When Sentiment Analysis Meets Social Network: A Holistic User Behavior Modeling in Opinionated Data

Lin Gong, Hongning Wang
Department of Computer Science, University of Virginia
{lg5bt,hw5x}@virginia.edu

ABSTRACT

User modeling is critical for understanding user intents, while it is also challenging as user intents are so diverse and not directly observable. Most existing works exploit specific types of behavior signals for user modeling, e.g., opinionated data or network structure; but the dependency among different types of user-generated data is neglected.

We focus on self-consistence across multiple modalities of user-generated data to model user intents. A probabilistic generative model is developed to integrate two companion learning tasks of *opinionated content modeling* and *social network structure modeling* for users. Individual users are modeled as a mixture over the instances of paired learning tasks to realize their behavior heterogeneity, and the tasks are clustered by sharing a global prior distribution to capture the homogeneity among users. Extensive experimental evaluations on large collections of Amazon and Yelp reviews with social network structures confirm the effectiveness of the proposed solution. The learned user models are interpretable and predictive: they enable more accurate sentiment classification and item/friend recommendations than the corresponding baselines that only model a singular type of user behaviors.

CCS CONCEPTS

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

KEYWORDS

User behavior modeling, sentiment analysis, social network

ACM Reference Format:
Lin Gong, Hongning Wang. 2018. When Sentiment Analysis Meets Social Network: A Holistic User Behavior Modeling in Opinionated Data. In *Proceedings of The 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining (KDD '18).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3219819.3220120

1 INTRODUCTION

User modeling is essential for understanding users' diverse preferences and intents, which in turn provides valuable insights for online service systems to adaptively maximize their service utility in a per-user basis [28, 36]. Numerous successes have proved its value in practical applications. For example, Yan et al. [38] reported that Click-Through Rate (CTR) of an ad can be averagely improved as high as 670% by properly segmenting users for behavioral targeted advertising in a sponsored search; and Zhang et al. [40] found that modeling users' review contents for explainable recommendation improved CTR by more than 34.7% and conversion rate by more than 25.5% in an online e-commence website.

User modeling is also challenging, as humans are self-interested actors with diverse decision making autonomy. Their intents are distinctive while not directly observable from the systems. Various behavior signals have been explored, with different focuses in exploiting information about users' intents. A large body of efforts explore opinionated text data to understand users' emphasis on specific entities or aspects [13, 26]. The distribution of words and their sentiment polarities are modeled in a statistical way to decipher the embedded user intents. System logged behavior data, such as opinion ratings and result clicks, provide direct supervision to infer users' latent preferences over the systems' outputs [28] or their decision making process [36, 40]. In parallel, social network structure among users has been proved to be useful in examining users' interactive behaviors. The proximity between a pair of users has been studied to understand social influence and information diffusion [19], and network structure has been analyzed to examine users' social grouping and belonging [1, 18, 23].

However, most existing solutions restrict the analysis within a specific modality of user behaviors, and fail to realize the dependency among these different types of user-generated data, which are essentially governed by the same intents in each user. We argue that in order to accurately and comprehensively understand users, user modeling should consist of multiple companion learning tasks focusing on different modalities of user-generated data, such that the observed behaviors (e.g., opinion ratings or social connections) can be mutually explained by the associated models. Our argument is also supported by the *Self Consistency Theory* [17] in social psychology studies, as it asserts that consistency of ideas and representation of the self are integral in humans.

In this work, we focus on user modeling in social media data, where users generate opinionated textual contents to express their opinions on various topics, and connect to others to form social network. It is therefore an ideal platform for collecting various types of user behavior data. We model distinct behavior patterns of *individual users* by taking a holistic view of sentiment analysis and social network analysis. In particular, we develop a probabilistic generative model to integrate two complementary tasks of *opinionated content modeling* for recognizing user preferences and *social network structure modeling* for understanding user relatedness, i.e.,

KDD '18, August 19–23, 2018, London, United Kingdom

ACM ISBN 978-1-4503-5552-0/18/08…$15.00
https://doi.org/10.1145/3219819.3220120

a multi-task learning approach [6, 11, 37]. In the first task, a statistical language model is used to model the generation of textual contents, and a logistic regression model maps the textual contents to the sentiment polarity. In the second task, a stochastic block model [12] is employed to capture the relatedness among users. To realize the diversity across individual users, we assume there are multiple instances of both learning tasks in a population of users, and different users are associated with different instances of them. And to encode our consistency assumption about user behaviors, we further assume an instance of opinionated content modeling task is always coupled with an instance of social network structure modeling task. For example, users who prefer history books tend to connect to those who like memoirs but not those who like makeup. Such pairing hence represents the shared user intents.

The problem of user modeling is thus formulated as assigning users to those instances of paired learning tasks, which best explain a particular user's observed behaviors in both modalities. To capture behavior heterogeneity of each individual user, such as a user might be in favor of both history and science fiction books, we model a user as a *mixture* over those instances of paired learning tasks. And to reflect the homogeneity across users, i.e., different users might share the same intent, we impose a globally shared Dirichlet Process (DP) prior [22] over the instances of paired learning tasks. The clustering property of DP is beneficial as draws from it often share some common values and therefore naturally form clusters. Thus, we do not need to specify the number of unique instances beforehand and we use a data-driven approach to explore the possible setting of potentially infinite number of instances.

We refer to the unique instance of paired tasks as a *collective identity* in this paper, as it characterizes the behavior norms in a collection of user-generated data [35]. To accommodate the variable number of collective identities that a user can associate with, we impose another DP prior over the mixing proportion of collective identities in each user, i.e., a hierarchical Dirichlet Process (HDP) [33] structure. Accordingly, we refer to this user-specific mixing proportion as his/her *personal identity*. This design is also supported by the social psychology theories about human's formation and evolution of behaviors. In particular, *Self-Categorization Theory* [24] asserts that human beings are able to act at individual level (i.e., their personal identity) and social group level (i.e., their collective identity). Overall, the objective of model learning is thus to infer the posterior distribution of those shared learning tasks and the belonging of each user to those tasks in a given population of users.

To investigate the effectiveness of the proposed model for user modeling, we performed extensive experiments on two different sets of user reviews collected from Amazon and Yelp, together with the social network structures. The results clearly demonstrate the advantages of the proposed solution: the learned user models are interpretable and unveil dominant behavior patterns across users; they also introduce improved predictive power which is verified by the improved performance in a diverse set of applications, such as sentiment classification, collaborative filtering based item recommendation and friend recommendation, compared with the state-of-the-art solutions in each of these problems.

2 RELATED WORK

A great deal of efforts have been devoted in the exploration of opinion-rich textual contents to understand users' decision making process [20, 25, 26]. Earlier works utilized both lexicon-based [13] and learning-based [25, 26] solutions to classify sentiment polarity of text data. Later finer-grained models were developed based on topic modeling techniques to predict users' detailed aspect-level opinions and preferences [34, 36]. The development in modeling user-generated text data directly enables personalized recommendation and retrieval. Zhang et al. [40] combined phrase-level sentiment analysis with matrix factorization for explainable recommendation. Ghose et al. [9] illustrated how user-generated content can be mined and incorporated into a demand estimation model to generate a new ranking system in product search engines.

In a parallel line of research, considerable efforts are made in utilizing social network structure for user modeling. Huo et al. [16] calculated the linking probability between a pair of users by incorporating social influence which is considered based on their activities. Community structure in social networks has been investigated [1, 39], in which users are clustered to reflect their social belongings and interactions. Leskovec et al. [18] studied signed networks in social media, with a focus on how the interplay between positive and negative relationships affects the structure of online social networks. Network embedding techniques [2, 32] have been developed to project a social network onto a continuous space to facilitate the study of affinity and grouping among users.

There are also works which combine opinionated text modeling with social network modeling to improve the fidelity of learned user models. Speriosu et al. [29] proposed to propagate labels from a supervised classifier over the Twitter follower graph to improve sentiment classification. Studies in [14, 30] incorporated user-user interactions as side information to regularize sentiment classification. Cheng et al. [3] leveraged signed social network to infer the sentiment of text documents in an unsupervised manner. Tang et al. [31] proposed to propagate emotional signals and text-based classification results via different relations in a social network, such as word-microblog relations, microblog-microblog relations. However, all the aforementioned works only treat the network as side information for opinionated text modeling, and they do not explicitly model the dependency between the two modalities of data. Our work proposes a holistic view to unify the modeling of opinionated text data and social network data, thus to understand user intents from multiple complementary perspectives.

We view the modeling of different modalities of user behaviors as a multi-task learning problem [6, 11, 37], which exploits the relatedness between tasks to mutually reinforce each other. Similar modeling approaches have been explored in user modeling before. Fei et al. [7] used multi-task learning to predict users' response (e.g., comment or like) to their friends' postings regarding the message content, where each user is modeled as a task and task relation is defined by content similarity between users. Evgeniou and Pontil [5] imposed task relatedness by constructing a common underlying representation across different tasks, e.g., assuming users assess product quality based on a common set of features describing these products. The most related work with ours is [11], in which personalized sentiment classification models are constructed via multi-task

learning. But the work studied sentiment classification problem without modeling the generation of text data nor the associated user interactions. It therefore cannot handle new users. We view the modeling of different types of user-generated data as companion learning tasks, and develop a generative model to integrate these tasks across users.

3 METHODOLOGY

Self-consistence across multiple modalities of user-generated data reflects user intents [17]. To exploit this unique property of user behavior, we develop a probabilistic generative model to integrate two companion learning tasks of *opinionated content modeling* and *social network structure modeling*. Individual users are modeled as a mixture over the instances of paired learning tasks to realize their heterogeneity, and the tasks are clustered by sharing a global prior distribution to capture the homogeneity among users. An efficient stochastic EM algorithm [4] is developed to infer the posterior distribution of task assignments and task structures.

3.1 Problem Definition

We focus on a typical type of social media data, user reviews, in conjunction with the social network among users. Formally, denote a collection of N users as $U = \{u_1, u_2, ...u_N\}$, in which each user u_i is associated with a set of review documents $D_i = \{(\boldsymbol{x}_d^i, y_d^i)\}_{d=1}^{|D_i|}$. Each document d is represented as a V-dimensional feature vector $\boldsymbol{x}_d$, and y_d is the corresponding sentiment label. We assume binary sentiment labels (i.e., +1 for positive and -1 for negative) to simplify the discussion, but the developed algorithm can be easily extended to multi-grade or continuous rating settings. In this collection, each user is connected to a set of other users, referred as friends. For a pair of users u_i and u_j, a binary variable e_{ij} denotes the affinity between them: $e_{ij} = 1$ indicates they are directly connected in the network, i.e., friends, and otherwise $e_{ij} = 0$. For user u_i, we denote the complete set of his/her social connections as $E_i = \{e_{ij}\}_{j\neq i}^N$.

The task of opinionated content modeling is to specify the generation of review contents and sentiment labels in each individual user, i.e., $p(D_i)$. And the task of social network modeling is to specify the generation of friendship relations, i.e., $p(E_i)$. We pair these two learning tasks across users to capture the consistency among different modalities of user behaviors. We assume multiple instances of these paired tasks exist in a collection of users, to reflect behavior heterogeneity across individual users. This is also supported by the *Self-Categorization Theory* as it states that "self-categorization is comparative, inherently variable, fluid and context dependent". As a result, the problem of user modeling is formulated as learning a distribution over these paired tasks in each individual user, i.e., $p(D_i, E_i) = \int p(D_i, E_i|\boldsymbol{\pi}_i)p(\boldsymbol{\pi}_i|u_i)d\boldsymbol{\pi}_i$, where the latent variable $\boldsymbol{\pi}_i$ indicates the distribution of those paired tasks in user u_i, together with estimating the configurations of paired tasks across users.

3.2 A Holistic User Modeling via Multi-Task Learning

We model each individual user as a mixture over the instances of paired learning tasks, so that each of his/her review documents and social connections can be explained by different paired tasks. We refer to each instance of the paired tasks as a collective identity. In a given collection of users, we assume there are C unique collective identities shared across users. When modeling the opinionated content in user u_i, we use an indicator variable z_d^i to denote the assignment of a collective identity to his/her document (x_d^i, y_d^i). We employ a statistical language model to capture the generation of review content, i.e., $p(\boldsymbol{x}_d^i|z_d^i) \sim Multi(\boldsymbol{\psi}_{z_d^i})$, which is a V-dimensional multinomial distribution over the vocabulary. And we use a logistic regression model to map the textual contents to binary sentiment polarities as,

$$p(y_d^i|\boldsymbol{x}_d^i, z_d^i) = \frac{1}{1+\exp(-y_d^i \bar{\boldsymbol{\phi}}_{\boldsymbol{z_d^i}}^\top \boldsymbol{x}_d^i)}, \tag{1}$$

where $\bar{\boldsymbol{\phi}}_{\boldsymbol{z_d^i}}$ is a V-dimensional feature weight vector. Following the setting in [11] to handle data sparsity issue, where the authors suggest to further decompose the feature weight vector $\bar{\boldsymbol{\phi}}_{\boldsymbol{c}}$ into two parts, one global component $\boldsymbol{\phi}_{\boldsymbol{s}}$ shared by all models, and one local component $\boldsymbol{\phi}_{\boldsymbol{c}}$ just for this current model, we further decompose $\bar{\boldsymbol{\phi}}_{\boldsymbol{c}} = \boldsymbol{\phi}_{\boldsymbol{s}} + \boldsymbol{\phi}_{\boldsymbol{c}}$ in our logistic regression model.

Putting these two components together, the task of opinionated content modeling in user u_i is formalized as,

$$p(D_i|\boldsymbol{\pi}_i) = \prod_{d\in D_i} \sum_{z_d^i=1}^{C} p(y_d^i|\boldsymbol{x}_d^i, z_d^i)p(\boldsymbol{x}_d^i|z_d^i)p(z_d^i|\boldsymbol{\pi}_i) \tag{2}$$

where we assume the review documents are independent from each other given the collective identity assignments in user u_i.

Based on the notion of collective identity, we appeal to the stochastic block model [12] to realize the relatedness among users. We assume the connection between a pair of users is determined by the affinity strength between their corresponding collective identities, rather than specifically who they are. For example, history book lovers tend to connect to those who like memoirs. As a result, the observed social connection e_{ij} between user u_i and u_j is modeled as a Bernoulli random variable governed by the corresponding pairwise affinity, i.e., $e_{ij} \sim Bernoulli(B_{z_{i\to j}, z_{j\to i}})$, where B is a $C \times C$ matrix specifying the affinity between any pair of collective identities, and $z_{i\to j}$ and $z_{j\to i}$ denote the collective identities that user u_i and u_j choose when forming this connection. Without loss of generality, we do not assume the social affinity is symmetric. For example, history book lovers tend to connect with memoirs lovers, but it might not be true vice versa. As a result, the task of social network structure modeling in user u_i can be formalized as,

$$p(E_i|\boldsymbol{\pi}_i, z_{j\to i}, B) = \prod_{e_{ij}\in E_i} \sum_{z_{i\to j}=1}^{C} p(e_{ij}|B_{z_{i\to j}, z_{j\to i}})p(z_{i\to j}|\boldsymbol{\pi}_i) \tag{3}$$

again we assume that given the collective identity assignments on the user connections, user u_i's connections with other users are independent from each other.

Based on the above specifications, each collective identity indexed by c can be represented as a homogeneous generative model characterized by a set of parameters $\theta_c = (\boldsymbol{\psi}_c, \boldsymbol{\phi}_c, \boldsymbol{b}_c)$, where $\boldsymbol{\psi}_c$ is the parameter for the multinomial distribution in a language model, $\boldsymbol{\phi}_c$ is the feature weight parameter in a logistic regression model, and $\boldsymbol{b}_c$ is a C-dimensional parameter vector for the Bernoulli distributions specifying affinity between the collective identity c

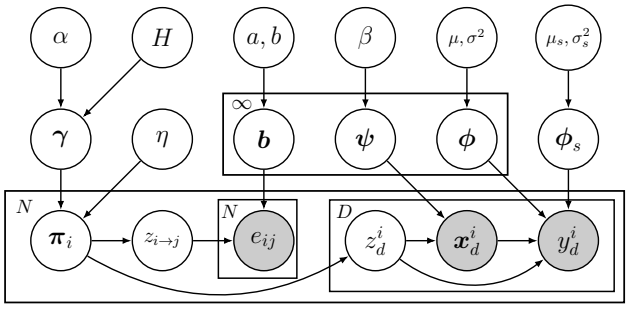

Figure 1: Graphical model representation of HUB. The upper plate indexed by ∞ denotes the unified model parameters for collective identities. The outer plate indexed by N denotes distinct users. The inner plates indexed by N and D denote each user's social connections and review documents respectively.

and all others. The affinity vectors of all the collective identities constitute the aforementioned affinity matrix $B_{C\times C}$. The next step is to specify the generation of the collective identities, such that they best characterize the behavior homogeneity across a collection of users.

Instead of manually selecting the number of collective identities for each given collection of users, we take a data-driven approach to jointly estimate the model structure embedded in the data and the allocation of those learned models in each individual user. In particular, we assume the parameter $\boldsymbol{\theta}_c$ itself is also a random variable drawn from a Dirichlet Process prior [8] with base distribution H and concentration parameter α. Each draw from DP is a discrete distribution consisting of weighted sum of point masses with locations drawn from H. Thus, draws from DP may share common values and form clusters naturally. As a result, the number of unique collective identities will be inferred from data automatically.

As a result, the global distribution of opinionated content and social connections across users follows $DP(H, \alpha)$, which can be described by the following stick-breaking representation:

$$p(D, E) = \sum_{c=1}^{\infty} \gamma_c \delta_{\boldsymbol{\theta}_c}, \tag{4}$$

where δ_{θ_c} is an indicator of the location centered at the sample $\theta_c \sim H$, and $\{\gamma_c\}_{c=1}^{\infty}$ represents the concentration of the unique samples θ_c in the whole collection. The corresponding stick-breaking process for $\boldsymbol{\gamma}$ is defined as: $\gamma_c' \sim Beta(1, \alpha), \gamma_c = \gamma_c' \prod_{t=1}^{c-1}(1 - \gamma_t)$, which is a generalization of multinomial distribution with a countably infinite number of components.

In particular, we impose a Dirichlet distribution, i.e., $Dirichlet(\boldsymbol{\beta})$, as the prior over the language model parameters $\{\boldsymbol{\psi}_c\}_{c=1}^{\infty}$; and an isometric Gaussian distribution $N(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ as the prior for $\{\boldsymbol{\phi}_c\}_{c=1}^{\infty}$ of logistic regression models. A Beta distribution is introduced as the prior over each element of the affinity matrix B, i.e., $b_{ij} \sim Beta(a, b)$. Outside the DP prior structure, we also impose an isometric Gaussian distribution $N(\boldsymbol{\mu}_s, \boldsymbol{\sigma}_s^2)$ over the globally shared logistic regression parameter $\boldsymbol{\phi}_s$.

The global mixture structure defined in Eq (4) is to capture the common user behavior patterns across all users; and the user-level mixture structure is to capture each user's specific characteristics. To afford a mixture over a possibly infinity number of collective identities, we introduce another layer of DP to model the mixture proportion $\boldsymbol{\pi}_i$ in user u_i, which is referred as the personal identity, with the global mixture $\boldsymbol{\gamma}$ as the base distribution and its own concentration parameter η. Another challenge introduced by this possibly infinite number of collective identities resides in the modeling of user social connections via the pairwise affinity between collective identities (i.e., in Eq (3)). As the structure of collective identities becomes unspecified under the DP prior, the affinity relation becomes undefined. Because Beta distribution is conjugate with the pairwise affinity measure matrix B, we can integrate out B without explicitly specifying it. We will provide more details about this special treatment in the later posterior inference discussions.

Putting all the developed components together, we obtain a full generative model describing multiple modalities of user-generated data in a holistic manner. We name the resulting model as **H**olistic **U**ser **B**ehavior model, or HUB in short; we illustrate our imposed dependency between different components of HUB in Figure 1, using a graphical model representation.

3.3 Posterior Inference

Since we formulate the problem of user modeling as assigning users to the instances of paired learning tasks, i.e., collective identity, for a given user u_i, we need to infer the latent collective identity z_d^i that he/she has used in generating the review document $(\boldsymbol{x}_d^i, y_d^i)$, and $z_{i\to j}$ taken by him/her when interacting with user u_j. Based on the inferred collective identities in a collection of users, we can estimate the posterior distributions of model parameters, which collectively specify latent intents of users. In particular, $\boldsymbol{\psi}_c$ characterizes the generation of textual contents under each collective identity; $\boldsymbol{\phi}_c$ and $\boldsymbol{\phi}_s$ capture the mapping from textual contents to sentiment polarities; B represents the affinity among collective identities.

Due to the conjugacy between Beta distribution in our DP prior and the Binomial distribution over the users' social connections, the posterior distribution of $z_{i\to j}$ can be analytically computed; but the lack of conjugate prior for logistic regression makes the exact inference for z_d^i impossible. This also prevents us to perform exact inference on $\boldsymbol{\phi}_c$ and $\boldsymbol{\phi}_s$. As a result, we appeal to a stochastic Expectation Maximization (EM) [4] based iterative algorithm for posterior inference in these three types of latent variables. More specifically, Gibbs Sampling method based on auxiliary variables [22] is utilized to infer the collective identity for each review document possessed by each user, i.e., $\{z_d^i\}_{d=1}^{D}$, and the group membership for each interaction, i.e., $\{z_{i\to j}\}_{j\neq i}^{N}$. This forms the E-step. Then, Maximum A Posterior (MAP) is utilized to estimate the language model parameters $\{\boldsymbol{\phi}_c\}_{c=1}^{\infty}$ and affinity matrix B, and Maximum Likelihood Estimation (MLE) is utilized to estimate the parameters $\{\boldsymbol{\phi}_c\}_{c=1}^{\infty}$ and $\boldsymbol{\phi}_s$ for logistic regression model. This forms the M-step. During the iterative process, we repeat the E-step and M-step until the likelihood on the training data converges.

We first describe the detailed inference procedures of z_d^i and $z_{i\to j}$ in each user's review documents and social connections.

• **Sampling z_d^i.** Given user u_i, the conditional distribution of z_d^i and the mixing proportion $\boldsymbol{\pi}_i$ is given by,

$$p(\boldsymbol{\pi}_i, z_d^i | D_i, E_i, \boldsymbol{\gamma}, \alpha, \eta, \Psi, \Phi) \propto p(\{z_d^i\}_{d=1}^D | \boldsymbol{\pi}_i) \quad (5)$$
$$p(\{z_{i\to j}\}_{j\neq i}^N | \boldsymbol{\pi}_i) p(\boldsymbol{\pi}_i | \boldsymbol{\gamma}, \eta) p(y_d^i, x_d^i | z_d^{u_i}, \boldsymbol{\psi}_{z_d^i}, \boldsymbol{\phi}_{z_d^i}).$$

Due to the conjugacy between Dirichlet distribution $p(\boldsymbol{\pi}_i|\boldsymbol{\gamma}, \eta)$ and multinomial distributions $p(\{z_d^i\}_{d=1}^D|\boldsymbol{\pi}_i)$ and $p(\{z_{i\to j}\}_{j\neq i}^N|\boldsymbol{\pi}_i)$, we can marginalize out $\boldsymbol{\pi}_i$ in Eq (5). This leaves us the conditional probability of z_d^i in user u_i given his/her rest collective identity assignments,

$$p(z_d^i | \boldsymbol{\gamma}, \eta) = \frac{\Gamma(\eta)}{\Gamma(\eta + n_{i\star} + l_{i\star\star})} \prod_{c=1}^{C} \frac{\Gamma(\eta\boldsymbol{\gamma}_c + n_{i,c} + l_{i\star,c})}{\Gamma(\eta\gamma_c)}, \quad (6)$$

where $n_{i,c}$ denotes the number of reviews in u_i assigned to collective identity c, $l_{i\star,c}$ denotes the number of interactions u_i and his/her friends assigned to collective identity c, $l_{i\star\star}$ denotes the total number of interactions u_i has, and C denotes the total number of unique collective identities at this moment. Thus, Eq (5) can be computed as follows:

$$p(z_d^i = c | D_i, E_i, \boldsymbol{\gamma}, \alpha, \eta, \Psi, \Phi)$$
$$\propto (n_{i,c}^{-d} + l_{i\star,c} + \eta\gamma_c) p(y_d^i, x_d^i | \boldsymbol{\psi}_c, \boldsymbol{\phi}_c), \quad (7)$$

where $n_{i,c}^{-d}$ represents the number of reviews from user u_i assigned to group c except the current review d.

Because of the dynamic nature of DP, we need to account for the possibility that new model components are needed to explain the observations. This requires us to compute the posterior predictive distribution of $p(y_d^i, \boldsymbol{x}_d^i | \boldsymbol{\psi}_c, \boldsymbol{\phi}_c)$, by marginalizing out $\boldsymbol{\psi}_c$ and $\boldsymbol{\phi}_c$. We leverage a sampling scheme proposed in [22] due to the lack of conjugate prior for logistic regression. We introduce a set of auxiliary random variables of size M serving as new possible collective identities, i.e., $\{\boldsymbol{\phi}_m^a\}_{m=1}^M$, to define a valid Markov chain for Gibbs sampling. On the other hand, due to the conjugacy between Dirichlet and multinomial distributions, the posterior predictive distribution of $p(\boldsymbol{x}_d^i | \boldsymbol{\psi}_{z_d^i})$ can be analytically computed to avoid sampling $\boldsymbol{\psi}$ when calculating likelihood in the auxiliary models. Therefore, the posterior distribution of z_d^i can be estimated by,

$$p(z_d^i = c | D_i, E_i, \Phi, \{\boldsymbol{\phi}_m^a\}_{m=1}^M, \Psi) \quad (8)$$
$$\propto (n_{i,c}^{-d} + l_{i\star,c} + \eta\gamma_c) p(y_d^i, x_d^i | \boldsymbol{\psi}_c, \boldsymbol{\phi}_c)$$
$$= \begin{cases} (n_{i,c}^{-i,d} + l_{i\star,c} + \eta\boldsymbol{\gamma}_c) p(y_d^i, x_d^i | \boldsymbol{\psi}_c, \boldsymbol{\phi}_c) & \text{for } 1 \le c \le C, \\ \frac{\eta\gamma_e}{M} p(y_d^i, x_d^i | \boldsymbol{\psi}_c, \boldsymbol{\phi}_c^a) & \text{for } C < c \le C + M. \end{cases}$$

where $\eta\gamma_e$ represents the total proportion for the remaining inactive components in the stick-breaking process. In particular, the $p(y_d^i, \boldsymbol{x}_d^i | \boldsymbol{\psi}_c, \boldsymbol{\phi}_c)$ under existing and new collective identities can be calculated respectively,

$$p(y_d^i, \boldsymbol{x}_d^i | \boldsymbol{\psi}_c, \boldsymbol{\phi}_c) = \quad (9)$$
$$\begin{cases} \frac{1}{1+\exp(-y_d^i \boldsymbol{\phi}_c^T \boldsymbol{x}_d^i)} \frac{m_{d\star}^i!}{\prod_{v=1}^V m_{d,v}^i!} \prod_{v=1}^V {\psi_{c,v}}^{m_{d,v}^i} & \text{for } 1 \le c \le C, \\ \frac{1}{1+\exp(-y_d^i \boldsymbol{\phi}_c^{aT} \boldsymbol{x}_d^i)} \frac{\Gamma(\beta)}{\Gamma(\beta + m_{d\star}^i)} \prod_{v=1}^V \frac{\Gamma(m_{d,v}^i + \beta_v)}{\Gamma(\beta_v)} & \text{for } C < c \le C + M \end{cases}$$

where $m_{d,v}^i$ denotes the frequency of word v in review d and $m_{d\star}^i$ indicates the total number of words in review d. Again, following the design in [11], we have $\bar{\boldsymbol{\phi}}_m^a = \boldsymbol{\phi}_m^a + \boldsymbol{\phi}_s$. Once an auxiliary component is sampled, it will be added to the global collection of collective identities; as a result, the configuration of collective identities is dynamic rather than predefined.

• **Sampling $z_{i\to j}$.** Given user u_i, the conditional distribution of $z_{i\to j}$ and $\boldsymbol{\pi}_i$ is given by,

$$p(\boldsymbol{\pi}_i, z_{i\to j} | D_i, E_i, \boldsymbol{\gamma}, \alpha, \eta, \Psi, \Phi) \propto p(\{z_d^i\}_{d=1}^D | \boldsymbol{\pi}_i) \quad (10)$$
$$p(\{z_{i\to j}\}_{j\neq i}^N | \boldsymbol{\pi}_i) p(\boldsymbol{\pi}_i | \boldsymbol{\gamma}, \eta) p(e_{ij} | z_{i\to j}, z_{j\to i}, B)$$

Similarly to the sampling procedure of z_d^i, we can integrate out $\boldsymbol{\pi}_i$ to obtain the conditional probability of $z_{i\to j}$ as follows,

$$p(z_{i\to j} = c | D_i, E_i, \boldsymbol{\gamma}, \alpha, \eta, \Psi, \Phi) \quad (11)$$
$$\propto (n_{i,c} + l_{i\star,c}^{-(i\to j)} + \eta\gamma_c) p(e^{ij} | z_{i\to j}, z_{j\to i}, B)$$

where $z_{j\to i}$ is the collective identify that u_j chose when interacting with u_i, B is the affinity matrix among all the collective identities, and $l_{i\star,c}^{-(i\to j)}$ represents the number of interactions assigned to collective identity c when user u_i interacts with others except the current interaction.

However, as we could have countably infinite number of collective identities in a collection of users, the dimension of the affinity matrix B is undefined, which makes the explicit calculation of Eq (11) impossible. Fortunately, because of the conjugacy between Beta distribution and Bernoulli distribution, we can integrate out the affinity matrix B to directly calculate the posterior predictive distribution of the collective identity assignment that user u_i has taken when interacting with user u_j,

$$p(z_{i\to j} = c | D_i, E_i, \boldsymbol{\gamma}, \alpha, \eta, \Psi, \Phi, z_{j\to i} = h) \quad (12)$$
$$\propto (n_{i,c} + l_{i\star,}^{-(i\to j)} + \eta\gamma_c) p(e_{ij} | z_{i\to j}, z_{j\to i=h}, B)$$
$$= \begin{cases} (n_{i,c} + l_{i\star,c}^{-(i\to j)} + \eta\gamma_c) B_{ch}^{e_{ij}} (1 - B_{ch})^{(1-e_{ij})} & \text{for } 1 \le c \le C, \\ \eta\gamma_e \frac{\Gamma(e_{ij}+a)\Gamma(1-e_{ij}+b)}{(a+b)\Gamma(a)\Gamma(b)} & \text{for } c = C + 1. \end{cases}$$

Based on the sampled results in E-step, we perform posterior inference of $\{\boldsymbol{\psi}_c\}_{c=1}^C$, B, $\{\boldsymbol{\phi}_c\}_{c=1}^C$ and $\boldsymbol{\phi}_s$ to capture the specification of those learning tasks in each identified collective identity. We should note that as the collective identities have been determined in each user, we do not need to handle the possible generation of new components at this step; and we assume at this stage we have in total C unique collective identities.

• **Estimating $\boldsymbol{\psi}_c$ and B.** We use the Maximum A Posterior principle to infer the configuration of language models and the affinity matrix, as conjugate priors have been postulated on them. Specifically, the posterior distribution of $\boldsymbol{\psi}_c$ follows a Dirichlet distribution: $\boldsymbol{\psi}_c \sim Dirichlet(\boldsymbol{\beta} + \boldsymbol{m})$, where each dimension m_v of $\boldsymbol{m}$ represents the frequency of word v occurring across all the reviews assigned to the collective identity c.

Similarly, the posterior distribution of each element in B follows a Beta distribution: $B_{gh} \sim Beta(a + e_1, b + e_0)$ where e_0 and e_1 denote the number of non-interactions and interactions generated between collective identity h and g accordingly.

• **Estimating $\boldsymbol{\phi}_c$ and $\boldsymbol{\phi}_s$.** As no conjugate prior exists for logistic regression, we appeal to the maximum likelihood principle to estimate $\boldsymbol{\phi}_c$ and $\boldsymbol{\phi}_s$. Given the collective identity assignments in all the

review documents across users, the complete-data log-likelihood over the opinionated contents can be written as,

$$L(\{\boldsymbol{\phi}_c\}_{c=1}^{C}, \boldsymbol{\phi}_s) = \sum_{i=1}^{N}\sum_{d=1}^{D} \log P(y_d^i|\boldsymbol{x}_d^i, \boldsymbol{\phi}_{z_d^i}, \boldsymbol{\phi}_s) + \sum_{c=1}^{C} \log p(\boldsymbol{\phi}_c|\boldsymbol{\mu}, \boldsymbol{\sigma}^2) + \log p(\boldsymbol{\phi}_s|\boldsymbol{\mu}_s, \boldsymbol{\sigma}_s^2). \quad (13)$$

Using a gradient-based optimizer, Eq (13) can be optimized efficiently. With respect to the complete-data log-likelihood, the gradients for $\boldsymbol{\phi}_c$ and $\boldsymbol{\phi}_s$ on a specific training instance $(\boldsymbol{x}_d^i, y_d^i)$ assigned to collective identity z_d^i can be formalized as follows:

$$\frac{\partial L(\cdot)}{\partial \boldsymbol{\phi}_c} = \sum_{i=1}^{N}\sum_{z_d^i=c} \boldsymbol{x}_d^i \left[y_d^i - p(y_d^i = 1|\boldsymbol{x}_d^i)\right] - \frac{(\boldsymbol{\phi}_c - \boldsymbol{\mu})}{\sigma^2},$$

$$\frac{\partial L(\cdot)}{\partial \boldsymbol{\phi}_s} = \sum_{i=1}^{N}\sum_{d=1}^{D} \boldsymbol{x}_d^i \left[y_d^i - p(y_d^i = 1|\boldsymbol{x}_d^i)\right] - \frac{(\boldsymbol{\phi}_s - \boldsymbol{\mu}_s)}{\sigma_s^2},$$

where the gradient for the globally shared sentiment model $\boldsymbol{\phi}_s$ is collected from all the opinionated documents, while the gradient for sentiment model $\boldsymbol{\phi}_c$ of each collective identity is only collected from the documents assigned to it. As a result, $\boldsymbol{\phi}_s$ captures the global pattern in which users express their opinions, and $\boldsymbol{\phi}_c$ captures group-specific properties that users express opinions.

HUB can also predict sentiment polarity in a user's unlabeled review documents, and missing connections between users.

• **Predicting y_d^i.** During the t-th iteration of stochastic EM, we use the newly inferred collective identity z_d^i and corresponding sentiment model to predict y_d^i in review $\boldsymbol{x}_d^i$ of u_i,

$$P(y_d^i|\boldsymbol{x}_d^i, \{\boldsymbol{\phi}_c^t\}_{c=1}^{C_t}, \boldsymbol{\phi}_s^t) = \sum_{c=1}^{C_t} P(z_d^i = c) P(y_d^i = 1|\boldsymbol{x}_d^i, \boldsymbol{\phi}_{z_d^i}^t, \boldsymbol{\phi}_s^t)$$

where $\left(\{\boldsymbol{\phi}_c^t\}_{c=1}^{C_t}, z_d^i, \boldsymbol{\phi}_s^t\right)$ are the inferred latent variables at the tth iteration, $P(z_d^i = c)$ is by Eq (8) for the inferred collective identity for review, and $P(y_d^i|\boldsymbol{x}_d^i, \boldsymbol{\phi}_{z_d^i}^t)$ is computed by Eq (1). The posterior of y_d^i can thus be estimated via an empirical expectation after T iterations,

$$P(y_d^i = 1|\boldsymbol{x}_d^i, \{\boldsymbol{\phi}_c^t\}_{c=1}^{C_t}, \boldsymbol{\phi}_s, \alpha, \eta, \boldsymbol{\gamma}) = \frac{1}{T}\sum_{t=1}^{T} P(y_d^i = 1|\boldsymbol{x}_d^i, \{\boldsymbol{\phi}_c^t\}_{c=1}^{C_t}, \boldsymbol{\phi}_s^t)$$

• **Predicting e_{ij}.** Similarly, for each pair of user u_i and u_j, who are not currently connected in the training data, we can predict their connectivity by,

$$P(e_{ij} = 1|B, \gamma, \eta, a, b) = \frac{1}{T}\sum_{t=1}^{T}\sum_{g,h}\int d\pi_i \int d\pi_j \int dB_{gh} P(e_{ij} = 1|B_{gh}) p(z_{i\to j} = g|\pi_i) p(z_{j\to i} = h|\pi_j) p(\pi_i|\gamma, \eta) p(\pi_j|\gamma, \eta) \quad (14)$$

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.

3.4 Discussion

• **Modeling Sparsity.** The social network is usually sparse. That is, they contain many zeros or non-interactions. We distinguish two sources of non-interactions between them: the rarity of interactions in general or the pair of users rarely interact. It is reasonable to expect a large portion of non-interactions is caused by the limited opportunity of contact instead of deliberate choices. Thus, we introduce a sparsity parameter ρ to accommodate the two resources where we define ρ as the proportion of deliberate choices among both interactions and non-interactions. Thus, the corresponding probability of generating a social connection can be rewritten as:

$$p(e_{ij}|z_{i\to j}, B, z_{j\to i}) = \begin{cases} \rho B_{z_{i\to j}, z_{j\to i}} & e_{ij} = 1, \\ 1 - \rho B_{z_{i\to j}, z_{j\to i}} & e_{ij} = 0. \end{cases}$$

• **Computational Complexity.** Inferring the latent collective identity z_d^i that each user has used in generating the review document $(\boldsymbol{x}_d^i, y_d^i)$ is computationally cheap. Specifically, by Eq (8), updating the membership of all the documents imposes a complexity of $O\left(N\bar{D}(C+M)\right)$, where $N\bar{D}$ is the total number of documents, C is the number of collective identities and M is the size for auxiliary collective identities. While inferring the latent collective identity for each interaction requires a complexity of $O(N^2C)$, as we need to consider interaction among each pair of users. With the consideration of sparsity, the computation for modeling interactions can be greatly reduced from $O(N^2C)$ to $O(\rho N^2C)$ as ρ usually takes a small value indicating the proportion of deliberate choices. The overall complexity for the proposed algorithm is thus $O\left(N\bar{D}(C+M) + \rho N^2C\right)$.

• **Summarization.** The proposed HUB model achieves the joint modeling of opinionated content and social network structure. As explicitly expressed in Eq (8), inferring the collective identity for each document of one user depends on not only the other documents of the current user, but also the interactions between this current user and other users. This is also true for inferring the collective identity for each pair of users as stated in Eq (12). This mutual effects of textual documents and social connections well align with the Self Consistency Theory: the user-generated documents and interactions can be understood as two different representations of self with the dependency being the integrality in humans.

As noticed, the assignment of latent collective identity for a particular document d in user u_i is determined by three factors: 1) the proportion of the current collective identity $\eta\boldsymbol{\gamma}_c$; 2) the number of documents and interactions of user u_i belongs to a particular collective identity $(n_{i,c}^{-i,d} + l_{i\star,c})$; 3) the likelihood of the given document under a candidate collective identity $p(y_d^i, x_d^i|\boldsymbol{\psi}_c, \boldsymbol{\phi}_c)$. As a result, the choice of a proper collective identity for each document/interaction not only relies on individual-level factors, i.e., whether the candidate collective identity can best explain the current document/interaction, but also aggregate-level factors, i.e., if the candidate collective identity closely aligns with the current user's other observations, together with its own popularity.

By inferring the posterior distributions of latent variables, important knowledge about each user can be discovered. First, the posterior distributions of the set of parameters $\theta_c = (\boldsymbol{\psi}_c, \boldsymbol{\phi}_c, \boldsymbol{b}_c)$ reveal the distribution of words, sentiment preferences and pair-wise affinities in a particular collective identity. Second, the posterior

distribution of each user's personal identity $p(\pi_i|u_i)$, which is defined as the assignment of collective identities for this particular user, depicts an individual user's intent in history. In fact, this distribution of each user, together with the learned affinities between different collective identities, provides pair-wise user affinity that is useful for many personalized applications.

4 EXPERIMENTS

We evaluated the effectiveness of our proposed user modeling solution on two large collections of Amazon and Yelp reviews, together with their network structures. Both quantitative and qualitative evaluations are performed to assess the effectiveness of the proposed solution for user modeling.

4.1 Datasets

We used two publicly available review datasets collected from Amazon [21] and Yelp [1], for our evaluation purpose. In these two datasets, each user is associated with a set of reviews, each of which contains various attributes such as author ID, review ID, timestamp, textual content, and an opinion rating in a discrete five-star range. The Yelp dataset provides user friendship imported from users' Facebook friend connections, while there is no explicit social network in Amazon dataset. We utilized the "co-purchasing" information to build the network structure for Amazon users.

We pre-processed the two datasets: 1) labeled the reviews with 1 and 2 stars as negative, and those with 4 and 5 stars as positive (reviews with 3 stars are considered as neutral, and thus ignored); 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 constructed the text feature vector for each review by both unigrams and bigrams based on a union of top features selected by Chi-square and information gain. We selected 5,000 and 3,071 features for Amazon and Yelp datasets respectively. From the resulting datasets, we randomly sampled 9,760 Amazon reviewers and 10,830 Yelp reviewers for evaluation. From 9,760 Amazon users, there are 105,472 positive and 37,674 negative reviews; and from 10,830 Yelp users, there are 157,072 positive and 51,539 negative reviews. Correspondingly, we have 269,180 edges and 113,030 edges in the resulting Amazon and Yelp social networks respectively, resulting in an average of 27.6 and 10.5 friends per user. This indicates most users are not directly connected with others.

4.2 The Formation of Collective Identities

First of all, it is important to study the inferred collective identities by the proposed model. We traced the complete-data log-likelihood during the iterative process, the number of inferred collective identities, together with the sentiment classification quality in hold-out testing reviews, during each iteration of posterior inference in HUB to assess the model's clustering property and predictive ability. The results on the two datasets are demonstrated in Figure 2. We collected results in every three iterations after the burn-in period.

It is clear that the likelihood keeps increasing during the iterative process and converges later on. It increases much faster at the earlier stage when more collective identities are generated to cover the diversity in user behaviors. Accordingly, the collective identities become stable and a more accurate estimate of behavior models can be achieved in the later stage, leading to the improved sentiment classification performance, especially for the negative class as its training observations are quite limited.

We are also interested in the unified behaviors of each collective identity learned through the paired tasks in HUB. As we examine the most frequently used words under different collective identities for generating review contents, it is easy to recognize the cohesive factor that defines them, such as the type of restaurants in Yelp and the category of products in Amazon. Correspondingly, each collective identity is associated with its own language style to express sentiment polarity. At the same time, the interactions among different collective identities are also discovered to indicate the affinity among them. In order to demonstrate the learned behaviors of collective identities, we selected a subset of collective identities learned from Yelp dataset and visualized their corresponding behavior patterns in Figure 3.

In Figure 3, each collective identity is represented as a node with three behavior patterns: a word cloud summarizes the frequently used words, a set of sentiment words depict the attitudes, and connections to other nodes represent the affinity. As shown in the word cloud, it is easy to tell the cohesive semantics of each collective identity, i.e., Asian v.s., Italian restaurants. The two sets of words on the left are the most representative words used to indicate the sentiment polarities under each collective identity, the words in orange are associated with positive learnt weights and those in blue are with negative weights. It is clear that different collective identities tend to use different words to express opinions. The green lines among collective identities indicate the affinity between each pair of them, with darker and thicker lines indicating stronger affinity. We can recognize that strong connections are detected in "similar" collective identities where the similarity can be understood from both their interested topics and the corresponding sentiment words. By jointly modeling different modalities of user-generated data, HUB recognized the collective identities which are interpretable and descriptive.

4.3 Personalized Sentiment Classification

In order to evaluate the effectiveness of opinionated content modeling by HUB, we compared the proposed HUB model with the following five baselines for sentiment classification: 1) **MT-SVM**: it is a state-of-the-art multi-task learning solution proposed in [6], which encodes the task relatedness via a shared linear kernel across tasks without accumulating similar behavior patterns for information sharing. 2) **MTLinAdapt+kMeans**: to verify the effects of proposed clustering algorithm, we followed [10] to perform k-means clustering of users based on training reviews to estimate sentiment model for each user group in a multi-task fashion. 3) **cLinAdapt**: the model [11] leveraged the Dirichlet Process to explore the clustering property among users while neither mixed membership for each user nor network structure is considered in the exploration. 4) **cLinAdapt+HDP**: the model considered individual's diversity inside a group by jointly modeling the generation of text contents and the sentiment labels, while social connections

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

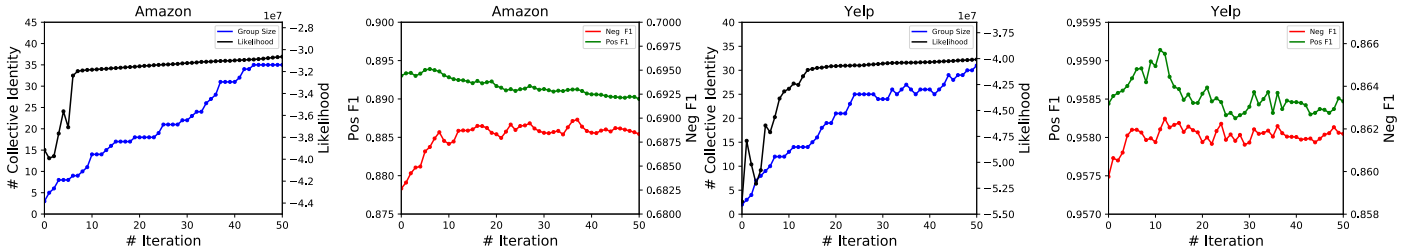

Figure 2: Trace of likelihood, model size and sentiment classification performance when training HUB on Amazon and Yelp.

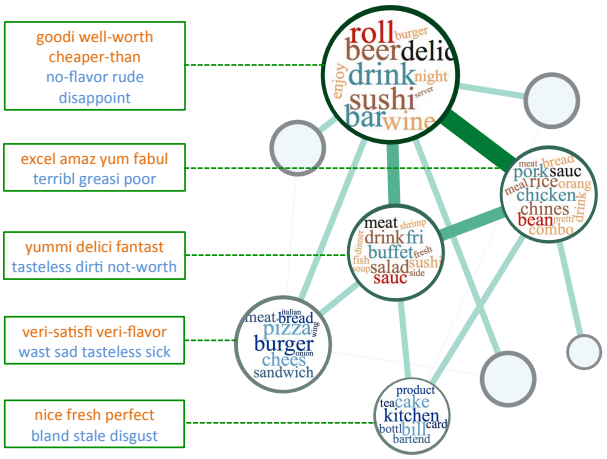

Figure 3: The identified behavior patterns among a subset of collective identities on Yelp dataset.

are neglected. 5) **Graph-Based Semi-Supervised Learning (GB-SSL)** : we followed [41] to construct a network among the textual reviews based on two layers of proximity between each pair of documents: the affinity between the owners of the two reviews and the proximity of the text contents of the two reviews.

In this experiment, we chronologically partitioned the review data into two parts: the first half for training and the second half for testing. Due to the biased class distribution in both datasets, we used F1 measure as evaluation metric for both classes in each user, and used macro average among users to compare the classification performance. Detailed performance can be found in Table 1. Due to the large variance of the review size among users, there exists large variance in the macro average F1 across all the users. Therefore, we utilized the Wilcoxon signed-rank test to verify whether their population mean ranks differ, i.e., whether the difference between a paired value is significant.

Overall, HUB achieves encouraging classification performance as it outperforms all the baselines except MT-SVM for Amazon dataset. Compared with the k-means based user grouping strategy, the automatically identified collective identities can better capture the commonality shared among all users. Different from the DP-based clustering algorithm, cLinAdapt, HUB relaxes the assumption that one user can only have one single membership and allows mixed membership, which directly helps capture the diversity existing in each individual user, thus yields better sentiment classification performance. cLinAdapt+HDP baseline assigns

Table 1: Personalized sentiment classification results.

Models	Amazon		Yelp	
	Neg F1	Pos F1	Neg F1	Pos F1
Base	0.6300	0.8858	0.8141	0.9385
MT-SVM	**0.6929***	**0.8992***	0.8633	0.9591
MTLinAdapt+kMeans	0.6224	0.8390	0.8453	0.9336
cLinAdapt	0.6842	0.8752	0.8574	0.9527
cLinAdapt+HDP	0.6846	0.8868	0.8556	0.9566
GBSSL	0.6179	0.8847	0.8303	0.9529
HUB	0.6905	0.8934	**0.8647***	**0.9595***

*: p-value<0.05 under Wilcoxon signed-rank test.

each user's mixed membership simply based on review contents, thus it cannot benefit from the information provided by social network. GBSSL leverages the social connections as regularization to enhance the sentiment prediction, while mutual influence between text contents and network structure are ignored.

In order to further diagnose the performance difference between MT-SVM and HUB, we looked into the classification performance with respect to user-specific statistics, i.e., the number of reviews and the number of friends. We found that MT-SVM performs well on those users with sufficient training review data. That is, either the user has a large amount of reviews to support an accurate estimation of his/her sentiment model; or the user's attitude is quite unified and a handful of reviews are sufficient. However, our proposed model is superior to MT-SVM for the users with limited amount of review data while possessing a rich number of friends, i.e., users with an average of 2.2 training reviews and an average of 40 friends. This observation clearly reflects the unique advantage of our proposed model: the social connections help users with limited training reviews identify appropriate collective identities, thus achieve more accurate sentiment classification results.

4.4 Serve for Collaborative Filtering

Collaborative filtering is popularly utilized in modern recommender systems to make predictions about the interests of a user by collecting information from others. The key component is to infer the similarity between users in order to achieve accurate recommendations. The learned distinct personal identity of each user from HUB, i.e., the mixture of collective identities, naturally serves as a good proxy of user preferences.

In this experiment, we evaluated the utility of learned personal identities of individual users and affinity among collective identities from HUB, in a collaborative filtering based recommendation. In order to construct a valid set of ranking candidates, we split each

user's reviewed items into two sets: the items in training reviews and those in testing reviews. The training reviews are utilized to train each user's personal identity while the testing reviews provide the relevant items for ranking. For a specific user u_i, we also selected irrelevant items from the users who have rated u_i's purchased items in their training set. Since many items are rarely rated, we utilized the popularity of each item as the threshold to filter the irrelevant items. The popularity is defined as the number of reviews the item received among all the users' training reviews and the same set of candidate items are maintained in all the algorithms. 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 overall ratings to act as ranking score for this item. Normalized discounted cumulative gain (NDCG) and mean average precision (MAP) are used to measure the quality of the recommendation.

To evaluate the recommendation performance, we selected three algorithms which achieve decent sentiment classification performance among the five baselines, i.e., MT-SVM, cLinAdapt, cLinAdapt+HDP, and leveraged their learned sentiment models for similarity calculation on both Amazon and Yelp datasets. For the proposed HUB model, each pair of users' personal identities, together with the corresponding social affinity, are utilized to calculate the similarity between them,

$$sim(u_i, u_j) = \sum_{g=1}^{G} \sum_{h=1}^{H} u_{i,g} \cdot u_{j,h} \cdot B_{gh} \tag{15}$$

We include a baseline that makes recommendations by the simple average of ratings from all the users who reviewed the item, and name it as Average. In addition, since low-rank matrix factorization based solutions have achieved decent empirical performance in collaborative filtering, we also include two baselines, SVD++ [15] and factorization machine (FM) [27] for comparison.

In the testing phrase, we selected users with at least one relevant ranking item, which resulted in 7216 and 9247 valid users for the Amazon and Yelp respectively. We selected 3 and 50 as popularity threshold, which ended up with 31 and 103 average ranking candidates for the two datasets. Because the average number of users who reviewed the same item in training data is 1.3 in Amazon and 5.8 in Yelp, we select top-4 neighbors for ranking score calculation for both datasets. We reported the NDCG and MAP performance across all users in Table 2. As we can see, HUB achieves encouraging recommendation performance on both datasets, which indicates the learned personal identity and social affinity accurately capture the relatedness among users regarding their preferences over the recommended items. Matrix factorization based methods can only exploit the observed association between users and items, but not the opinionated text contents. Due to the very sparse distribution of items in both datasets, matrix factorization based methods suffer in performance.

4.5 Serve for Friend Recommendation

The social network structure modeling in HUB helps friend recommendation. In particular, it is more important to provide friend recommendation to new users in a system: they may actively post textual contents but may have very few friends in the system, which may lead to poor friend recommendation from most existing network-based recommendation solutions. Our proposed model can overcome this limitation by utilizing user-generated textual contents to infer their personal identity, thus to provide helpful friend recommendation.

Table 2: Collaborative filtering results on Amazon and Yelp.

Models	Amazon		Yelp	
	NDCG	MAP	NDCG	MAP
Average	0.7813	0.6573	0.6606	0.4700
MT-SVM	0.7982	0.6798	0.7519	0.5847
cLinAdapt	0.7926	0.6725	0.7548	0.5898
cLinAdapt+HDP	0.7956	0.6766	0.7598	0.5989
SVD++	0.5502	0.3853	0.5731	0.3880
FM	0.4874	0.3110	0.4057	0.1979
HUB	**0.7993**	**0.6816**	**0.7685**	**0.6082**

Table 3: Friend recommendation results on Yelp.

Train Size	BoW		SVM		HUB	
	NDCG	MAP	NDCG	MAP	NDCG	MAP
4000	1.0003	1.0230	1.0314	1.3130	**1.1017**	**1.8779**
6000	1.0002	1.0419	1.0128	0.9222	**1.1137**	**1.5928**
8000	1.0010	1.0887	1.0602	1.4194	**1.1428**	**2.6532**

In order to verify the effectiveness of the proposed model with respect to friend recommendation, we split the whole set of users into varying sizes of training users and a fixed set of testing users. As we only have the friendship for Yelp dataset, we performed this experiment on this dataset. More specifically, we selected 4000, 6000 and 8000 users for training and utilized another set of 2830 users for testing. For the training users, all their textual reviews, together with the social connections are used for model training. Based on the trained model, each testing user's personal identity is inferred based on their review contents. Then, both the social affinity and the personal identity are utilized to calculate the similarity between each pair of users to serve as the ranking score for friend recommendation, as described in Eq (14).

We include a baseline which utilized SVM to estimate the affinity among different collective identities. We also include another baseline which represents each user with a BoW representation by aggregating all their reviews. Though matrix factorization based methods are widely used in recommender systems, they do not apply in this case as there is no direct friendship connection between training and testing users. NDCG and MAP are utilized to evaluate the effectiveness and we reported the performance in Table 3 by comparing against a random solution, i.e., divide the performance by the performance of a random recommendation.

It is clear the proposed model achieves the best performance in friend recommendation as the accurate proximity between pairs of users are properly identified. A simple BoW representation cannot well represent users and therefore leads to poor similarity measurement between users. Compared with the SVM based learning method, our model can benefit from the affinity between distinct collective identities, thus to provide an accurate approximation of user similarity. This experiment further verifies the effectiveness of the identified affinity among collective identities. At the same time, it proves the necessity for joint modeling of opinionated content

modeling and network structure modeling in order to get an overall understanding of users.

5 CONCLUSION AND FUTURE WORK

In the paper, we studied the problem of user behavior modeling by utilizing multiple types of user generated data. We proposed a generative model HUB to integrate two companion learning tasks of opinionated content modeling and social network structure modeling, for a holistic modeling of user intents. The learning tasks are paired and clustered to reflect the homogeneity among users while each user is modeled as a mixture over the instances of paired tasks to indicate heterogeneity. The learned user behavior models are interpretable and predictive in enabling more accurate sentiment classification and item/friend recommendations on two large collections of review documents from Amazon and Yelp with corresponding social network structures.

Several areas are left open for our future explorations. In the current work, we have not considered the timestamp of the generation of reviews or social connections while they are generated sequentially. Utilizing the temporal information can further help capture the evolution of social network and users' behaviors. We have assumed each opinionated document is associated with one specific language model, which only allows us to identify categorical patterns of word distribution, rather than the fine-grained topical patterns. Introducing topic models would enable us to study more detailed user opinions at an aspect level.

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.

REFERENCES

[1] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. 2008. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research* 9, Sep (2008), 1981–2014.
[2] Simon Bourigault, Cedric Lagnier, Sylvain Lamprier, Ludovic Denoyer, and Patrick Gallinari. 2014. Learning social network embeddings for predicting information diffusion. In *Proceedings of the 7th ACM WSDM*. ACM, 393–402.
[3] Kewei Cheng, Jundong Li, Jiliang Tang, and Huan Liu. 2017. Unsupervised Sentiment Analysis with Signed Social Networks. In *AAAI*. 3429–3435.
[4] Jean Diebolt and Eddie HS Ip. 1996. Stochastic EM: method and application. In *Markov chain Monte Carlo in practice*. Springer, 259–273.
[5] A Evgeniou and Massimiliano Pontil. 2007. Multi-task feature learning. *Advances in neural information processing systems* 19 (2007), 41.
[6] Theodoros Evgeniou and Massimiliano Pontil. 2004. Regularized multi–task learning. In *Proceedings of the 10th ACM SIGKDD*. ACM, 109–117.
[7] Hongliang Fei, Ruoyi Jiang, Yuhao Yang, Bo Luo, and Jun Huan. 2011. Content based social behavior prediction: a multi-task learning approach. In *Proceedings of the 20th ACM CIKM*. ACM, 995–1000.
[8] Thomas S Ferguson. 1973. A Bayesian analysis of some nonparametric problems. *The annals of statistics* (1973), 209–230.
[9] Anindya Ghose, Panagiotis G Ipeirotis, and Beibei Li. 2012. Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Marketing Science* 31, 3 (2012), 493–520.
[10] Lin Gong, Mohammad Al Boni, and Hongning Wang. 2016. Modeling social norms evolution for personalized sentiment classiofication. In *Proceedings of the 54th ACL*, Vol. 1. 855–865.
[11] Lin Gong, Benjamin Haines, and Hongning Wang. 2017. Clustered Model Adaption for Personalized Sentiment Analysis. In *Proceedings of the 26th WWW*. 937–946.
[12] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. 1983. Stochastic blockmodels: First steps. *Social networks* 5, 2 (1983), 109–137.
[13] Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD*. ACM, 168–177.
[14] Xia Hu, Lei Tang, Jiliang Tang, and Huan Liu. 2013. Exploiting social relations for sentiment analysis in microblogging. In *Proceedings of the 6th WSDM*. ACM, 537–546.
[15] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *IEEE ICDM*. Ieee, 263–272.
[16] Zepeng Huo, Xiao Huang, and Xia Hu. 2018. Link Prediction with Personalized Social Influence. (2018).
[17] Prescott Lecky. 1945. Self-consistency; a theory of personality. (1945).
[18] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. 2010. Signed networks in social media. In *Proceedings of the SIGCHI*. ACM, 1361–1370.
[19] David Liben-Nowell and Jon Kleinberg. 2007. The link-prediction problem for social networks. *journal of the Association for Information Science and Technology* 58, 7 (2007), 1019–1031.
[20] Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies* 5, 1 (2012), 1–167.
[21] Julian McAuley, Rahul Pandey, and Jure Leskovec. 2015. Inferring networks of substitutable and complementary products. In *Proceedings of the 21th ACM SIGKDD*. ACM, 785–794.
[22] Radford M Neal. 2000. Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics* 9, 2 (2000), 249–265.
[23] Mark EJ Newman. 2006. Modularity and community structure in networks. *Proceedings of the national academy of sciences* 103, 23 (2006), 8577–8582.
[24] Rina S Onorato and John C Turner. 2004. Fluidity in the self-concept: the shift from personal to social identity. *European Journal of Social Psychology* 34, 3 (2004), 257–278.
[25] Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd ACL*. ACL, 115–124.
[26] Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval* 2, 1-2 (2008), 1–135.
[27] Steffen Rendle. 2010. Factorization machines. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 995–1000.
[28] Xuehua Shen, Bin Tan, and ChengXiang Zhai. 2005. Implicit user modeling for personalized search. In *Proceedings of the 14th ACM CIKM*. ACM, 824–831.
[29] Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. 2011. Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the First workshop on Unsupervised Learning in NLP*. Association for Computational Linguistics, 53–63.
[30] Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li. 2011. User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD*. ACM, 1397–1405.
[31] Jiliang Tang, Chikashi Nobata, Anlei Dong, Yi Chang, and Huan Liu. 2015. Propagation-based sentiment analysis for microblogging data. In *Proceedings of the 2015 SIAM International Conference on Data Mining*. SIAM, 577–585.
[32] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *Proceedings of the 24th WWW*. 1067–1077.
[33] Yee W Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2005. Sharing clusters among related groups: Hierarchical Dirichlet processes. In *Advances in neural information processing systems*. 1385–1392.
[34] Ivan Titov and Ryan T McDonald. 2008. A Joint Model of Text and Aspect Ratings for Sentiment Summarization.. In *ACL*, Vol. 8. Citeseer, 308–316.
[35] John C Turner. 1985. Social categorization and the self-concept: A social cognitive theory of group behavior. *Advances in group processes* 2 (1985), 77–122.
[36] Hongning Wang, Yue Lu, and ChengXiang Zhai. 2011. Latent aspect rating analysis without aspect keyword supervision. In *Proceedings of the 17th ACM SIGKDD*. ACM, 618–626.
[37] Ya Xue, Xuejun Liao, Lawrence Carin, and Balaji Krishnapuram. 2007. Multi-task learning for classification with Dirichlet process priors. *Journal of Machine Learning Research* 8, Jan (2007), 35–63.
[38] Jun Yan, Ning Liu, Gang Wang, Wen Zhang, Yun Jiang, and Zheng Chen. 2009. How much can behavioral targeting help online advertising?. In *Proceedings of the 18th international conference on World wide web*. ACM, 261–270.
[39] Jaewon Yang, Julian McAuley, and Jure Leskovec. 2013. Community detection in networks with node attributes. In *Data Mining (ICDM), 2013 IEEE 13th international conference on*. IEEE, 1151–1156.
[40] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th ACM SIGIR*. ACM, 83–92.
[41] Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. 2003. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*. 912–919.