized sentiment classification models at a group level. Our solution roots in the *social comparison theory* that humans tend to form groups with others of similar minds and ability, and the *cognitive consistency theory* that mutual influence inside groups will eventually shape group norms and attitudes, with which group members will all shift to align. We formalize personalized sentiment classification as a multi-task learning problem. In particular, to exploit the clustering property of users' opinions, we impose a non-parametric Dirichlet Process prior over the personalized models, in which group members share the same customized sentiment model adapted from a global classifier. Extensive experimental evaluations on large collections of Amazon and Yelp reviews confirm the effectiveness of the proposed solution: it outperformed user-independent classification solutions, and several state-of-the-art model adaptation and multi-task learning algorithms.

CCS Concepts

•**Information systems** → **Sentiment analysis; Clustering and classification;**

Keywords

Sentiment analysis, model adaptation, multi-task learning

1. INTRODUCTION

Traditional solutions for text-based sentiment modeling mostly focus on building population-level supervised classifiers [29, 28, 36], which estimate and apply a shared classifier across all users' opinionated data. This postulates a strong assumption that the joint probability of sentiment labels and text content is independent and identical across users. However, this assumption is usually undermined in practice: it is well known in social psychology and linguistic studies that sentiment is personal and humans have diverse ways of expressing attitudes and opinions [37]. Hence, a single generic sentiment model can hardly capture the heterogeneity among users, and it will inevitably lead to inaccurate opinion mining results. Explicitly modeling the heterogeneity to capture individualized opinions is thus of particular importance.

Estimating a personalized sentiment model is challenging. Sparsity of individual users' opinionated data prevents us from estimating supervised classifiers on a per-user basis. Some existing works utilize semi-supervised methods to address the sparsity issue. For example, [18, 33] utilized user-user and user-document relations as regularizations to perform transductive learning. However, only one global sentiment model is estimated in such solutions, and it cannot capture the nuance in which individual users express their diverse opinions. [1] developed a transfer learning solution to adapt a global sentiment model to each individual user, but limited improvement is achieved on users with few observations, who form a major portion of the user population.

In this work, we take a new perspective to build personalized sentiment models by exploiting social psychology theories about humans' dispositional tendencies. First, the theory of social comparison [7] states that the drive for self-evaluation can lead people to associate with others of similar opinions and abilities, thus to form groups. This guarantees the relative homogeneity of opinions and abilities within groups. In our solution, we capture such clustering property of different users' opinions by postulating a non-parametric Dirichlet Process (DP) prior [12] over the individualized models, such that those models automatically form latent groups. In the posterior distribution of this postulated stochastic process, users join groups by comparing the likelihood of generating their own opinionated data in different groups (i.e., realizing self-evaluation and group comparison). Second, according to the cognitive consistency theory [25], once the groups are formed, members inside the same group will be influenced by other in-group members mutually through both implicit and explicit information sharing, which leads to the development of group norms and attitudes [32]. We formalize this by adapting a global sentiment model to individual users in each latent user group, and jointly estimating the global and group-wise sentiment models. The shared global model can be interpreted as the global social norm, because it is estimated based on observations from all users. It thus captures homogenous sentimental regularities across users. The group-wise adapted models capture heterogenous sentimental variations among users across groups. Because of this two-level information grouping and sharing, the complexity of preference learning will be largely reduced. This is of particular value for sentiment analysis in tail users, who only possess a handful of observations but take the major proportion in user population.

We should note that our notion of user group is different from those in traditional social network analysis, where user interaction or community structure is observed. In our solution, user groups are *latent*: they are formed based on the textual patterns in users'

WWW 2017, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4913-0/17/04.
http://dx.doi.org/10.1145/3038912.3052693

sentimental expressions, i.e., implicit sentimental similarity instead of direct influence, such that members inside the same latent group are not necessarily socially connected. This aligns with our motivating social psychology theories: people who have similar altitudes or behavior patterns might not know each other, while they interact via implicit influence, such as being exposed to the same social norms or read each others' opinionated texts. Being able to quantitatively identify such latent user groups also provides a new way of social network analysis – content-based community detection. But this is beyond the scope of this paper.

Our proposed solution can also be understood from the perspective of multi-task learning [10, 19, 39]. In particular, the problem of personalized sentiment classification can be considered as estimating a set of related classifiers across users. In our solution, we formalize this idea as *clustered model sharing and adaptation across users*. We assume the distinct ways in which users express their opinions can be characterized by different configurations of a linear classifier's parameters, i.e., the weights of textual features. Individualized models can thus be achieved via a series of linear transformations over a globally shared classifier, e.g., shifting and scaling the weight vector [1]. Moreover, we enforce the relatedness among users via the automatically identified user groups – users in the same group would receive the same set of model adaptation operations. The user groups are jointly estimated with the group-wise and global classifiers, such that information is shared across users to conquer data sparsity in each user and non-linearity is achieved when performing sentiment classification across users.

We performed extensive experimentations on two large collections of Amazon and Yelp reviews to evaluate our solution. It outperformed user-independent classification methods, and several state-of-the-art model adaption and multi-task learning algorithms.

2. RELATED WORK

Building personalized sentiment classifiers can be considered as a multi-task learning problem, which exploits the relatedness among multiple learning tasks to benefit each individual task. Tasks can be related in various ways. A typical assumption is that all models learned are close to each other in some matrix norm of their model parameters [10, 19]. This assumption has been empirically proved to be effective for modeling consumer preferences in market research [11]. [8] proposed a simultaneous co-clustering algorithm between customers and products considering the dyadic property of the data. Some recent efforts suggest that relatedness between tasks should also be estimated to restrict information sharing only within similar tasks [34, 3]. Dirichlet Process prior [12] naturally satisfies this goal: it associates related tasks into groups via exploiting the clustering property of data. [21] utilized the property to achieve content personalization of users by generating both the latent domains and the mixture of domains for each user. And they also trained the personalized models using the multi-task learning idea to capture heterogeneity and homogeneity among users with respect to the content. Their solution is different from ours as we consider clustering users regarding to opinionated sentiment models. [39, 31] estimated a set of linear classifiers in automatically identified groups. However, sparsity of personal opinionated data in the sentiment analysis scenario still limits the practical value of conventional multi-task learning algorithms, since in each task a full set of model parameters still have to be estimated. Our solution instead only learns simple model transformations over groups of features in each task [1], which greatly reduces the overall model learning complexity. And because the number of groups is automatically identified from data, it naturally balances sample complexity in learning group-wise models.

The proposed solution is also closely related to model adaptation, which is an important topic in transfer learning [27]. In the opinion mining community, model adaptation techniques are mostly exploited for domain adaptation, e.g., adapting sentiment classifiers trained on book reviews to DVD reviews [6, 26, 38]. There are also some recent works that attempt to perform model adaptation on a per-user basis for sentiment classification. Li et al. proposed an online learning algorithm to continue training personalized classifiers from a shared global model [20]. [1] applied the idea of linear transformation based model adaptation for personalized sentiment classifier training. [17] adapted individual user models from a updated global model to achieve user personalization. However, no existing work in model adaptation considers the relatedness among users, and thus adaptations are performed in an isolated manner. Our solution enforces users in the same group to share the same set of adaptation parameters and links models in different user groups by a globally shared model, which propagates information among users to overcome the data sparsity issue.

3. METHODOLOGY

Our solution roots in the social comparison theory and cognitive consistence theory. Specifically, we build personalized sentiment classification models via a set of shared model adaptations for both a global model and individualized models in groups. The latent user groups are identified by imposing a Dirichlet Process prior over the individual models. In the following, we first discuss the motivating social behavior theories, and then carefully describe how we formulate these social concepts to computational models for personalized sentiment analysis.

3.1 Group Formation and Group Norms

In social science, the theory of social comparison explains how individuals evaluate their own opinions and abilities by comparing themselves to others in order to reduce uncertainty when expressing opinions and learn how to define themselves [13]. In the context of sentiment analysis, we consider building personalized sentiment models as a set of inductive tasks. Because of the explicit and implicit comparisons users have performed when generating the opinionated data, those learning tasks become related. [23] further suggested the drive for self-evaluation leads people to associate with others of similar minds to form (latent) groups, and this guarantees the relative homogeneity of opinions within groups. In sentiment analysis, this can be translated as model regularization among users in the same group. Correspondingly, the process of self-definition can be considered as people recognizing a specific group after comparison, i.e., joining an existing similar group or creating a new distinct group after evaluating both self and group information. This further suggests us to build personalized models in a group-wise manner and identify the latent groups by exploiting the clustering property of users' opinionated data.

Once the groups of similar opinions are formed, cognitive consistency theory [14, 25] suggests that members in the same group interact mutually in order to reduce the inconsistency of opinions, and this eventually leads to group norms that all members will shift to align with. Group norms thus act as powerful force that dramatically shapes and exaggerates individuals' emotional responses [4]. Such groups are not necessarily defined by observed social networks, as the influence can take forms of both implicit and explicit interactions. In the context of sentiment analysis, we capture group norms by enforcing users in the same group to share *identical* sentiment models. Heterogeneity is thus characterized by the distinct sentiment models across groups. This reduces the learning complexity from per-user model estimation to per-group. Besides the group norms, the simultaneously estimated global model provides the basis for group norms to evolve from, which represents the homogeneity among all users.

3.2 Personalized Model Adaptation

We assume the diverse ways in which users express their opinions can be characterized by different settings of a linear classifier, i.e., the weight vector of textual features. We choose to estimate a linear classifier for each user to model sentiment, because of its empirically superior performance in text-based sentiment analysis [29, 28]. But the proposed solution can be easily extended to non-linear classification models, with the constraints that the model takes a linear combination of features in its core computation and its likelihood function can be readily evaluated at given data points.

Formally, denote a collection of N users as $U = \{u_1, u_2, ...u_N\}$, in which each user u is associated with a set of opinionated text documents as $D^u = \big\{(x^u_d, y^u_d)\big\}_{d=1}^{|D^u|}$. Each document d is represented by a V-dimension vector x_d of textual features, and y_d is the corresponding sentiment label. We assume each user is associated with a sentiment model $f(x;\omega^u) \rightarrow y$, which is characterized by the individualized feature weight vector ω^u. Estimating $f(x;\omega^u)$ for users in U is the inductive learning task of our focus.

Instead of assuming $f(x;\omega^u)$ is solely estimated from the user's own opinionated data, we further assume it is obtained from a global sentiment model $f(x;\omega^s)$ via a series of linear model transformations [1, 35], i.e., shift and scale the shared model parameter ω^s into ω^u based on D^u. To simplify the discussions in this paper, we assume binary sentiment classification, i.e., $y \in \{0, 1\}$, and we will use logistic regression as the reference model in the following discussions. To handle sparse observations in each individual users' opinionated data, we further assume that model adaptations can be performed in feature groups [35]. Specifically, features in the same group will be updated synchronously by performing the same set of shifting and scaling operations, i.e., shift and scale the model weights. This enables information propagation from seen features to unseen features in the same feature group. Various feature grouping methods have been explored in [35], and we directly employed their methods for this purpose, since feature grouping is not the contribution of this work.

We define $g(i) \rightarrow k$ as the feature grouping method, which maps feature i in $\{1, 2, \ldots, V\}$ to feature group k in $\{1, 2, \ldots, K\}$. The set of personalized model adaptation operations in user u can then be represented as a $2K$-dimension vector $\theta^u = (a^u_1, a^u_2, \ldots, a^u_K, b^u_1, b^u_2, \ldots, b^u_K)$, where a^u_k and b^u_k represent the scaling and shifting operations in feature group k for user u. This gives us a one-to-one mapping of feature weights from global model ω^s to personalized model ω^u as $\forall i \in \{1, 2, \ldots, V\}$, $\omega^u_i = a^u_{g(i)}\omega^s_i + b^u_{g(i)}$. Because θ^u uniquely determines the personalized feature weight vector ω^u, we will then refer to θ^u as the personalized sentiment model for user u in our discussions.

Different from what has been explored in [1, 35], where the global model ω^s is predefined and fixed, we assume ω^s is unknown and dynamic. Therefore, it needs to be learnt based on the observations from all the users in U. This helps us capture the variability of people's sentiment, such as the dynamics of social norms. In particular, we apply the same linear transformation method to adapt ω^s from a predefined sentiment model ω^0. ω^0 can be empirically set based on a separate user-independent training set, e.g., pooling opinionated data from different but related domains. Since this transformation will be jointly estimated across all users, a different feature mapping function $g'(\cdot)$ can be used to organize features into more groups to increase the resolution of sentiment classification in the global model. We denote the corresponding global model adaptation as $\theta^s = (a^s_1, a^s_2, \ldots, a^s_L, b^s_1, b^s_2, \ldots, b^s_L)$, in which additional degree of freedom is given to the feature group size L. The benefit of this second-level model adaptation is two-fold. First, the predefined sentiment model ω^0 can serve as a prior for global sentiment classification [1]. This benefits multi-task learning when the overall observations are sparse. Second, non-linearity among features is introduced when the global model and personalized models employ different feature groupings. This enables observation propagation across features in different user groups.

Plugging this two-level linear transformation based model specification into the logistic function, we can materialize the personalized logistic regression model for user u as,

$$P(y^u_d = 1|x^u_d, \theta^u, \theta^s, \omega^0) = \sigma\Big(\sum_{k=1}^{K}\sum_{g(i)=k}(a^u_k\omega^s_i + b^u_k)x^u_{d,i}\Big) \quad (1)$$

where $\omega^s_i = a^s_{g'(i)}\omega^0_i + b^s_{g'(i)}$ and $\sigma(x) = \frac{1}{1+\exp(-x)}$.

3.3 Non-parametric Modeling of Groups

The inductive learning task in each user u hence becomes to estimate θ^u that maximizes the likelihood of the user's own opinionated data defined by Eq (1). Accordingly, a shared task for all users is to estimate θ^s with respect to the likelihood over all of their observations. As we discussed in the related social theories about humans' dispositional tendencies, people tend to automatically form groups of similar opinions, and follow the mutually reinforced group norms in their own behavior. Therefore, instead of estimating the personalized model adaptation parameters $\{\theta^u\}^N_{u=1}$ independently, we assume they are grouped and those in the same group share *identical* model adaptation parameters.

Determining the task grouping structure in multi-task learning is challenging, because the optimal setting of individual models is unknown beforehand and it will also be affected by the imposed task grouping structure. Ad-hoc solutions approximate the group structure by first performing clustering in the feature space [5] or individually trained models [16], and then restarting the learning tasks with the fixed task structure as additional regularization. Unfortunately, such solutions have serious limitations: 1) they isolate the learning of task relatedness structure from the targeted learning tasks; 2) one has to manually exhaust the number of clusters;p and 3) the identified task grouping structure introduces unjustified bias into multi-task learning. To avoid such limitations, we appeal to a non-parametric approach to jointly estimate the task grouping structure and perform multi-task learning across users.

Motivated by the social comparison theory, in our solution instead of considering the optimal setting of $\{\theta^u\}^N_{u=1}$ as fixed but unknown, we treat it as stochastic by assuming each user's model parameter θ^u is drawn from a Dirichlet Process prior [12, 2]. A Dirichlet Process (DP), $DP(\alpha, G_0)$ with a base distribution G_0 and a scaling parameter α, is a distribution over distributions. An important property of DP is that samples from it often share some common values, and therefore naturally form clusters. The number of unique draws, i.e., the number of clusters, varies with respect to the data and therefore is random, instead of being pre-specified.

Introducing the DP prior thus imposes a generative process over the learning task in each individual user in our problem. This process can be formally described as follows,

$$\begin{aligned} G &\sim DP(\alpha, G_0),\\ \theta^u|G &\sim G, \\ y^u_d|x^u_d, \theta^u, \theta^s, \omega^0 &\sim P(y^u_d = 1|x^u_d, \theta^u, \theta^s, \omega^0). \end{aligned} \quad (2)$$

where the hyper-parameter α controls the concentration of unique draws from the DP prior, the base distribution G_0 specifies the prior distribution of the parameters in each individual model, and G represents the mixing distribution of the sampled results of θ^u. To simplify the notations for discussion, we define $\boldsymbol{a}^u$ and $\boldsymbol{b}^u$ as the scaling and shifting components in θ^u, such that $\theta^u = (\boldsymbol{a}^u, \boldsymbol{b}^u)$. We impose an isometric Gaussian distribution in G_0 over θ^u as $\theta^u \sim N(\mu, \sigma^2)$, where $\mu = (\mu^{\boldsymbol{a}}, \mu^{\boldsymbol{b}})$ and $\sigma = (\sigma^{\boldsymbol{a}}, \sigma^{\boldsymbol{b}})$ accordingly. That is, we allow the shifting and scaling operations to be generated from different prior distributions. Correspondingly, we also treat the globally shared model adaptation parameter θ^s as a latent random variable, and impose another isometric Gaussian prior over it as $\theta^s \sim N(\mu_s, \sigma_s^2)$, where μ_s and σ_s^2 are also decomposed with respect to the shifting and scaling operations.

By integrating out G in Eq (2), the predictive distribution of θ^u conditioned on the individualized models in the other users, denoted as $\theta^{-u} = \{\theta^1, .., \theta^{u-1}, \theta^{u+1}, ...\theta^N\}$, can be analytically computed as follows,

$$p(\theta^u|\theta^{-u}, \alpha, G_0) = \frac{\alpha}{N-1+\alpha}G_0 + \frac{1}{N-1+\alpha}\sum_{j\neq i}^{N}\delta_{\theta^u}(\theta^j) \quad (3)$$

where $\delta_{\theta^u}(\cdot)$ is the distribution concentrated at θ^u.

This predictive distribution well captures the idea of social comparison theory. On the one hand, the second part of this predictive distribution captures the process that a user compares his/her own sentiment model against the other users' models, as the distribution $\delta_{\theta^u}(\cdot)$ takes probability one only when $\theta^j = \theta^u$, i.e., they hold the same sentiment model. Hence, a user tends to join groups with established sentiment models, and this probability is proportional to the popularity of this sentiment model in overall user population. On the other hand, the first part of Eq (3) captures the situation that a user decides to form his/her own sentiment model, but this probability is small when the user population is large. As a result, the imposed DP prior encourages users to form shared groups.

We denote the unique samples in G as $\{\phi_1, \phi_2, \ldots, \phi_c\}$, i.e., the group models, where the group index c takes value from 1 to ∞, and ϕ_i represents the homogeneity of sentiment models in user group i. We should note that the notion of an infinite number of groups is only to accommodate the possibility of generating new groups during the stochastic process. As the sample distribution G resulting from the DP prior in Eq (2) only has finite supports at the points of $\{\theta^1, \theta^2, \ldots, \theta^N\}$, the maximum value for c is N, i.e., all users have their own unique sentiment models. Then the likelihood of the opinionated data in user u can be computed under the stick-breaking representation of DP [30] as follows:

$$\begin{aligned} & P(\boldsymbol{y}^u|\boldsymbol{x}^u, \omega^0, \alpha, G_0) \quad (4) \\ = & \int d\phi \int d\theta^s \int d\pi \sum_{c_u=1}^{\infty} \prod_{d=1}^{|D^u|} P(y_d^u|x_d^u, \phi_{c_u}, \theta^s, \omega^0) p(c_u|\boldsymbol{\pi}) \\ & p(\phi_{c_u}|\mu, \sigma^2) p(\theta^s|\mu_s, \sigma_s^2) p(\pi|\alpha) \end{aligned}$$

where $\boldsymbol{\pi} = (\pi_c)_{c=1}^{\infty} \sim Stick(\alpha)$ captures the proportion of unique sample ϕ_c in the whole collection. And the stick-breaking process $Stick(\alpha)$ for π is defined as: $\pi'_c \sim Beta(1, \alpha), \pi_c = \pi'_c \prod_{t=1}^{c-1}(1-\pi'_t)$, which is a generalization of multinomial distribution with a countably infinite number of components.

As the components to be estimated in each latent puser group (i.e., $\{\phi_c\}_{c=1}^{\infty}$) is a set of linear model transformations, we name the resulting model defined by Eq (4) as *Clustered Linear Model Adaptation*, or *cLinAdapt* in short. And using the language of graphical models, we illustrate the dependency between different components of cLinAdapt in Figure 1. We should note that our cLinAdapt model is not a fully generative model: as defined in Eq (4), we treat the documents $\{\boldsymbol{x}^u\}_{u=1}^N$ as given and do not specify any generation process on them. The group membership variable c_u can thus only be inferred for users with at least one labeled document, since that is the only supervision for group membership inference. As a result, we assume the group membership for each user is *stationary*: once inferred from training data, it can be used to guide personalized sentiment classification in the testing phase. Modeling the dynamics in such latent groups is outside the scope of this work.

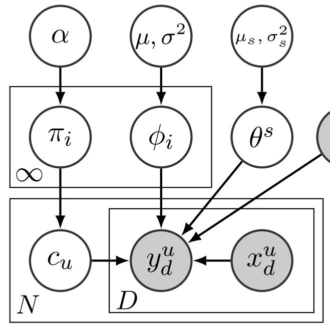

Figure 1: Graphical model representation of cLinAdapt. Light circles denote the latent random variables, and shadow circles denote the observed ones. The outer plate indexed by N denotes the users in the collection, the inner plate indexed by D denotes the observed opinionated data associated with user u, and the upper plate denotes the parameters for the countably infinite number of latent user groups in the collection.

3.4 Posterior Inference

To apply cLinAdpat for personalized sentiment classification, we need to infer the posterior distributions of: 1) group-wise model adaptation parameters $\{\phi_c\}_{c=1}^{\infty}$, each one of which captures the homogeneity of personalized sentiment models in a corresponding latent user group; 2) global model adaptation parameter θ^s, which is shared by all users' sentiment models; 3) group membership variable c_u for user u; and 4) sentiment labels $\boldsymbol{y}^u$ for testing documents in user u. However, because there is no conjugate prior for the logistic regression model, exact inference for cLinAdapt becomes intractable. In this work, we develop a stochastic Expectation Maximization (EM) [9] based iterative algorithm for posterior inference in cLinAdapt. In particular, Gibbs sampling is used to infer the group membership $\{c_u\}_{u=1}^N$ for all users based on the current group models $\{\phi_c\}_{c=1}^{\infty}$ and global model θ^s, and then maximum likelihood estimation for $\{\phi_c\}_{c=1}^{\infty}$ and θ^s is performed based on the newly updated group membership $\{c_u\}_{u=1}^N$ and corresponding observations in users. These two steps are repeated until the likelihood on the training data set converges. During the iterative process, the posterior of $\boldsymbol{y}^u$ in testing documents in user u is accumulated for final prediction.

Next we will carefully describe the detailed procedures of each step in this iterative inference algorithm.

• **Inference for $\{c_u\}_{u=1}^N$:** Following the sampling scheme proposed in [24], we introduce a set of auxiliary random variables of size m, i.e., $\{\phi_i^a\}_{i=1}^m$, drawn from the same base distribution G_0 to define a valid Markov chain for Gibbs sampling over $\{c_u\}_{u=1}^N$. To facilitate the description of the developed sampling scheme, we assume that at a particular step in sampling c_u for user u, there are in total C active user groups (i.e., groups that associate with at least one user, excluding the current user u), and by permuting the indices, we can index them from 1 to C. By denoting the number of users in group c as n_c^{-u} (excluding the current user u), the posterior distribution of c_u can be estimated by,

$$P\big(c_u = c|\boldsymbol{y}^u, \boldsymbol{x}^u, \{\phi_i\}_{i=1}^C, \{\phi_j^a\}_{j=1}^m, \theta^s, \omega^0\big) \propto \tag{5}$$
$$\begin{cases} n_c^{-u} \prod_{d=1}^{|D^u|} P(y_d^u|x_d^u, \phi_c, \theta^s, \omega^0) & \text{for} \quad 1 \leq c \leq C, \\ \frac{\alpha}{m} \prod_{d=1}^{|D^u|} P(y_d^u|x_d^u, \phi_c^a, \theta^s, \omega^0) & \text{for} \quad 1 < c \leq m. \end{cases}$$

If an auxiliary variable is chosen for c_u, it will be appended to $\{\phi_i\}_{i=1}^C$ as one extra active user group.

Because of the introduction of auxiliary variables $\{\phi_i^a\}_{i=1}^m$, the integration of $\{\phi_c\}_{c=1}^\infty$ with respect to the base distribution G_0 is approximated by a finite sum over the current active groups and auxiliary variables. Therefore, the number of sampled auxiliary variables affects accuracy of this posterior. To avoid bias in sampling c_u, we will draw a new set of auxiliary variables from G_0 every time when sampling. As the prior distributions for θ^u in G_0 are Gaussian, sampling the auxiliary variables is efficient.

We should note that the sampling step derived in Eq (5) for cLinAdapt is closely related to the social comparison theory. The auxiliary variables can be considered as pseudo groups: no user has been assigned to them but they provide options for constructing new sentiment models. When a user develops his/her own sentiment model, he/she will evaluate the likelihood of generating his/her own opinionated data under all candidate models together with such a model's current popularity among other users. In this comparison, the likelihood function serves as a similarity measure between users. Additionally, new sentiment models will be created if no existing model can well explain this user's opinionated data. This naturally determines the proper size of user groups with respect to the overall data likelihood during model update.

• **Estimate for $\{\phi_c\}_{c=1}^\infty$ and θ^s:** Once the group membership $\{c_u\}_{u=1}^N$ is sampled for all users, the grouping structure among individual learning tasks is known, and the estimation for $\{\phi_c\}_{c=1}^\infty$ and θ^s can be readily performed by maximizing the complete-data likelihood based on the current group assignments.

Specifically, assume there are C active user groups after the sampling of $\{c_u\}_{u=1}^N$, the complete-data log-likelihood over $\{\phi_c\}_{c=1}^C$ and θ^s can be written as,

$$L\big(\{\phi_c\}_{c=1}^C, \theta^s\big) = \sum_{u=1}^N \log P(\boldsymbol{y}^u|\boldsymbol{x}^u, \phi_{c_u}, \theta^s, \omega^0) \tag{6}$$
$$+ \sum_{c=1}^C \log p(\phi_c|\mu, \sigma^2) + \log p(\theta^s|\mu_s, \sigma_s^2)$$

As the global model adaptation parameter θ^s is shared by all the users (as defined in Eq (1)), it makes the estimation of $\{\phi_c\}_{c=1}^C$ dependent across all the user groups, i.e., information sharing across groups in cLinAdapt.

In Section 3.3, we did not specify the detailed configuration of the prior distributions on θ^u and θ^s, i.e., Gaussian's mean and standard deviation. But given θ^u and θ^s stand for linear transformations in model adaptation, proper assumption can be postulated on their priors. In particular, we believe the scaling parameters should be close to one and shifting parameters should be close to zero, i.e., $\mu^{\boldsymbol{a}} = 1$ and $\mu^{\boldsymbol{b}} = 0$, to encourage individual models to be close to the global model (i.e., reflecting social norm). The standard deviations control the confidence of our belief and can be empirically tuned. The same treatment also applies to μ_s and σ_s^2 for the global model adaptation parameter θ^s.

Eq (6) can be efficiently maximized by a gradient-based optimizer, and the actual gradients of Eq (6) reveal the insights of our proposed two-level model adaptation in cLinAdapt. For illustration purpose, we only present the decomposed gradients with respect to the complete-data log-likelihood for scaling operation in ϕ_c and θ^s on a specific training instance (x_d^u, y_d^u) in user u:

$$\frac{\partial L(\cdot)}{\partial a_k^{c_u}} = \Delta_d^u \sum_{g(i)=k} \Big(a_{g'(i)}^s \omega_i^0 + b_{g'(i)}^s\Big) x_{di}^u + \frac{a_k^{c_u} - 1}{\sigma^2} \tag{7}$$

$$\frac{\partial L(\cdot)}{\partial a_l^s} = \Delta_d^u \sum_{g'(i)=l} a_{g(i)}^{c_u} \omega_i^0 x_{di}^u + \frac{a_l^s - 1}{\sigma_s^2} \tag{8}$$

where $\Delta_d^u = y_d^u - P(y_d^u = 1|x_d^u, \phi_{c_u}, \theta^s, \omega^0)$, and $g(\cdot)$ and $g'(\cdot)$ are the feature grouping functions for individual users' and global model adaptation. First, observations from all group members will be aggregated to update the group-wise model adaptation parameter ϕ_c (as users in the same group share the same model padaptations). This can be understood as the mutual interactions within groups to form group norms and attitudes. Second, the group-wise observations are also utilized to update the globally shared model adaptations among all the users (as shown in Eq (8)), which adds another dimension of task relatedness for multi-task learning. Also as illustrated in Eq (7) and (8), when different feature groupings are used in $g(\cdot)$ and $g'(\cdot)$, nonlinearity is introduced to propogate information across features.

• **Predict for $\boldsymbol{y}^u$:** During the t-th iteration of stochastic EM, we use the newly inferred group membership and sentiment models to predict the sentiment labels $\boldsymbol{y^u}$ in user u's testing documents by,

$$P(y_d^u = 1|x_d^u, \{\phi_c^t\}_{c=1}^{C_t}, \theta_t^s, \omega^0) = \tag{9}$$
$$\sum_{c=1}^{C_t} P(c_u^t = c) P(y_d^u = 1|x_d^u, \phi_{c_u^t}^t, \theta_t^s, \omega^0)$$

where $\big(\{\phi_c^t\}_{c=1}^{C_t}, c_u^t, \theta_t^s\big)$ are the estimates of latent variables at the tth iteration, $P(c_u^t = c)$ is estimated in Eq (5) and $P(y_d^u = 1|x_d^u, \phi_{c_u^t}, \theta^s, \omega^0)$ is computed by Eq (1). Then the posterior of $\boldsymbol{y}^u$ can thus be estimated via empirical expectation after T iterations,

$$P(y_d^u = 1|x_d^u, \omega^0, \alpha, G_0) = \frac{1}{T} \sum_{t=1}^T P(y_d^u = 1|x_d^u, \{\phi_c^t\}_{c=1}^{C_t}, \theta_t^s, \omega^0)$$

To avoid auto-correlation in the Gibbs sampling chain, samples in the burn-in period are discarded and proper thinning of the sampling chain is performed in our experiments.

4. EXPERIMENTS AND DISCUSSIONS

We performed empirical evaluations to validate the effectiveness of our proposed personalized sentiment classification algorithm. Extensive quantitative comparisons on two large-scale opinionated review datasets collected from Amazon and Yelp confirmed the effectiveness of our algorithm against several state-of-the-art model adaptation and multi-task learning algorithms. Our qualitative studies also demonstrated the automatically identified user groups recognized the diverse use of vocabulary across different users.

4.1 Experimental Setup

• **Datasets.** We used two publicly available review datasets, Amazon [22] and Yelp[1], for our evaluation purpose. In these two datasets, each review is associated with various attributes such as author ID, review ID, timestamp, textual content, and an opinion rating in a discrete five-star range. Specifically, the Amazon dataset is extremely sparse: 89.8% reviewers only have one or two reviews and

[1] Yelp dataset challenge. http://www.yelp.com/dataset_challenge

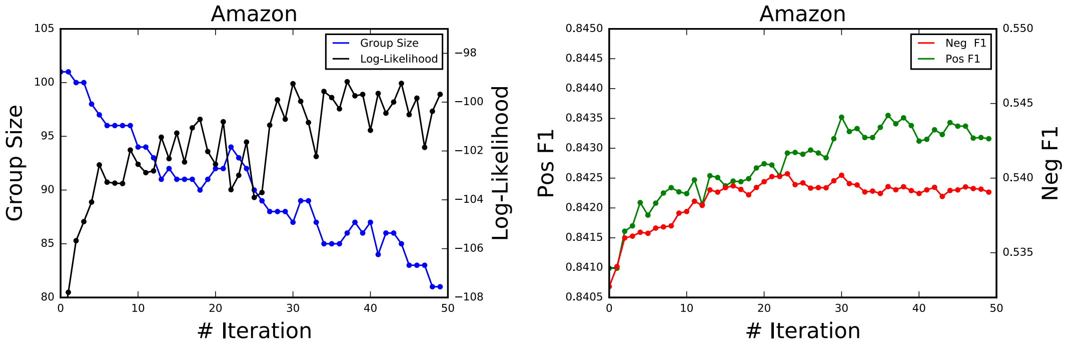

Figure 2: Trace of likelihood, group size and performance during iterative posterior sampling in cLinAdapt for Amazon.

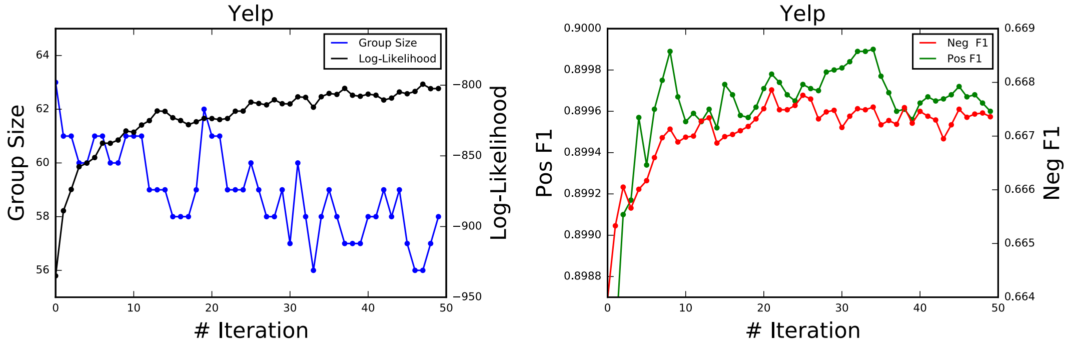

Figure 3: Trace of likelihood, group size and performance during iterative posterior sampling in cLinAdapt for Yelp.

only 0.85% of them have more than 50 reviews. This raises a serious challenge for personalized sentiment analysis.

We performed the following pre-processing steps on both datasets: 1) labeled the reviews with less than 3 stars as negative, and those with more than 3 stars as positive; 2) excluded reviewers who posted more than 1,000 reviews and those whose positive or negative review proportion is greater than 90% (little variance in their opinions and thus easy to classify); 3) ordered each user's reviews with respect to their timestamps. We then constructed feature vector for each review with both unigrams and bigrams after stemming, and performed feature selection by taking the union of top features ranked by Chi-square and information gain metrics [40]. The final controlled vocabulary consists of 5,000 and 3,071 text features for Amazon and Yelp datasets respectively; and we adopted TF-IDF as the feature weighting scheme. From the resulting datasets, we randomly sampled 9,760 Amazon reviewers and 10,830 Yelp reviewers for evaluation purpose. There are 105,472 positive and 37,674 negative reviews in the selected Amazon dataset; 157,072 positive and 51,539 negative reviews in the selected Yelp dataset.

• **Baselines.** We compared the proposed cLinAdapt algorithm with nine baselines, covering several state-of-the-art model adaptation and multi-task learning algorithms. Below we briefly introduce each one of them and discuss their relationship with our algorithm.

1) **Base**: In order to perform the proposed clustered model adaptation, we need a user-independent classification model to serve as the prior model (i.e., ω^0 in Eq (1)). We randomly selected a subset of 2,500 users outside the previously reserved evaluation dataset in Amazon and Yelp to estimate logistic regression models for this purpose accordingly. 2) **Global SVM**: We trained a global linear SVM classifier by pooling all users' training data together to verify the necessity of personalized classifier training. 3) **Individual SVM**: We estimated an independent SVM classifier for each user based on his/her own training data as a straightforward personalized baseline. 4) **LinAdapt**: This is a linear transformation based model adaptation solution for personalized sentiment classification proposed in [1]. 5) **LinAdapt+kMeans**: To verify the effectiveness of our proposed user grouping method in personalized sentiment model learning, we followed [5] to first perform k-means clustering of users based on their training documents, and then estimated a shared LinAdapt model in each identified user group. 6) **LinAdapt+DP**: We also introduced DP prior to LinAdapt to perform joint user grouping and model adaptation training. Because LinAdapt directly adapts from the predefined Base model, *no* information is shared across user groups. 7) **RegLR+DP**: It is an extension of regularized logistic regression for model adaptation [15] with the introduction of DP prior for automated user grouping. In this model, a new logistic regression model will be estimated in each group with the predefined Base model as prior. As a result, this baseline is essentially the same algorithm as that in [39]. 8) **MT-SVM**: It is a state-of-the-art multi-task learning solution proposed in [10]. It encodes the task relatedness via a shared linear kernel across tasks. Comparing to our learning scheme, it only estimates shifting operation in each user without user grouping nor feature grouping. 9) **MT-RegLR+DP**: This baseline identifies groups of similar tasks that should be learnt jointly while the extend of similarity among different tasks are learned via a Dirichlet process prior. Instead of estimating individual group models from the Base model in RegLR+DP independently, the same task decomposition used in MT-SVM is introduced. As a result, the learning tasks will be decomposed to group-wise model learning and global model learning. But it estimates a full set of model parameters of size V in each individual task and global task, such that it requires potentially more training data.

• **Evaluation Settings.** In our experiment, we split each user's review data into two parts: the first half for training and the rest for testing. As we introduced in Section 3.3 and 3.4, the concentration parameter α in DP together with the the number of auxiliary variables m in sampling of $\{c_u\}_{u=1}^N$ play an important role in determining the number of latent user groups in all DP-based models. We empirically fixed $\alpha = 1.0$ and $m = 6$ in all such models. Due to the biased class distribution in both datasets, we compute F1 measure for both positive and negative class in each user, and take macro average among users to compare the different models' classification performance.

4.2 Feasibility of Automated User Grouping

First of all, it is important to verify our stochastic EM based posterior inference in cLinAdapt is converging, as only one sample was taken from the posterior of $\{c_u\}_{u=1}^N$ when updating the group sentiment models $\{\phi_c\}_{c=1}^\infty$ and global model θ^s. We traced the complete-data log-likelihood, the number of inferred latent user groups, together with the testing performance (by Eq (9)) during each iteration of posterior inference in cLinAdapt over all users from both datasets. We reported the results for the two datasets in Figure 2 and 3, where for visualization purpose the illustrated results were collected in every five iterations (i.e., thinning the sampling chain) after the burn-in period (the first ten iterations).

As observed from the results on both datasets, the likelihood kept increasing during the iterative posterior sampling process and converged later on. In the meanwhile, the group size fluctuated a lot at the beginning of sampling and became more stable near the end of iterations. On the other hand, the classification performance on the testing collection kept improving as more accurate sentiment models were estimated from the iterative sampling process. This verifies the effectiveness of our posterior inference procedure. We also looked into the automatically identified groups and found many of them exhibited unique characteristics. The median number of reviews per user in these two datasets were only 7 and 8, while in some groups the average number of reviews per user is as large as 22.1, with small variances. This indicates active users were grouped together in cLinAdapt. In addition, the overall positive class ratio on these two datasets is 74.7% and 75.3% respectively, but in many identified groups the class distribution was extremely biased: some towards negative, as low as 62.1% positive; and some towards positive, as high as 88.2% (note users with more than 90% positive or negative reviews have been removed). This suggests users with similar opinions were also successfully grouped in cLinAdapt. In addition, small fluctuation in the number of sampled user groups near the end of iterations is caused by a small number of users keeping switching groups (as new groups were created for them). This is expected and reasonable, since the group assignment is modeled as a random variable and multiple latent user groups might fit a user's opinionated data equally well. This provides us the flexility to capture the variance in different users' opinions.

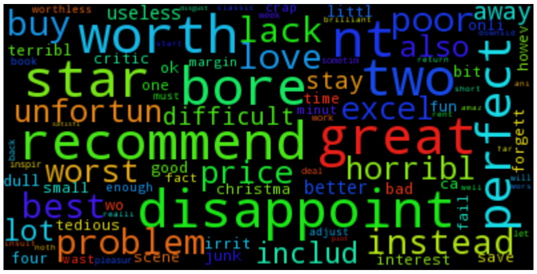

Figure 4: Word clouds on Amazon.

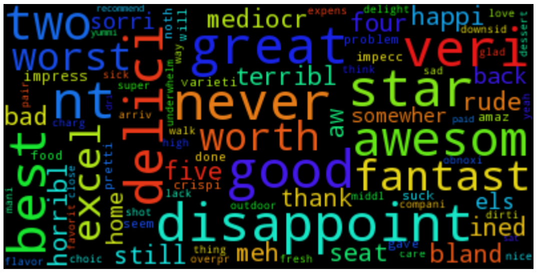

Figure 5: Word clouds on Yelp.

Table 1: Effect of different feature groupings in cLinAdapt.

Method	Amazon		Yelp	
	Pos F1	Neg F1	Pos F1	Neg F1
Base	0.8092	0.4871	0.8809	0.6284
400-1600	0.8313	0.5033	0.8942	0.6563
400-all	0.8405	0.5213	0.8981	0.6632
800-1600	0.8325	0.5115	0.8959	0.6592
800-all	0.8437	**0.5478**	**0.9010**	**0.6694**
1600-all	**0.8440**	0.5334	0.8993	0.6674
all-all	0.8404	0.5391	0.8995	0.6681

In addition to the above quantitative measures, we also looked into the learnt word sentiment polarities reflected in each group's sentiment classifier to further investigate the automatically identified user groups. Most of the learnt feature weights followed our expectation of the words' sentiment polarities, and many words indeed exhibited distinct polarities across groups. We visualized the variance of learnt feature weights across all the groups using word clouds and demonstrated the top 10 words with largest variance and top 10 words with smallest variance in Figure 4 and 5 for Amazon and Yelp datasets respectively. Considering the automatically identified groups were associated with different number of users, we normalized the group feature weight vector by its $L2$ norm. The displayed size of the selected features in the word cloud is proportional to their variances. From the results we can find that, for example, the words "*bore*, *lack*, *worth*" conveyed quite different sentiment polarities among diverse latent user groups in Amazon dataset, while the words like "*pleasure, deal, fail*" had quite consistent polarities. This is also observed in the Yelp dataset, as we can find words like "*star*, *good*, *worth*" were used quite differently across groups, while the words like "*horribl*, *sick*, *love*" are used more consistently.

4.3 Effect of Feature Grouping

We then investigated the effect of feature grouping in cLinAdapt. As discussed in Section 3.3, different feature groupings can be applied to the individual models and global model, such that nonlinearity is introduced when different grouping functions are used in these two levels of model adaptation.

We adopted the most effective feature grouping method named "*cross*" from [35]. Following their design, we first evenly spilt the hold-out training set (for Base model training) into N non-overlapping folds, and estimated a single SVM model on each fold. Then, we created a $V \times N$ matrix by collecting the learned SVM weights from the N folds, on which k-means clustering was applied to group V features into K and L feature groups. We compared the performance of varied combinations of feature groups for individual and global models in cLinAdapt. The experiment results are demonstrated in Table 1; and for comparison purpose, we also included the base classifier's performance in the table. In Table 1, the first column indicates the feature group sizes in the personalized models and global model respectively. And *all* indicates one feature per group (i.e., no feature grouping). All adapted models in cLinAdapt achieved promising performance improvement against the Base model. In addition, further improved performance in cLinAdapt's was achieved when we increased the feature group size in the global model. Under a fixed feature group size in the global model, a moderate size of feature groups in personalized models was more advantageous.

These observations follow our expectation. Since the global model is shared across all users, the whole collection of training data can be leveraged to adapt the global model to overcome sparsity. This allows cLinAdapt to afford more feature groups in the global model, and leads to a more accurate model adaptation. But at the group level, data sparsity remains as the major bottleneck for accurate estimation of model parameters, although observations have already been shared in groups. Hence, the trade-off between observation sharing among features and estimation accuracy has to be made. Based on this analysis, we selected the combination of **800-all** feature grouping methods in the following experiments.

4.4 Personalized Sentiment Classification

We compared cLinAdapt against all nine baselines on both Amazon and Yelp datasets, and the detailed performance is reported in Table 2. Overall, cLinAdapt achieved the best performance against all baselines, except the prediction of positive class in Amazon dataset. Considering these two datasets are heavily biased towards positive class, improving the prediction accuracy in negative class is arguably more challenging and important.

It is meaningful to compare different algorithms' performance according to their model assumptions. First, as the Base model was trained on an isolated collection, though from the same domain, it failed to capture individual users' opinions. Global SVM benefited from gathering large collection of data from the targeted user population but was short of personalization, thus it performed well on positive class while suffered in negative class. Individual SVM could not capture each user's own sentiment model due to serious data sparsity issue; and it was the worst solution for personalized sentiment classification.

Second, as a state-of-the-art model adaptation based baseline, LinAdapt slightly improved over the Base model; but as the user models were trained independently, its performance was limited by the sparse observations in each individual user. The arbitrary user grouping by k-means barely helped LinAdapt in personalized classification, though more observations became available for model training. The joint user grouping with LinAdapt training finally achieved substantial performance improvement (especially on the Yelp dataset). Similar result was achieved in RegLR+DP as well. This confirms the necessity of joint task relatedness estimation and model training in multi-task learning.

Third, global information sharing is essential. All methods with a jointly estimated global model, i.e., MT-SVM, MT-RegLR+DP, cLinAdapt and also Global SVM, achieved significant improvement over others that do not have such a globally shared component. Additionally, as the class prior was against negative class in both datasets, observations of negative class became even rare in each user. As a result, compared with MT-SVM and MT-RegLR+DP baselines, cLinAdapt achieved improved performance in this class by sharing observations across features via its unique two-level feature grouping mechanism. However, comparing to MT-SVM, although no user grouping nor feature grouping was performed, its performance was very competitive. We hypothesized it was because on both datasets we had overly sufficient training signals for the globally shared model in MT-SVM. To verify this hypothesis, we reduced the number of users in the evaluation data set when training MT-SVM and cLinAdapt. Both models' performance decreased, but cLinAdapt decreased much slower than MT-SVM. When we only had five thousand users, cLinAdapt significantly outperformed MT-SVM in both classes on these two evaluation datasets. This result verifies our hypothesis and demonstrates the distinct advantage of cLinAdapt: when the total number of users (i.e., inductive learning tasks) is limited, properly grouping the users and leveraging information from a pre-trained model help improve overall classification performance.

Table 2: Personalized sentiment classification results.

Method	Amazon		Yelp	
	Pos F1	Neg F1	Pos F1	Neg F1
Base	0.8092	0.4871	0.8809	0.6284
Global SVM	0.8386	0.5245	0.8982	0.6596
Individual SVM	0.5582	0.2418	0.5691	0.3492
LinAdapt	0.8091	0.4894	0.8811	0.6281
LinAdapt+kMeans	0.8096	0.4990	0.8836	0.6461
LinAdapt+DP	0.8157	0.4721	0.8878	0.6391
RegLR+DP	0.8256	0.5021	0.8929	0.6528
MT-SVM	**0.8484**	0.5367	0.9002	0.6663
MT-RegLR+DP	0.8466	0.5247	0.8998	0.6630
cLinAdapt	0.8437	**0.5478**	**0.9010**	**0.6694**
Oracle-cLinAdapt	0.9049	0.6791	0.9268	0.7358

One limitation of cLinAdapt is that the latent group membership can only be inferred for users with at least one labeled training instance. This limits its application in cases where new users keep emerging for analysis. This difficulty is also known as cold-start, which concerns the issue that a system cannot draw any inferences for users about which it has not yet gathered sufficient information. One remedy is to acquire a few labeled instances from the testing users for cLinAdapt model update. But it would be prohibitively expensive if we do so for every testing user. Instead, we decide to only infer the group membership for the new users based on their disclosed labeled instances, while keep the previously trained cLinAdapt model intact (i.e., perform sampling defined in Eq (5) without changing the group structure). This implicitly assumes the previously identified user groups are comprehensive and the new users can be fully characterized by one of those groups.

In order to verify this testing scheme, we randomly selected 2,000 users with at least 4 reviews to create hold-out testing sets on both Amazon and Yelp reviews accordingly, and used the rest users to estimate the cLinAdapt model. During testing in each user, we held the first three reviews' labels as known, and gradually disclosed them to cLinAdapt to infer this user's group membership and classify in the rest reviews. For comparison purpose, we also included Individual SVM, LinAdapt and MT-SVM trained and tested in the same way on these two newly collected evaluation datasets for cold-start, and reported the results in Table 3. From the results, it is clear that Individual SVM's performance was almost random due to the limited amount of training data in this testing scenario. LinAdapt benefited from a predefined Base model, while the independent model adaptation in single users still led to suboptimal performance. The same reason also limited MT-SVM: it treats users independently by only sharing the global model among them, so that the newly available labeled instances could not effectively help individual models at beginning. cLinAdapt better handled cold-start by reusing the learned user groups for new users. Significant improvement was achieved for negative class, as the observations in negative class were even more scarce in those newly disclosed labeled instances of each testing user.

Table 3: Effectiveness of model sharing for cold-start on Amazon and Yelp.

Obs.	Amazon								Yelp							
	Individual SVM		LinAdapt		MT-SVM		cLinAdapt		Individual SVM		LinAdapt		MT-SVM		cLinAdapt	
	Pos F1	Neg F1	Pos F1	Neg F1	Pos F1	Neg F1	Pos F1	Neg F1	Pos F1	Neg F1	Pos F1	Neg F1	Pos F1	Neg F1	Pos F1	Neg F1
1^{st}	0.0000	0.4203	0.8587	0.5898	0.8588	0.5073	**0.8925**	**0.6675**	0.0000	0.4101	0.9322	0.7724	0.9251	0.7285	**0.9582**	**0.8335**
2^{nd}	0.4683	0.3831	0.8455	0.5495	0.8534	0.5267	**0.8795**	**0.6076**	0.7402	0.3116	0.9243	0.7176	0.9291	0.7027	**0.9501**	**0.7726**
3^{rd}	0.7362	0.1751	0.8113	0.4863	0.8283	0.4919	**0.8440**	**0.5402**	0.7812	0.1608	0.8873	0.6639	0.8954	0.6619	**0.9116**	**0.7147**

Table 4: Collaborative filtering results on Amazon and Yelp.

Models	Amazon		Yelp	
	NDCG	MAP	NDCG	MAP
Average	0.7758	0.5587	0.6798	0.3867
LinAdapt	0.8046	0.6640	0.7445	0.4945
LinAdapt+kMeans	0.8030	0.6635	0.7399	0.4901
LinAdapt+DP	0.8004	0.6597	0.7454	0.4986
RegLR+DP	0.8023	0.6614	0.7460	0.4991
MT-SVM	0.8050	0.6646	0.7439	0.4935
MT-RegLR+DP	0.8030	0.6626	0.7419	0.4935
cLinAdapt	**0.8052**	**0.6652**	**0.7473***	**0.5001**

$^*p$-value<0.05 under binomal test

Another observation in Table 3 is that all models' testing performance decreased with more labeled instances disclosed from the testing users. This is unexpected and might indicate the consistence assumption about a user's sentiment model does not hold. To verify this, we tested an oracle setting of cLinAdapt in the original evaluation set: we revealed the labels of testing data when inferring group assignments in testing, and this greatly boosted the test performance of cLinAdapt. We appended the result in Table 2. This indicates the performance bottleneck of cLinAdapt is the accuracy of inferred group membership in testing phase. We assumed this membership is stationary in each user, but this might not be true given the reviews were generated in a chronological order and users' sentiment model might change over time. In our future work, we plan to also model the generation of document content in cLinAdapt, such that the inferred group membership can be calibrated for each testing document accordingly.

4.5 Serve for Collaborative Filtering

Collaborative filtering technique has been successfully applied in many recommendation systems. One of its key components is to infer the similarity between users. The learnt personalized sentiment model for each user naturally serves as a good proxy of their preference; and the distance between the model weights can therefore characterize the similarity between users. In this experiment, we evaluated the utility of learnt personalized models in collaborative filtering based recommendation. To create an evaluation data set, besides the items that a user has reviewed, we randomly selected a set of items from other users and label them as irrelevant in recommendation evaluation. We fixed this random item set to be four times large as a user's actually reviewed item size and maintained the same random candidate items in all the algorithms. In addition, we also removed the items that were only reviewed by one user. For each candidate item, we selected the target user's top K most similar neighbors who also reviewed this item, and calculated the weighted average of neighbors' actual ratings as ranking score for this item. Normalized discounted cumulative gain (NDCG) and mean average precision (MAP) are used to measure the recommendation quality. In particular, NDCG takes the user's original five star rating as a multi-scale relevance judgment, and MAP takes reviews with higher than 3 star as relevant and the rest as irrelevant.

We compared the recommendation performance based on the user similarity computed by different personalized sentiment classification methods on both Amazon and Yelp datasets. We also included a baseline that makes recommendations by the simple average of ratings from all the users who reviewed the item, and named it as Average. The average number of users who reviewed the same item in Amazon is 3.2 and in Yelp it is 10. Correspondingly, we selected top-4 neighbors and top-8 neighbors in Amazon and Yelp datasets respectively based on the cosine similarity between the users' personalized models. We report the resulting MAP and NDCG performance across all users in Table 4. As we can find from the Table 4, cLinAdapt achieved encouraging recommendation performance on both datasets, which indicates its learned sentiment models better captured the relatedness among users in their preferences over the recommended items. Despite the very sparse distribution of reviews in both datasets, cLinAdapt correctly recognized the preference of different users, and found the best neighbors for collaborative filtering.

5. CONCLUSION AND FUTURE WORK

In this paper, we developed a clustered model adaptation solution for personalized sentiment classification. Our work is inspired by the well-established social theories about humans' dispositional tendencies, i.e., social comparison and cognitive consistence. By exploiting the clustering property of users' sentiment models, empirically improved sentiment classification performance was achieved on two large collections of opinionated review documents.

Several areas are left open for our future explorations. In the current work, we assumed a user's latent group membership is stationary: once inferred from training data, it could be repeatedly used in testing. However, a user's group membership and even sentiment model might evolve over time. It is beneficial to efficiently update the model when new labeled data and users become available. Also, the current model is unable to inference group memberships over users with no labeled instances. This could be overcome if the generative model also accounts for the generation of review content in each user. In addition, it is interesting to study how to identify the feature grouping together with the user groups, such that the balance can be automatically adjusted with respect to the available training data in each latent user group.

6. ACKNOWLEDGMENTS

We thank the anonymous reviewers for their insightful comments. This paper is based upon work supported by the National Science Foundation under grant IIS-1553568.

7. REFERENCES

[1] Mohammad Al Boni, Keira Qi Zhou, Hongning Wang, and Matthew S Gerber. Model adaptation for personalized opinion analysis. *In Proceedings of ACL*, 2015.

[2] Charles E Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics*, pages 1152–1174, 1974.

[3] Bart Bakker and Tom Heskes. Task clustering and gating for bayesian multitask learning. *The Journal of Machine Learning Research*, 4:83–99, 2003.

[4] Sigal G Barsäde and Donald E Gibson. Group emotion: A view from top and bottom. *Research on managing groups and teams*, 1:81–102, 1998.

[5] Jiang Bian, Xin Li, Fan Li, Zhaohui Zheng, and Hongyuan Zha. Ranking specialization for web search: a divide-and-conquer approach by using topical ranksvm. In *Proceedings of the 19th WWW*, pages 131–140. ACM, 2010.

[6] John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 EMNLP*, pages 120–128. ACL, 2006.

[7] John Bruhn. The concept of social cohesion. In *The Group Effect*, pages 31–48. Springer, 2009.

[8] Meghana Deodhar and Joydeep Ghosh. Scoal: A framework for simultaneous co-clustering and learning from complex data. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(3):11, 2010.

[9] Jean Diebolt and Eddie HS Ip. Stochastic em: method and application. In *Markov chain Monte Carlo in practice*, pages 259–273. Springer, 1996.

[10] Theodoros Evgeniou and Massimiliano Pontil. Regularized multi–task learning. In *Proceedings of the 10th ACM SIGKDD*, pages 109–117. ACM, 2004.

[11] Theodoros Evgeniou, Massimiliano Pontil, and Olivier Toubia. A convex optimization approach to modeling consumer heterogeneity in conjoint estimation. *Marketing Science*, 26(6):805–818, 2007.

[12] Thomas S Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

[13] Leon Festinger. A theory of social comparison processes. *Human relations*, 7(2):117–140, 1954.

[14] Leon Festinger. *A theory of cognitive dissonance*, volume 2. Stanford university press, 1962.

[15] Bo Geng, Yichen Yang, Chao Xu, and Xian-Sheng Hua. Ranking model adaptation for domain-specific search. *TKDE*, 24(4):745–758, 2012.

[16] Giorgos Giannopoulos, Ulf Brefeld, Theodore Dalamagas, and Timos Sellis. Learning to rank user intent. In *Proceedings of the 20th CIKM*, pages 195–200. ACM, 2011.

[17] Lin Gong, Mohammad Al Boni, and Hongning Wang. Modeling social norms evolution for personalized sentiment classification.

[18] Xia Hu, Lei Tang, Jiliang Tang, and Huan Liu. Exploiting social relations for sentiment analysis in microblogging. In *Proceedings of the 6th WSDM*, pages 537–546. ACM, 2013.

[19] Laurent Jacob, Jean-philippe Vert, and Francis R Bach. Clustered multi-task learning: A convex formulation. In *NIPS*, pages 745–752, 2009.

[20] Guangxia Li, Steven CH Hoi, Kuiyu Chang, and Ramesh Jain. Micro-blogging sentiment detection by collaborative online learning. In *ICDM*, pages 893–898. IEEE, 2010.

[21] Yucheng Low, Deepak Agarwal, and Alexander J Smola. Multiple domain user personalization. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 123–131. ACM, 2011.

[22] Julian McAuley, Rahul Pandey, and Jure Leskovec. Inferring networks of substitutable and complementary products. In *Proceedings of the 21th ACM SIGKDD*, pages 785–794. ACM, 2015.

[23] Brian Mullen and George R Goethals. *Theories of group behavior*. Springer Science & Business Media, 2012.

[24] Radford M Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

[25] Theodore M Newcomb. *The acquaintance process.* Holt, Rinehart & Winston, 1961.

[26] Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th WWW*, pages 751–760. ACM, 2010.

[27] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359, 2010.

[28] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd ACL*, pages 115–124. ACL, 2005.

[29] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pages 79–86, 2002.

[30] Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[31] Babak Shahbaba and Radford Neal. Nonlinear models using dirichlet process mixtures. *Journal of Machine Learning Research*, 10(Aug):1829–1850, 2009.

[32] Muzafer Sherif. *The psychology of social norms.* Harper, 1936.

[33] Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li. User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD*, pages 1397–1405. ACM, 2011.

[34] Sebastian Thrun and Joseph O'Sullivan. Discovering structure in multiple learning tasks: The tc algorithm. In *ICML*, volume 96, pages 489–497, 1996.

[35] Hongning Wang, Xiaodong He, Ming-Wei Chang, Yang Song, Ryen W White, and Wei Chu. Personalized ranking model adaptation for web search. In *Proceedings of the 36th ACM SIGIR*, pages 323–332. ACM, 2013.

[36] Hongning Wang, Yue Lu, and ChengXiang Zhai. Latent aspect rating analysis without aspect keyword supervision. In *Proceedings of the 17th ACM SIGKDD*, pages 618–626. ACM, 2011.

[37] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210, 2005.

[38] Fangzhao Wu and Yongfeng Huang. Sentiment domain adaptation with multiple sources. In *Proceedings of the 54th Annual Meeting on Association for Computational Linguistics*, pages 301–310, 2016.

[39] Ya Xue, Xuejun Liao, Lawrence Carin, and Balaji Krishnapuram. Multi-task learning for classification with dirichlet process priors. *Journal of Machine Learning Research*, 8(Jan):35–63, 2007.

[40] Yiming Yang and Jan O Pedersen. A comparative study on feature selection in text categorization. In *ICML*, volume 97, pages 412–420, 1997.